UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Milena M. Šošić

Modelovanje moralnih i emocionalnih aspekata jezika u klasifikaciji konverzacionih tekstova

DOKTORSKA DISERTACIJA

Beograd, 2025

Mentor:

prof. dr Jelena Graovac, vanredni profesor

Matematički fakultet, Univerzitet u Beogradu

Članovi komisije:

prof. dr Nenad Mitić, redovni profesor

Matematički fakultet, Univerzitet u Beogradu

prof. dr Mladen Nikolić, vanredni profesor

Matematički fakultet, Univerzitet u Beogradu

prof. dr Ranka Stanković, redovni profesor

Rudarsko-geološki fakultet, Univerzitet u Beogradu

Datum odbrane: _____

UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Milena M. Šošić

Modeling moral and emotional language aspects in conversational texts classification

DOCTORAL DISSERTATION

Belgrade, 2025

Advisor:

prof. dr Jelena Graovac, associate professor

Faculty of Mathematics, University of Belgrade

Committee members:

prof. dr Nenad Mitić, full professor

Faculty of Mathematics, University of Belgrade

prof. dr Mladen Nikolić, associate professor

Faculty of Mathematics, University of Belgrade

prof. dr Ranka Stanković, full professor

Faculty of Mining and Geology, University of Belgrade

Defense date: _____

Posvećeno mojim roditeljima.

,,Reči božanskog saznanja: moralno uzdizanje; trajna osećanja uzvišenosti, radosti, veselja; ubrzanje moralnog osećaja, koji se čini važnijim od intelektualnog razumevanja stvari; usklađenost univerzuma duž moralnih linija, a ne intelektualnih; spoznaja da je osnovni princip postojanja ono što nazivamo ljubavlju, koja se ponekad ne ispoljava jasno, ne čisto, ne odmah, ipak neizbežno"

– Jan Martel, Pijev život

,,Words of divine consciousness: moral exaltation; lasting feelings of elevation, elation, joy; a quickening of the moral sense, which strikes one as more important than an intellectual understanding of things; an alignment of the universe along moral lines, not intellectual ones; a realization that the founding principle of existence is what we call love, which works itself out sometimes not clearly, not cleanly, not immediately, nonetheless ineluctably."

– Yann Martel, Life of Pi

Zahvalnica

Veliku zahvalnost upućujem mentoru, prof. dr Jeleni Graovac, na pažljivom i posvećenom vođenju kroz sve faze ovog istraživanja, koje je u velikoj meri oblikovalo i usmerilo ovaj naučno-istraživački rad. Prof. dr Ranki Stanković upućujem iskrenu zahvalnost za dragocenu pomoć u sticanju znanja i veština iz oblasti računarske obrade srpskog jezika. Izuzetno sam zahvalna prof. dr Nenadu Mitiću, koji me je svojim znanjem i posvećenošću uveo u svet naučno-istraživačkog rada, podstakao moju radoznalost i otvorio put daljem profesionalnom usavršavanju. Veliku zahvalnost dugujem i prof. dr Mladenu Nikoliću, čiji su konstruktivni saveti i stručne sugestije iz oblasti mašinskog učenja u značajnoj meri doprineli kvalitetu ovog rada.

Takođe, želim da izrazim posebnu zahvalnost profesorima Matematičkog fakulteta Univerziteta u Beogradu na bogatstvu znanja i temeljnim osnovama koje sam tokom školovanja stekla, a koje su mi omogućile da sa sigurnošću i stručnošću pristupim ovom istraživačkom poduhvatu. Posebnu zahvalnost izražavam članovima multidisciplinarne akademske zajednice za razvoj jezičkih resursa JeRTeh, koji su svoje resurse podelili sa mnom, čime su omogućili da ovo istraživanje bude uspešno sprovedeno.

Najdublju zahvalnost dugujem majci Sibinki čija je neiscrpna ljubav, razumevanje i podrška bila moj stalni oslonac i podsticaj za sticanje novih saznanja. Sa zahvalnošću se obraćam i mojim rođacima, prijateljima i kolegama, koji su verovali u mene i ohrabrivali me da istrajem na ovom istraživačkom putu.

U Beogradu, septembra 2025. *Milena Šošić*

Naslov doktorske disertacije: Modelovanje moralnih i emocionalnih aspekata jezika u klasifikaciji konverzacionih tekstova

Rezime: Konverzacione tekstualne poruke predstavljaju važan oblik digitalne komunikacije u savremenom društvu. Sa razvojem informacionih tehnologija, pojavili su se različiti alati za komunikaciju, kao što su elektronska pošta, društvene mreže, alati za brze poruke i sistemi za automatsko generisanje odgovora. Poruke koje nastaju u ovim alatima, za razliku od standardnih tekstova, imaju specifičnu strukturu koja omogućava klasifikaciju pojedinačnih poruka ili skupova poruka koje zajedno čine jednu konverzaciju. Klasifikaciona obeležja su određena zadatkom koji se rešava i mogu biti jednoznačna ili višeznačna, čime se omogućava prepoznavanje složenih međuzavisnosti koje mogu postojati između kategorija.

Uvođenje moralnih i emocionalnih dimenzija jezika u istraživanje je od suštinskog značaja za razumevanje složenih obrazaca ljudske komunikacije, posebno u kontekstu digitalnih platformi i društvenih medija. Metode mašinskog učenja (eng. *Machine Learning*, ML), poput dubokih neuronskih mreža (eng. *Deep Neural Networks*, DNN), omogućavaju korišćenje i preciznije prepoznavanje ovih aspekata, i istovremeno pružaju efikasan način za klasifikaciju emocija i moralnih vrednosti koje su izražene u tekstovima. Primetna složenost u izražavanju ljudskih emocija i moralnih vrednosti, koje su često implicitno izražene i zavise od konteksta, čini njihovo prepoznavanje naročito izazovnim.

Jedan od velikih izazova je nedostatak ili ograničenost postojećih resursa po veličini i raznovrsnosti kod jezika sa nedovoljno razvijenim resursima (eng. *low resource languages*), kojima pripada i srpski jezik. Izgradnja jezičkih resursa, kao što su obeleženi leksikoni i korpusi, igra važnu ulogu u ovom procesu, jer se na taj način obezbeđuju neophodni izvori znanja za izgradnju i unapređivanje postojećih ML modela. Jezički resursi omogućavaju modelima da nauče kako različita emocionalna izražavanja i iskazane moralne vrednosti utiču na ton i značenje komunikacije. U tom cilju, za srpski jezik je najpre razvijen semantički leksikon intenziteta sentimenta *SentiWords.SR* od ∼15k reči, kao i pridruženi programski alat *SRPOL* za merenje intenziteta sentimenta u tekstualnim sekvencama na srpskom jeziku. Nadalje, razvijeni su semantički leksikon emocionalnog afekta *EmoLex.SR* koji sadrži ∼9.8k reči sa pridruženim intenzitetom emocionalnog afekta, kao i semantički leksikon moralnih vrednosti *MFD.SR* koji sadrži ∼4.3k reči sa pridruženim težinama osnovnih moralnih vrednosti. Najzad, značajan napor je uložen u obeležavanje prvih konverzacionih korpusa sa društvenih mreža u emocionalne i moralne kategorije. U tom pogledu, razvijen je obeležen korpus *Social-Emo.SR* (∼34.6k poruka), sastavljen od podkorpusa *Twitter-Emo.SR* (∼16.7k poruka) i *Reddit-Emo.SR* (∼17.9k poruka) preuzetih sa Tviter (eng. *Twitter*) i Redit (eng. *Reddit*) društvenih mreža, u tom redosledu. Analogno, pretraživanjem značajnih ključnih reči, iz korpusa *Social-Emo.SR* izdvojen je deo poruka sa sadržajima u kojima se potencijalno iskazuju moralni stavovi. Ovaj skup poruka, pod nazivom *Social-Mor.SR* (∼13.6k poruka), ručno su proverili i obeležili (ljudski) anotatori, a korpus je sastavljen od podkorpusa *Twitter-Mor.SR* (∼6.1k Tviter poruka) i *Reddit-Mor.SR* (∼7.5k Redit poruka).

U kontekstu DNN mreža, modeli zasnovani na arhitekturama kao što su rekurentne mreže ili transformeri, obučavani nad jezičkim resursima, omogućavaju korišćenje i prepoznavanje emocionalnih i moralnih aspekata jezika u različitim kontekstima. Napredni algo-

ritmi podrazumevaju korišćenje dvosmerne rekurentne mreže sa dugom kratkoročnom memorijom (eng. *Bidirectional Long Short-Term Memory*, BiLSTM) uz opciono uključivanje dodatnog mehanizma pažnje (eng. Attention mechanism, Att). Kombinacija ovih algoritama sa jezički i kulturološki prilagođenim resursima, odnosno atributima teksta koji se pomoću tih resursa kreiraju (Meta atributi), otvara mogućnosti za analizu moralnih i emocionalnih aspekata jezika, sa širokom primenom u klasifikacionim zadacima kao što su prepoznavanje ličnog konteksta, istinitosti objava ili tipa delovanja u digitalnim komunikacijama. U prepoznavanju ličnog konteksta, odnosno klasifikaciji poruka korporativne elektronske pošte na poslovnu i ličnu, rezultati su pokazali da korišćenje pažljivo osmišljenog hibridnog pristupa (BiLSTM-Att+Meta) nad svim porukama konverzacione grane daje najbolje rezultate koji su u rangu drugih objavljenih rezultata na istom zadatku. U eksperimentima prepoznavanja istinitosti glasina, kao i tipa delovanja na glasinu, pokazano je da moralni i emocionalni atributi teksta kreirani pomoću semantičkih leksikona (EmoAttr, MorAttr $\subseteq$ Meta) doprinose poboljšanju tačnosti klasifikacije za +4.2% i +3.8%, u tom redosledu, u odnosu na metode bez uključivanja ovih atributa.

Na zadatku prepoznavanja emocionalnog afekta u konverzacionim tekstovima na srpskom jeziku, eksperimenti su pokazali da modeli transformer arhitekture nastali doobučavanjem osnovnih modela, sa dostignutim vrednostima F_1 mere od $\sim$53%, dostižu rezulate prijavljene na zadataku višeznačne klasifikacije na istom skupu emocionalnih kategorija. Eksperimenti su, takođe, pokazali da dodatno procesiranje i balansiranje podataka doprinosi poboljšanju performansi ovih modela. Na zadatku prepoznavanja moralnih vrednosti i moralnog sentimenta, korišćenjem korpusa *Social-Mor.SR* i njegovih podkorpusa, dostignute vrednosti F_1 mere od $\sim$46% za prepoznavanje moralne vrednosti, i F_1 mere od $\sim$38% za prepoznavanje moralnog sentimenta. Ovi rezultati ukazuju na prihvatljiv nivo tačnosti, ali istovremeno naglašavaju potrebu za daljim podešavanjem modela kako bi se postigle optimalne performanse. Eksperimenti doobučavanja *LLaMA* (eng. *Large Language Models Meta AI*) modela postigli su prihvatljive, ali nešto niže rezultate u odnosu na modele *BERT* (eng. *Bidirectional Encoder Representations from Transformers*) arhitekture. Sve performase, s obzirom na direktnu zavisnost od podataka nad kojima su izgrađeni, imaju potencijal daljeg unapređenja, nakon provere i balansiranja inicijalnih obeležja u korišćenim korpusima.

Ključne reči: Moralnost, emocionalnost, konverzacioni tekstovi, klasifikacija

Naučna oblast: Računarstvo

Uža naučna oblast: Računarska obrada teksta

UDK broj:

Disertation title: Modeling moral and emotional language aspects in conversational texts classification

Abstract: Conversational text messages represent an important form of digital communication in modern society. With the development of information technologies, various communication tools have emerged, such as email, social media, instant messaging tools, and automated response systems. Messages generated within these tools, unlike standard texts, have a specific structure that allows for the classification of individual messages or sets of messages that form a conversation. Classification labels are defined by the specific task being addressed and can be either single-label or multi-label, which enables the recognition of complex interrelationships between the categories.

Introducing moral and emotional dimensions of language into research is crucial for understanding the complex patterns of human communication, particularly in the context of digital platforms and social media. Machine learning (ML) methods, such as deep neural networks (DNN), facilitate the utilization and more precise recognition of these aspects while simultaneously providing an efficient way to classify emotions and moral values expressed in texts. The noticeable complexity in the expression of human emotions and moral values, which are often conveyed implicitly and depend heavily on context, makes their recognition particularly challenging.

One of the major challenges is the lack of or limited availability of resources in terms of size and diversity for low-resource languages, including Serbian. The development of linguistic resources, such as annotated lexicons and corpora, plays a crucial role in this process by providing the necessary knowledge sources for building and improving existing ML models. Linguistic resources enable models to learn how different emotional expressions and moral values influence the tone and meaning of communication. To support this, a semantic lexicon for sentiment intensity, *SentiWords.SR*, containing approximately 15k words, was developed for the Serbian language, along with the associated tool *SRPOL* for measuring sentiment intensity in textual sequences in Serbian. Additionally, a semantic lexicon for emotional affect, *EmoLex.SR*, comprising around 9.8k words with assigned emotional intensity values, and a semantic lexicon for moral values, *MFD.SR*, consisting of approximately 4.3k words with associated moral value weights, were developed. Significant efforts were also made in annotating the first conversational corpora from social media with emotional and moral categories. In this regard, the *Social-Emo.SR* corpus ($\sim$34.6k messages) was developed, consisting of the *Twitter-Emo.SR* subcorpus ($\sim$16.7k messages) and the *Reddit-Emo.SR* subcorpus ($\sim$17.9k messages), collected from Twitter and Reddit, respectively. Furthermore, by searching for key moral-related terms, a subset of messages expressing potential moral stances was extracted from *Social-Emo.SR*. This subset, named *Social-Mor.SR* ($\sim$13.6k messages), was manually verified and annotated by human annotators and consists of the *Twitter-Mor.SR* subcorpus ($\sim$6.1k Twitter messages) and the *Reddit-Mor.SR* subcorpus ($\sim$7.5k Reddit messages).

In the context of DNN architectures, models based on recurrent networks or transformers, trained on these resources, enable the recognition and utilization of emotional and moral aspects of language in various contexts. The combination of advanced algorithms, such as Bidirectional Long Short-Term Memory (BiLSTM) networks and the attention mechanism with linguistically and culturally adapted resources (Meta) opens new possibilities

for analyzing moral and emotional aspects of language. This has broad applications in classification tasks such as recognizing personal context, truthfulness of posts, or types of engagement in digital communication. For personal context recognition, i.e. classifying corporate emails as either business-related or personal, results show that using a carefully designed hybrid approach (BiLSTM-Att+Meta) across entire conversation branches yields the best results, comparable to published benchmarks on the same task. In experiments related to rumor veracity classification and identifying engagement types in response to rumors, it was demonstrated that moral and emotional attributes derived from semantic lexicons (EmoAttr, MorAttr $\subseteq$ Meta) improve classification accuracy by +4.2% and +3.8% respectively, compared to methods without these attributes.

For emotion recognition in Serbian conversational texts, experiments revealed that transformer-based models fine-tuned on the task achieved F_1-scores of approximately 53%, reaching performance levels reported for multi-label classification on the same emotional category set. Additionally, experiments showed that further data preprocessing and balancing improved model performance. In moral value and moral sentiment classification tasks, using the *Social-Mor.SR* corpus and its subcorpora, an F_1-score of $\sim$46% was achieved for moral value recognition and $\sim$38% for moral sentiment recognition, indicating acceptable results but also the need for further model optimization. Fine-tuning LLaMA models yielded reasonable but slightly lower performance compared to BERT-based architectures. Since model performance is directly dependent on the data they are trained on, there is potential for further improvements by refining and balancing initial annotations in the utilized corpora.

Keywords: Morality, emotions, conversational texts, classification

Scientific field: Computer science

Scientific subfield: Natural language processing

UDK number:

Sadržaj

1. Uvod

Konverzacione tekstualne poruke predstavljaju jedan od najvažnijih načina razmene informacija u savremenom digitalnom okruženju. Sa razvojem savremenih tehnologija, posebno interneta i mobilnih uređaja, ova vrsta komunikacije postala je nezamenljiva u svakodnevnom životu. Kroz različite platforme, poput društvenih mreža, alata za brzo komuniciranje, alata elektronske pošte ili alata za automatsko generisanje odgovora, korisnici mogu da brzo i efikasno razmenjuju neophodne informacije. Konverzacioni alati omogućavaju razmenu sadržaja različitog formata, čime se simulira izražavanje emocija i iskazivanje vrednosnih stavova iz realnog života. U pogledu grupne komunikacije, digitalni konverzacioni alati omogućavaju efikasniju organizaciju i koordinaciju među učesnicima, čime se umanjuju ograničenja vezana za geografsku udaljenost. Međutim, porast korišćenja digitalnih konverzacija donosi nove izazove u automatizaciji njihove obrade, kao što je potreba za automatskom klasifikacijom razmenjenih poruka, u cilju daljeg unapređenja kvaliteta digitalne komunikacije.

Specifična struktura konverzacionih poruka zahteva pažljiv pristup prilikom klasifikacije ove vrste podataka. Klasifikacija se može vršiti na nivou pojedinačne poruke ili čitavog niza poruka, u različite vrste klasa definisanih zadatkom klasifikacije. Pored tradicionalnih metoda mašinskog učenja (eng. *Machine Learning, ML*) primenjenih na pojedinačne poruke, razvoj specijalizovanih algoritamskih arhitektura dubokih neuronskih mreža (eng. *Deep Neural Networks, DNN*) predstavlja poseban izazov za modelovanje sekvencijalnih zavisnosti između poruka u konverzacionim nizovima. Dodatni izazovi uključuju promenljivu dužinu konverzacionih nizova, ponavljanje istih poruka u različitim sekvencama, naglašavanje pojedinih delova sekvence, kao i uključivanje pridruženih atributa sadržaja poruka.

Za primenu ML metoda, odnosno metoda veštačke inteligencije (eng. *Artificial Intelligence, AI*) u širem smislu, neophodno je tekstualne sadržaje poruka predstaviti kao numeričke vektore koje ML algoritmi zahtevaju. Ovaj postupak, poznat kao vektorizacija teksta, podrazumeva predstavljanje reči ili sekvence reči kao numeričke vektore na osnovu njihove učestalosti pojavljivanja u pojedinačnom dokumentu i korpusu. Pored vektorizacije, tekstualni sadržaji mogu biti obogaćeni pridruženim atributima, kao što su oni izračunati korišćenjem specijalizovanih leksikona. Atributi afektivnih kategorija, kao što su intenziteti sreće, tuge ili straha, pružaju ML algoritmima dodatno semantičko znanje koje može unaprediti performanse klasifikacije i doprineti boljem razumevanju tekstualnih sadržaja. Modelovanje moralnih i emocionalnih aspekata jezika u konverzacionim tekstovima je važno za unapređenje razumevanja ljudske komunikacije, posebno u kontekstu klasifikacije konverzacionih tekstova. Sadržaj koji se kreira u dijalozima ne prenosi samo informacije, već odražava i dublje vrednosti, uverenja i emocionalna stanja njihovih autora. Ovi aspekti mogu uticati na način tumačenja sadržaja i razumevanje društvenih interakcija. S obzirom na uočene složenosti, prepoznavanje moralnih i emocionalnih dimenzija jezika predstavlja značajan izazov, ali istovremeno pruža velike mogućnosti za napredak na opštem zadatku analize konverzacionih tekstova.

Značaj moralnih i emocionalnih aspekata jezika u klasifikaciji konverzacionih tekstova može zavisiti od osobina učesnika, tema diskusije, stila interakcije ili vrste konverzacije, odnosno platforme ili alata koji se koristi. Karakterne osobine učesnika, poput ekstrovertnosti ili introvertnosti, utiču na to kako pojedinci izražavaju emocije i moralne stavove [144, 84]. Na primer, ekstroverti su skloniji izražavanju pozitivnih emocija, dok introverti preferiraju

oprezniji pristup komunikaciji [117]. Teme razgovora, kao što su poslovne, geo-političke, sportske, ili informativne, određuju ton i emocionalni intenzitet komunikacije [165, 225, 43]. Dodatno, stil interakcije tokom razgovora značajno oblikuje moralne i emocionalne aspekte jezika. Agresivni pristupi (napad ili suprotstavljanje) izazivaju intenzivnije emocionalne reakcije, dok konstruktivno-asertivni stilovi (podrška i postavljanje pitanja) podstiču empatiju i pozitivne moralne reakcije [27, 28, 157]. Najzad, formalnost konverzacije značajno utiče na način izražavanja emocija i moralnih stavova. U formalnim razgovorima, učesnici su skloniji korišćenju sentimentalno neutralnog jezika, dok neformalne konverzacije pokazuju intenzivnije emocionalne sadržaje [93]. Iz tog razloga, konverzacije na društvenim mrežama imaju poseban značaj za analizu moralnih i emocionalnih aspekata jezika. Društvene mreže omogućavaju masovnu i brzu razmenu informacija, čime se stvaraju velike količine tekstualnih podataka koji odražavaju širok spektar emocija i moralnih stavova njihovih autora. Ovi podaci predstavljaju bogat izvor za istraživanje i razvoj algoritama za automatsku klasifikaciju i analizu teksta. Konverzacije na društvenim mrežama često sadrže neformalni i emocionalno obojen jezik, što ih čini posebno relevantnim za proučavanje afektivnih i moralnih dimenzija komunikacije [77]. Osim toga, društvene mreže pružaju mogućnost za analizu dinamike grupnih interakcija, što je važno za razumevanje kolektivnih emocionalnih i moralnih reakcija u društvu i pruža uvid u kompleksne društvene procese i trendove.

Iako je u oblasti računarske lingvistike postignut značajan napredak u prepoznavanju moralnih i emocionalnih aspekata za jezike poput engleskog, srpski jezik još uvek nema dovoljno razvijene jezičke resurse u ovim oblastima što ukazuje na potrebu za istraživanjima u cilju njihovog razvoja. Napredni pristupi, poput velikih jezičkih modela (eng. *Large Language Model, LLM*), se sve više koriste za automatsku analizu složenih moralno-emocionalnih aspekata komunikacije, čime je omogućeno njihovo bolje razumevanje [81, 158, 55]. Iako ovi modeli donose značajna unapređenja u obradi prirodnih jezika (eng. *Natural Language Processing, NLP*), na zadacima klasifikacije teksta LLM zahtevaju dodatno obučavanje zasnovano na specijalizovanim resursima. U tom cilju, specijalizovani jezički resursi za emocionalni i moralni aspekt jezika, kao što su specijalizovani leksikoni i obeleženi korpusi, mogu poslužiti za razvoj naprednih arhitektura za prepoznavanje ovih aspekata, čime se značajno unapređuje razumevanje i omogućava istraživanje digitalne komunikacije na srpskom jeziku.

1.1. Predmet, cilj i hipoteze istraživanja

Predmet ovog istraživanja jeste analiza uticaja moralnih i emocionalnih aspekata jezika u zadacima klasifikacije konverzacionih tekstova. Uticaj ovih aspekata će se meriti po stepenu njihovog doprinosa i značaja na uspešnost klasifikacije odnosno na uspešnost podele podataka u kategorije definisane zadatkom klasifikacije. Primenom matematički zasnovanih metoda za pronalaženje značajnih atributa klasifikacije utvrdiće se u kojoj meri moralni i emocionalni atributi uzimaju učešće u ovako utvrđenim skupovima. Istraživanje u ovom radu će biti zasnovano na analizi konverzacionih tekstova na engleskom i srpskom jeziku primenom matematički zasnovanih računarskih metoda, i to prvenstveno metoda i algoritama mašinskog učenja, kako tradicionalnog, tako i naprednih metoda dubokog učenja.

Glavni **cilj** ovog istraživanja jeste predlaganje generičkog hibridnog pristupa za klasifikaciju konverzacionih tekstova. Poseban cilj predstavlja razvoj novih resursa za srpski jezik - leksikona i korpusa sa obeleženim moralnim i emocionalnim kategorijama. Obeleženi korpusi će biti iskorišćeni za pravljenje klasifikacionog modela zasnovanog na algoritmima

mašinskog učenja, kao posebno značajnog resursa, za prepoznavanje moralnih i emocionalnih aspekata u tekstovima na srpskom jeziku. Jedan od važnih ciljeva jeste i utvrđivanje značaja moralnih i emocionalnih aspekata jezika na klasifikaciju pojedinačnih ili nizova konverzacionih poruka, kao i utvrđivanje stepena njihove međusobne zavisnosti.

Hipoteze od kojih se polazi u ovom istraživanju su:

H1 ***Uključivanje moralnih i emocionalnih aspekata jezika značajno doprinosi tačnosti modela klasifikacije konverzacionih tekstova, što ukazuje na njihov značaj u interpretaciji i razumevanju konverzacionih tekstova.***

H2 ***Razvijeni semantički leksikoni moralnih i emocionalnih afekata reči za srpski jezik mogu doprineti prepoznavanju ovih aspekata u tekstu.***

H3 ***Kvantitativnom analizom atributa moguće je utvrditi korelaciju između iskazanih moralnih stavova i emocionalnih reakcija u konverzacionim porukama.***

H4 ***Moguće je razviti klasifikacione modele prihvatljive tačnosti koji na osnovu karakteristika tekstualnih sadržaja iz razvijenih obeleženih konverzacionih korpusa na srpskom jeziku, mogu da prepoznaju moralne i emocionalne aspekte jezika.***

Pregled literature iz istraživanja moralnosti i emocionalnosti na koje se oslanja ovo istraživanje je dato u poglavlju 2, sa posebnim osvrtom na istaknute načine obeležavanja i razvijene resurse u drugim jezicima koji su predstavljeni u poglavlju 6. Različite vrste konverzacionih tekstova su analizirane u poglavlju 3. Osnovne naučne metode koje će biti korišćene u istraživanju zasnovane su na statističkim metodama i klasifikacionim algoritmima u oblasti mašinskog učenja, koje su predstavljene u poglavlju 4. Predložena metodologija klasifikacije konverzacionih poruka korišćenjem hibridnog pristupa je predstavljena u poglavlju 7, kojoj prethodi poglavlje 5 sa opisanim načinima vektorizacije teksta koji se u predloženoj metodi koriste. Poglavlje 9 prikazuje proveru razultata klasifikacije nad dva različita zadatka klasifikacije konverzacionih poruka pomoću metode predložene u ovom radu, kao i proveru ispravnosti razvijenih resursa za srpski jezik. U poglavlju 10 se diskutuju dobijeni rezultati i mogućnost potvrđivanja postavljenih hipoteza u ovom istraživanju, nakon kojeg sledi poglavlje 11 kojim se zaključuje ovaj rad. Deo rezultata ove disertacije sadržan je u objavljenim radovima [184], [185] i [186], dok su drugi materjali pokriveni ovim radom, a koji se odnose na razvoj drugih jezičkih resursa za srpski jezik, planirani za objavljivanje u narednom periodu.

2. Moralnost i emocije

2.1. Moralnost

Moral predstavlja jedan od osnovnih koncepata u ljudskom društvu koji obuhvata principe i vrednosti koji oblikuju etičko ponašanje i donošenje odluka svakog pojedinca. Ovaj složen koncept nastaje pod uticajem kulturnih, religioznih i filozofskih vrednosti, kao i individualnih verovanja i iskustava. Izučavanje moralnosti ima dugu istoriju u filozofiji i teologiji, a poslednjih decenija sve više postaje predmet naučnih istraživanja. Naučnici iz različitih oblasti, uključujući psihologiju, antropologiju i sociologiju, pokušavaju da razumeju prirodu i poreklo moralnog rasuđivanja i ponašanja. Na pitanje o prirodi morala postoji nekoliko teorijskih pristupa, svaki sa specifičnim utemeljenjem i perspektivom. Prvi pristup, konstrukcionistički, smatra moral objektivnim i univerzalnim, pri čemu moralna načela i vrednosti važe za sve pojedince i kulture bez obzira na lična uverenja; ovo je u skladu sa deontološkim etičkim teorijama, poput Kantove etike, koje naglašavaju značaj moralnih pravila utemeljenih na racionalnosti [36]. Suprotno tome, relativistički pristup tvrdi da je moral subjektivan i da se razlikuje zavisno od kulturnih, istorijskih i društvenih konteksta; ovaj stav je povezan s kulturnim relativizmom, koji ističe da ne postoji univerzalni moralni standard primenljiv na sve ljude [73]. Treći pristup je evolucionistički, koji moral posmatra kao rezultat biološke i evolucione adaptacije, gde su moralna načela i vrednosti razvijeni radi promocije društvene saradnje u cilju opstanka [111].

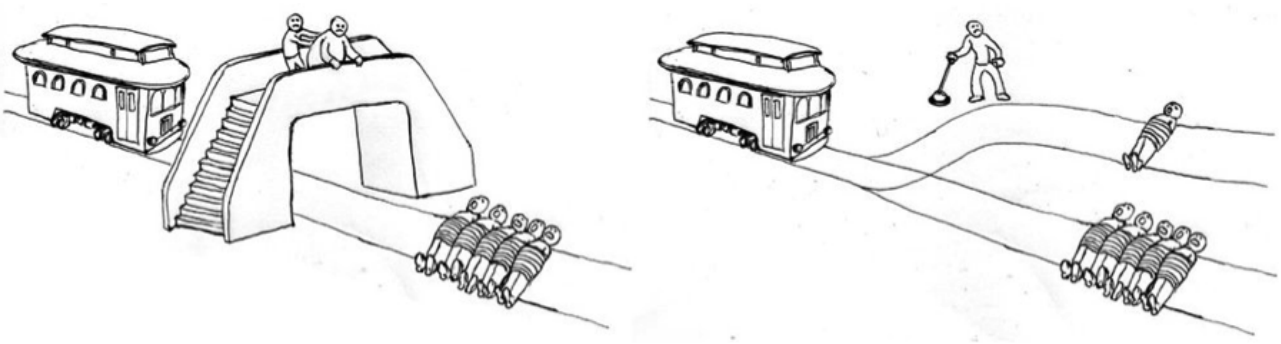
Pod direktnim uticajem relativističkog pristupa, razvijena je Teorija o moralnim osnovama (eng. *Moral Foundations Theory, MFT*) koja tvrdi da moralna uverenja nisu univerzalna već da su izgrađena nad kombinacijama osnovnih gradivnih elemenata [66]. Teorija potvrđuje da se moralne procene koje ljudi prave mogu kategorizovati u pet moralnih osnova razloženih na dihotomne parove: **`briga/povreda`** (eng. **`care/harm`**), **`pravednost/prevara`** (eng. **`fairness/cheating`**), **`lojalnost/izdaja`** (eng. **`loyalty/betrayal`**), **`autoritet/subverzija`** (eng. **`authority/subversion`**), **`svetost/degradacija`** (eng. **`sanctity (purity)/degradation`**), čiji su prevodni ekvivalenti[1] i opisi prikazani u tabeli 2.1. Na ovaj način predstavljena teorija omogućava opisivanje stepena moralnosti kod pojedinaca ili čitavih društvenih zajednica. Teorija je pod neprestanom analizom naučnika koji u svojim radovima predlažu njeno proširivanje drugim moralnim osnovama kao što je, na primer, podela moralne osnove pravednosti na proporcionalnost i jednakost [15].

Jedan od važnijih moralnih problema u filozofskoj etici, poznat kao i problem tramvaja (eng. *trolley problem*), je klasičan misaoni eksperiment koji postavlja moralnu dilemu u pogledu izbora između dva hipotetička pravca delovanja. Postoji više varijacija ovog problema, a osnovni scenario uključuje tramvaj koji se kreće prema grupi ljudi pri čemu je neophodno doneti odluku šta uraditi: da se ne učini ništa i dozvoli da tramvaj nastavi svojim putem ili da se tramvaj preusmeri na drugu stazu, u kom slučaju bi ubio manje ljudi [61]. Razvijene su mnoge varijacije problema sa tramvajem, uključujući verziju pešačkog mosta, verziju petlje i verziju debelog čoveka (pogledati sliku 2.1). Problem tramvaja se naširoko koristi u filozofiji, psihologiji i neuronauci, a koristi se i kao alat za prikazivanje moralnih dilema, istraživanje moralnog odlučivanja, prirode morala i odnosa između

[1] Za kategorizaciju tekstova u moralne vrednosti (moralne osnove) i moralni sentiment (razlaganje moralnih osnova na dihotomne parove) biće korišćeni nazivi kategorija na engleskom jeziku u cilju standardizovanja dobijenih rezultata i omogućavanja njihovog neposrednog korišćenja u drugim istraživanjima.

Tabela 2.1: *Osnovne moralne vrednosti prema Teoriji o moralnim osnovama predstavljene dihotomnim parovima*

Moralna vrednost	Kategorija	Opis
briga/povreda	**care/harm**	Senzitivnost prema patnji i dobrobiti drugih, zasnovana na empatiji i emocionalnoj povezanosti. Motiviše ponašanja koja uključuju zaštitu, negu i pomoć.
pravednost/prevara	**fairness/cheating**	Težnja ka pravičnim i ravnopravnim odnosima među pojedincima, koja je utemeljena na principima jednakosti, reciprociteta i poštenja.
lojalnost/izdaja	**loyalty/betrayal**	Identifikacija i posvećenost sopstvenoj grupi - podrazumeva spremnost na žrtvu u cilju očuvanja kohezije i stabilnosti grupe.
autoritet/subverzija	**authority/subversion**	Poštovanje legitimnih autoriteta, društvenih hijerarhija i tradicije, koje doprinosi očuvanju reda, discipline i institucionalne stabilnosti.
svetost/degradacija	**sanctity/degradation**	Ideal čistote i uzdržavanja od telesnih i moralno nepoželjnih uticaja, često povezan sa religijskim normama i konceptima duhovnog uzdizanja.



Slika 2.1: *Problem tramvaja u filozofskoj etici [61]*

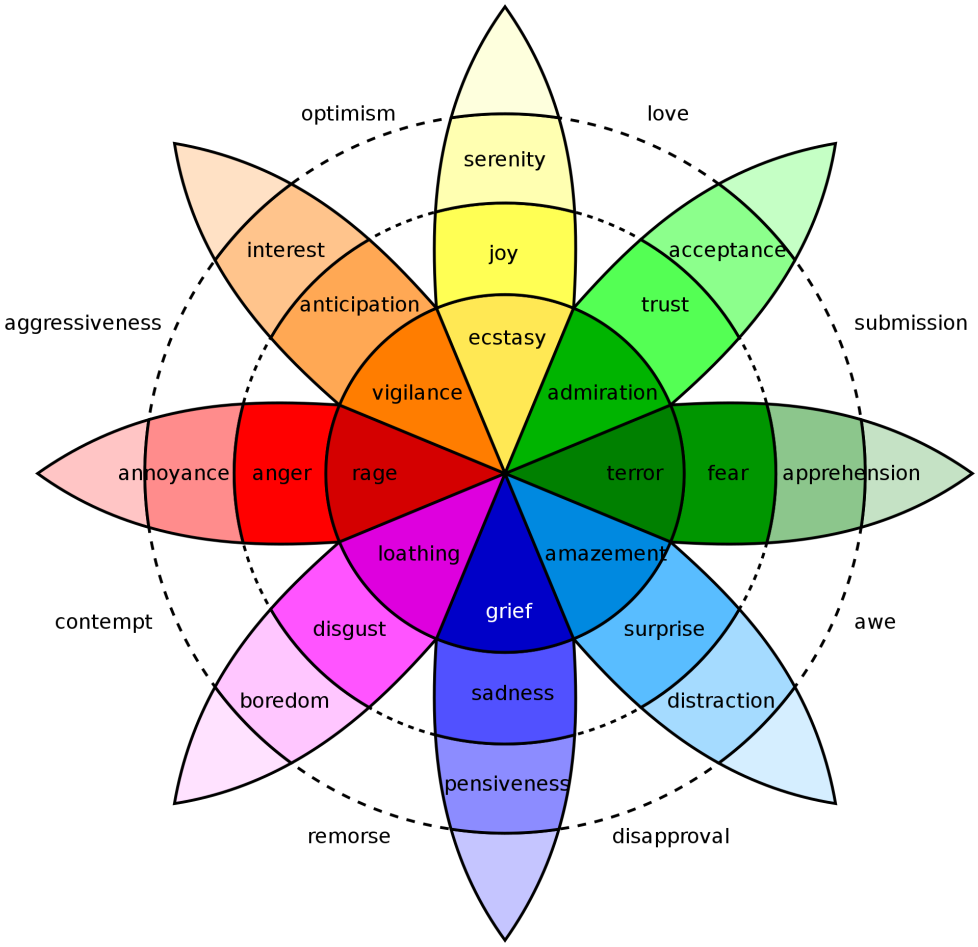
morala i emocija.

Tokom protekle decenije, polje moralne psihologije je istraživalo uticaje na moralno rasuđivanje, otkrivajući vitalnu ulogu koju emocije imaju u pokretanju ovih procesa [198]. Istraživanja su takođe pokazala da emocije kao što su empatija, krivica, stid i gađenje mogu uticati na moralno rasuđivanje i motivisati pojedince da se ponašaju u skladu sa moralnim principima [70]. Na primer, empatija, koja je sposobnost da se razumeju osećanja drugih, može navesti pojedince da dožive moralne emocije kao što su saosećanje i briga za druge. Emocije kao što su krivica i stid, koje se povezuju sa pogrešnim postupanjem ili kršenjem moralnih normi, mogu motivisati pojedince da se iskupe za svoje postupke i da izbegnu buduće prestupe. S druge strane, emocije poput ljutnje ili straha ponekad mogu ometati moralno rasuđivanje i dovesti do kršenja moralnih vrednosti. Na primer, bes može dovesti do agresivnog i osvetničkog ponašanja, dok strah može dovesti do izbegavanja ili zanemarivanja moralnih obaveza [58]. Autor u [205] ispituje moguće uticaje i sličnosti između razvijenih teorija o moralnosti i emocionalnosti i daje preporuku kako se one mogu iskoristiti za bolje razumevanje porekla ovih, jako složenih mehanizama, kao i rešavanje postojećih teoretskih neslaganja.

2.2. Emocije

Emocije su složena psihološka i fiziološka iskustva koja se pokreću raznim unutrašnjim i spoljašnjim stimulansima. Odlikuju ih subjektivna osećanja, fiziološki odgovori i izrazi ponašanja. Emocije mogu uticati na naše misli, ponašanje i interakcije sa drugima i igraju ključnu ulogu u životnim procesima. Najuticajnije psihološke teorije o emocijama obuhvataju različite pristupe razumevanju odnosa između fizioloških reakcija i emocio-

nalnih iskustava. Džejms-Langeova teorija tvrdi da emocije nastaju kao rezultat fizioloških reakcija na spoljašnje stimuluse, gde osećamo emocije zbog telesnih promena poput ubrzanog rada srca. S druge strane, Kanon-Bard teorija sugeriše da emocije i fiziološki odgovori nastaju istovremeno i nezavisno jedan od drugog kao odgovori na isti stimulus. Šahter-Singerova teorija dva faktora predlaže da emocije proizilaze iz kombinacije fiziološkog uzbuđenja i kognitivne interpretacije situacije, gde specifična emocija zavisi od interpretacije. Konačno, teorija procene objašnjava emocije kao rezultat naše procene situacije, gde emociju određuje način na koji smo situaciju vrednovali, kao pozitivnu ili negativnu.



Slika 2.2: *Plutčikov točak emocija [153]*

Poslednjih decenija, pojavile su se nove psihološke teorije koje uzimaju primat u istraživanju emocija, a nastale su pod direktnim uticajem prethodnih. Takva je, na primer, teorija Ekmanovih emocija koja pretpostavlja postojanje šest osnovnih emocija - bes, strah, tuga, gađenje, sreća i iznenađenje, koje je psiholog Pol Ekman identifikovao kao univerzalne u različitim kulturama i jezicima [51]. Ekmanovo istraživanje facijalnih ekspresija je imalo značajan uticaj na proučavanje emocija koje je doprinelo boljem razumevanju kako se emocije izražavaju i prepoznaju. Ekmanova teorija je zapravo proširenje Džejms-Langeove teorije koja fiziološke reakcije posmatra kao osnovnu komponentu emocionalnog iskustva.

S druge strane, Plutčikove emocije se odnose na teoriju koju je predložio psiholog Robert Plutčik, a koja opisuje osam osnovnih emocija za koje se veruje da su univerzalne u svim kulturama i društvima [153]. Osam emocija koje Plutčik prepoznaje jesu

Tabela 2.2: *Kategorizacija emocija prema intenzitetu u Plutčikovom modelu*

Primarna emocija	Intenzitet	Sekundarna emocija	Složena emocija
			agresivnost / *aggressiveness*
iščekivanje / anticipation	1	zainteresovanost / *interest*	
	2	iščekivanje / *anticipation*	
	3	opreznost / *vigilance*	optimizam / *optimism*
radost / joy	1	spokoj / *serenity*	
	2	radost / *joy*	
	3	ushićenje / *ecstasy*	ljubav/ *love*
poverenje / trust	1	prihvatanje / *acceptance*	
	2	poverenje / *trust*	
	3	divljenje / *admiration*	potčinjenost / *submission*
strah / fear	1	zabrinutost / *apprehension*	
	2	strah / *fear*	
	3	užas / *terror*	strahopoštovanje / *awe*
iznenađenje / surprise	1	uznemirenost / *distraction*	
	2	iznenađenje / *surprise*	
	3	zaprepašćenje / *amazement*	neodobravanje / *disapproval*
tuga / sadness	1	zamišljenost / *pensiveness*	
	2	tuga / *sadness*	
	3	patnja / *grief*	pokajanje / *remorse*
gađenje / disgust	1	dosađivanje / *boredom*	
	2	gađenje / *disgust*	
	3	gnušanje / *loathing*	prezir / *contempt*
ljutnja / anger	1	nerviranje / *annoyance*	
	2	ljutnja / *anger*	
	3	bes / *rage*	agresivnost / *aggressiveness*

ljutnja (eng. **anger**), **iščekivanje** (eng. **anticipation**), **gađenje** (eng. **disgust**), **strah** (eng. **fear**), **radost** (eng. **joy**), **tuga** (eng. **sadness**), **iznenađenje** (eng. **surprise**) i **poverenje** (eng. **trust**). Plutčikova teorija emocija sugeriše da su ovih osam emocija primarni gradivni blokovi svih ljudskih emocija, a da se druge emocije mogu posmatrati kao kombinacije ili varijacije osnovnih emocija. Plutčik je predložio da predstavljanje emocija na dijagramu koji podseća na točak sa 8 grana koje označavaju primarne emocije, pri čemu su dihotomne emocije postavljene na granama koje se nalaze jedna nasuprot drugoj, kao što je prikazano na slici 2.2. Na primer, radost je pozicionirana nasuprot tuge (**joy** $\leftrightarrow$ **sadness**), poverenje je nasuprot odvratnosti (**trust** $\leftrightarrow$ **disgust**), a strah je nasuprot ljutnje (**fear** $\leftrightarrow$ **anger**). Plutčikov model, takođe, prepoznaje različite intenzitete primarnih emocija, odnosno sekundarne emocije, pri čemu su emocije sa većim intenzitetom prikazane u sektoru na grani koji je bliži sredini točka. Tako je, na primer, nerviranje emocija koja je manjeg intenziteta od ljutnje, koja je, sa druge strane, manjeg intenziteta od besa (nerviranje $\ll$ ljutnja $\ll$ bes). Pored toga, Plutčikov model prepoznaje postojanje složenih emocija koje nastaju kao kombinacija primarnih emocija različitog intenziteta (pogledati tabelu 2.2). Na primer, optimizam je prema Plutčikovom modelu prepoznat kao složena emocija koja nastaje kao kombinacija primarnih emocija iščekivanja i radosti. Ova vrsta racionalnog predstavljanja emocija sugeriše na direktan uticaj Šahter-Singerove teorije na razvoj Plutčikove emocionalne teorije. U tabeli 2.2 predstavljeni su nazivi primarnih, sekundarnih i složenih emocija sa

prevodnim ekvivalentima na srpskom jeziku[2].

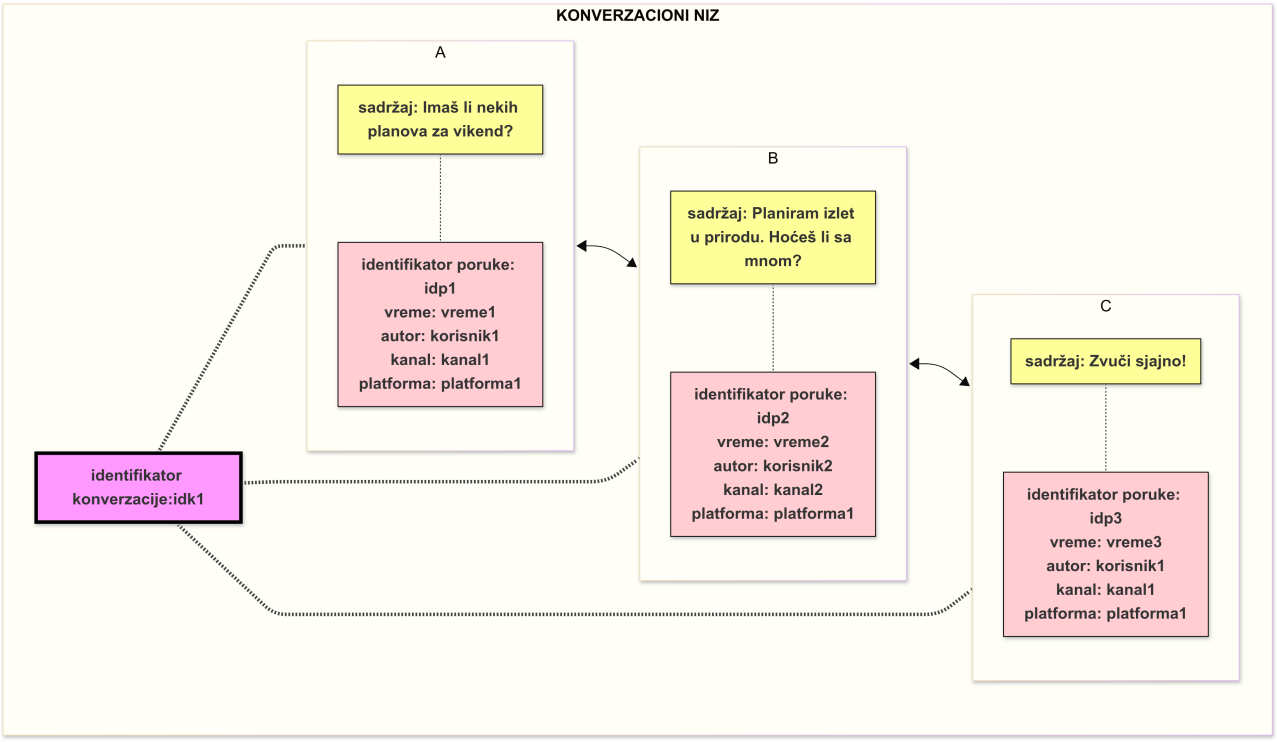
Poređenjem Ekmanove i Plutčikove teorije možemo uvideti da obe identifikuju osnovni skup kulturološki nezavisnih i univerzalnih emocija. Međutim, dok je Ekman identifikovao šest osnovnih emocija, Plutčik je identifikovao osam, čije se značenje u dve teorije ne poklapa u potpunosti. Pored toga, Plutčikova teorija naglašava ideju da se emocije mogu posmatrati kao kombinacije ili varijacije osnovnih emocija, dok se Ekmanova teorija ne bavi eksplicitno ovom idejom. Još jedna ključna razlika između ove dve teorije je način na koji konceptualizuju emocije. Ekmanova teorija se zasniva na ideji da su emocije diskretne i da se mogu identifikovati preko izraza lica, dok Plutčikova teorija naglašava ideju da su emocije složene i višestruke i da na njih može uticati širok spektar faktora, koji uključuju misli, verovanja i dosadašnja iskustva. Uprkos ovim razlikama, i Ekmanova i Plutčikova teorija su uticajne u oblasti psihologije i na njima se zasniva veliki broj savremenih istraživanja.

[2]Za klasifikaciju tekstova u kategorije emocionalnog afekta (primarne emocije) u okviru ovog istraživanja biće korišćeni nazivi kategorija na engleskom jeziku u cilju standardizovanja dobijenih rezultata i omogućavanja njihovog neposrednog korišćenja u drugim istraživanjima.

3. Konverzacioni tekstovi

Konverzacioni tekstovi predstavljaju bilo koju vrstu pisane komunikacije koja uključuje dve ili više osoba koje učestvuju u razgovoru. Konverzacioni niz se sastoji od sekvence međusobno povezanih poruka koje zajedno formiraju dijalog. U strukturi poruke u različitim vrstama konverzacionih tekstova mogu se identifikovati sledeći glavni segmenti (slika 3.1):

- Sadržaj konverzacione poruke;
- Potpis konverzacione poruke – deo strukture poruke koji sadrži propratne informacije o poruci kao što su jedinstveni identifikator, vremenska odrednica slanja, autor, kanal komunikacije, platforma konverzacije i druge;
- Veza sa prethodnom porukom u nizu.



Slika 3.1: *Identifikovani segmenti u neposrednim porukama u konverzacionom nizu*

Sa razvojem internet tehnologija u poslednjim godinama, pojavili su se i brojni alati za komuniciranje među ljudima. Prema formatu sadržaja, konverzacione poruke se mogu svrstati u tri glavne kategorije: tekstualne, audio i video poruke. Iz poslednje dve kategorije, tehnikama ručne ili automatske transkripcije, mogu se dobiti tekstualni sadržaji koji predstavljaju njihov najvažniji deo, pri čemu se zanemaruju prateći vizuelni i zvučni efekti. Poruke, takođe, mogu sadržati i kombinaciju navedenih formata kada kažemo da imaju višemodalni (eng. *multi-modal*) karakter. U poslednje vreme, sve je više primera skupova konverzacionih podataka koji sadrže višemodalne podatke [156].

Tekstualne konverzacije se mogu analizirati na više nivoa i to kao:

- Pojedinačne poruke – predstavljaju osnovne jedinice tekstualne konverzacije. Analiza pojedinačnih poruka se vrši radi utvrđivanja njihove sintakse i semantike, jezika,

kategorizacije kao što su namera, način delovanja, moralne vrednosti, sentiment ili emocije koje prenosi.

- Parovi poruka – predstavljaju neposredne poruke u konverzaciji koje su najčešće poslali različiti korisnici. Analiza na ovom nivou moze otkriti zakonitosti u interakcijama izmedju korisnika.
- Konverzacioni niz ili deo konverzacionog niza – predstavljaju grupu povezanih i neposrednih poruka u nizu koje pripadaju celokupnoj konverzaciji ili jednom njenom delu. Analiza konverzacionog niza može odgovoriti na pitanja o strukturi, pokrivenosti teme i načinu razvoja konverzacije kroz vreme i drugim faktorima koji na nju mogu da utiču.

Tri navedena nivoa analize konverzacionih tekstova mogu dati različite uvide u karakteristike konverzacije i pomoći u razumevanju različitih aspekata njenog toka.

3.1. Vrste konverzacionih poruka

3.1.1 Poruke elektronske pošte

Tokom poslednje dekade, poruke elektronske pošte su postale jedan od najznačajnijih medija za poslovnu i privatnu komunikaciju. I pored rasta popularnosti socijalnih mreža i alata za brzo komuniciranje, korišćenje alata elektronske pošte za razmenu poruka neprestano raste sa više od 4.4 milijarde korisničkih naloga i prosečnim brojem od 361.6 milijardi razmenjenih poruka po danu zabeleženih u 2024. godini širom sveta [3]. Razlog za ovakvu popularnost poruka elektronske pošte se može pronaći u njihovoj efikasnosti, niskoj ceni i kompatibilnosti sa različitim vrstama informacija koje se pomoću njih mogu proslediti. Jedan od najznačajnijih trendova koji se u statističkim izveštajima može primetiti je da upotreba imejl alata za komuniciranje nastavlja da raste svake godine. Ovakav trend ukazuje i da se potreba za automatskom obradom poruka elektronske pošte sve više povećava.

Programski alati za komunikaciju putem elektronske pošte su značajno unapređivani tokom godina, uvođenjem velikog broja funkcionalnosti koje podržavaju efikasno upravljanje poštanskim sandučetom i neposredniju saradnju sa drugim korisnicima, kao što su predlaganje sadržaja poruke ili automatsko slanje odgovora. Većina ovih alata u ovom trenutku ima funkcionalnosti za bogato formatiranje, sortiranje, prioritizaciju i selekciju poruka, kao i predefinisane obrasce za sastavljanje sadržaja poruka. Ovi alati takođe sadrže funkcionalnosti za lakšu komunikaciju sa drugim korisnicima u realnom vremenu, kao što su funkcionalnosti za slanje brzih poruka, organizovanje video konferencija i deljenje zajedničkih dokumenata. Sve ove mogućnosti obezbeđene su sa ciljem da unaprede, odnosno olakšaju i ubrzaju komunikaciju između korisnika, kako individualnih, tako i grupnih. Ovi alati uključuju sve naprednije funkcionalnosti za enkripciju i višefaktorsku autentifikaciju, razvijene s ciljem očuvanja privatnosti i zaštite poruka od neovlašćenog i zlonamernog pristupa.

Mnogi od alata elektronske pošte korisnicima pružaju funkcionalnost za analizu ličnih podataka, koji im mogu pomoći u razumevanju sopstvenih komunikacionih obrazaca i identifikovanju načina za njegovo poboljšanje. Ova analitička funkcionalnost najčešće sadrži statističke pokazatelje kao što su prosečno vreme odgovora na poruku, broj primljenih i poslatih poruka po kategoriji i vremenskom intervalu, prosečna dužina poruke i trajanje konverzacije u unapred definisanom periodu posmatranja. Svi ovi podaci se mogu koristiti

[3] https://www.radicati.com/?p=18519

za optimizaciju korisnikovih navika, poboljšanje njegove produktivnosti i podešavanje alata za ličnu upotrebu.

Sa povećanjem broja korisnika i korisničkih podataka, metode mašinskog učenja sve više dobijaju na značaju u optimizaciji korišćenja alata elektronske pošte. U naučnim istraživanjima i praktičnim implementacijama, algoritmi mašinskog učenja primenjeni nad porukama elektronske pošte se koriste za:

- Identifikovanje lažne (eng. *spam*) [141, 126] i mamac (eng. *phishing*) [6] poruke u cilju redukovanje broja njihovog pojavljivanja u korisnikovom sandučetu;
- Kategorizaciju poruka u organizacione direktorijume (promocije, društvena povezivanja, obaveštenja o promenama i dr.) [179];
- Prioritizaciju poruka na osnovu njihovog sadržaja [41];
- Analizu sentimenta poruke [7].

Jedan od primera kategorizacije poruka elektronske pošte jeste i njihovo klasifikovanje prema poslovnom ili ličnom kontekstu koji može pomoći u boljem upravljanju poštanskog sandučeta i smanjenju vremena koje korisnici alata za elektronsku poštu svakog dana utroše na tu aktivnost. Nasuprot drugim zadacima klasifikacije, kao što je, na primer, klasifikovanje lažnih poruka, ovom problemu do sada nije posvećeno dovoljno pažnje i još uvek predstavlja nedovoljno istražen zadatak. Jedan od razloga za to jeste i nedostatak odgovarajućih podataka - lični mejlovi su obično zaštićeni zakonima o zaštiti privatnosti, te su stoga nedostupni i za potrebe istraživanja. Za potrebe analize konverzacionih poruka elektronske pošte u toku ovog istraživanja, korišćen je korpus *Enron* [102] - jedini javno dostupan korpus konverzacija elektronske pošte, objavljen pod licencom *Creative Commons Attribution 3.0 United States*[4]. Iz razloga lake dostupnosti, korpus *Enron* korišćen je u okviru mnogobrojnih istraživanja. Prvi pokušaji automatske kategorizacije korporativnih poruka elektronske pošte na poslovnu i ličnu kategoriju sprovedeni su u radu [89], u kojem su autori postavili temelje za razvoj metoda koje omogućavaju efikasnu klasifikaciju elektronske komunikacije u zavisnosti od njenog sadržaja. Glavni doprinos ovih istraživanja je projekat označavanja jednog dela korpusa *Enron* od približno 12,500 poruka koje su klasifikovane u poslovnu ili ličnu kategoriju. U radu je korišćen klasifikator zasnovan na distribuciji reči, kako bi utvrdila izvodljivost razdvajanja poslovnih i ličnih poruka pomoću mašina. U [8] i [9], autori su obučili svoje modele nad obeleženim porukama korpusa *Enron* i testirali ih nad porukama iz korpusa *Enron* i *Avocado*. U [8], autori su kombinovali karakteristike grafovskih mreža sa unapred obučenim ugnježdenim vektorima *GloVe*[5] kao leksičkih karakteristika iz sadržaja poruke kako bi poboljšali performanse postojećih klasifikatora. Kao dodatni doprinos ovom radu, autori su takođe ručno obeležili skup *Enron* poruka. U [9], isti autori su dodatno razmotrili strukturu nizova poruka što je unapredilo performanse klasifikacije u odnosu na prethodne pristupe.

Pored zadataka klasifikacije, nad porukama elektronske pošte se primenjuju i druge metode mašinskog učenja u cilju poboljšanja efikasnosti korišćenja alata za razmenjivanje poruka. Neki od najznačajnijih predloženih naučnih metoda su sledeće:

- Automatsko generisanje personalizovanog odgovora [199];
- Dopunjavanje rečenice i pronalaženje pravopisnih grešaka [38];
- Sumarizacija pojedinačne poruke ili celog konverzacionog niza [222];

[4] https://enrondata.org/
[5] https://nlp.stanford.edu/projects/glove/

- Kreiranje liste zadataka na osnovu sadržaja poruke [140];
- Automatsko generisanje sadržaja naslova [221].

3.1.2 Poruke na društvenim mrežama

Sa pojavom društvenih mreža poslednjih godina, objave (eng. *post*) i komentari (eng. *comments*) na društvenim mrežama su postali jedna od najpopularnijih vrsta pisanih konverzacionih tekstova koja podrazumeva komunikaciju između dve ili više osoba. Ove vrste konverzacionih tekstova imaju svoje karakteristike kao što su postavljanje pitanja i odgovaranje na njih, izražavanje mišljenja i razmena ličnih stavova. Društvene mreže dodatno podstiču konverzacione interakcije putem svojih funkcionalnosti kao što su izražavanje podrške (eng. *like*), deljenje objave, komentarisanje i odgovaranje na komentare, označavanje korisnika, korišćenje emotikona i drugih vizuelnih elemenata za izražavanje emocija i podsticanje reakcija. Sve ove funkcionalnosti napravljene su sa namerom da oponašaju glavne karakteristike tradicionalnog načina komuniciranja.

Međutim, i pored niza sličnosti sa drugim medijima za komunikaciju i konverzacionim tekstovima koji na njima nastaju, konverzacioni tekstovi na društvenim mrežama imaju niz svojih karakteristika. Ove poruke su obično kratke dužine (eng. *short messages*) i pisane su neformalnim jezikom izražavanja. Putem poruka se mogu proslediti različite vrste sadržaja koji mogu uključivati tekst, sliku, audio i video format. Korisnici mogu sakriti svoj pravi identitet korišćenjem pseudonima, što podstiče lakše izražavanje mišljenja. Dodatno, za konverzacije na društvenim mrežama je karakterističan asinhroni način komunikacije koji može uticati na razvoj konverzacionog toka. Sve ove karakteristike je važno uzeti u obzir prilikom analize ovih vrsta konverzacionih tekstova.

Istraživanje fenomena ljudskih interakcija na društvenim mrežama je postalo veoma aktuelno u naučnim krugovima. Polja istraživanja obuhvataju komunikaciju, politička delovanja, sociološke i psihološke uticaje, načine povezivanja i analizu ponašanja korisnika u grupama. Kao programski sistemi koji su okrenuti ka komunikaciji između korisnika, podaci prikupljeni sa društvenih mreža Tviter i Redit zauzimaju posebno mesto u računarskoj lingvistici [11, 161]. Pored standardnih funkcionalnosti kao što su mogućnost deljenja višemodalnog sadržaja (tekst, slika, audio i video zapis), komentarisanja, podrške i označavanja, ove mreže imaju i neke svoje specifičnosti koje mogu uticati na izbor teme i rezultata istraživanja koja se nad njima izvode:

- Tviter
  - Poseduje ograničenja u dužini objave koji je inicijalno postavljeno na 140, a kasnije produženo na 280 karaktera;
  - Poseduje funkcionalnost za prikupljanje korisničkih reakcija na objavu sa karakteristikama za merenje sviđanja (eng. *favourite*, *like*), deljenja (eng. *retweet*) i deljenja sa pridruženim komentarom (eng. *quote*);
  - Usmeren je na tekuće vesti i obaveštenja.
- Redit
  - Objave su grupisane u Subredit (eng. *Subreddit*) zajednice koje su okrenute određenoj temi;
  - Poseduje složen sistem za glasanje sa funkcionalnostima za davanje podrške i neslaganje;

– Korisnici češće koriste pseudonime, prikrivajući na taj način svoj pravi identitet;

– Okrenut je postavljenim temama bez obzira na njihovu aktuelnost.

Konverzacioni podaci sa društvene mreže Tviter se koriste u analizi sentimenta [227], kategorizaciji tema [46], utvrđivanju načina delovanja [171], postojanja govora mržnje [16] i tona prosleđene poruke [40].

3.1.3 Ostale vrste konverzacionih poruka

Sa razvojem interneta, pojavljuju se i alati za razmenu brzih poruka (eng. *instant messaging*) u realnom vremenu. Ovi alati pružaju mogućnost brze i efikasne neformalne komunikacije sa drugim korisnicima korišćenjem tekstualnih poruka, glasovnih i video poziva, deljenja dokumenata i grupnih razgovora. Glavna karakteristika ovakvih konverzacija je mogućnost komuniciranja u realnom vremenu i dostupnost za korišćenje na različitim platformama (telefonima, tabletima i računarima). Ove vrste poruka možemo svrstati u sledeće kategorije:

- Direktne poruke na platformama društvenih mreža (*Facebook Messenger*, *Twitter Direct Messages*, *Instagram Chat*);
- Poruke u alatima specijalizovanim za direktno komuniciranje (*WhatsApp*, *Viber*, *Teams*, *Slack*);
- Poruke za direktnu komunikaciju sa korisnicima na komercijalnim platformama (eng. *live chat*);
- Poruke u alatima za automatsko generisanje odgovora (čatbot, eng. *chat bot*);
- Poruke kao komentari u deljenim dokumentima.

Posebna vrsta konverzacionih tekstova predstavljaju tekstovi nastali u alatima za automatsko generisanje odgovora. Još od nastanka računara, ljudi su pokušavali da komuniciraju sa njima. Prvi takvi pokušaji bili su putem razvoja nižih i viših programskih jezika pomoću kojih se računaru daju instrukcije šta bi trebalo da uradi. Drugi načini komuniciranja sa računarima jesu putem davanja instrukcija operativnom sistemu korišćenjem sistema komandi (eng. *command line interface - CLI*) ili njegovih vizuelnih alata (eng. *graphical user interface - GUI*). Dalja nastojanja išla su u pravcu razvoja alata koji bi ljudima omogućili da sa mašinama komuniciraju na jeziku ljudi. Jedan od prvih takvih pokušaja je bio razvoj ELIZA programa koji je, korišćenjem tehnika za podudaranje i zamenu obrazaca u obradi prirodnih jezika, simulirao razgovor psihijatra sa korisnicima [212]. Alati za automatsko generisanje odgovora i konverzacioni tekstovi koji iz njih nastaju su projektovani tako da imitiraju tradicionalan dijalog između ljudi. Prilikom razvoja ovih alata glavni cilj je da komunikacija bude razumljiva, relevantna, efikasna i pouzdana za krajnjeg korisnika. Automatsko generisanje odgovora u ovim alatima se zasniva na nekoj od narednih tehnika i metoda:

- Skupovi pravila – koristi predefinisana pravila i obrasce za kreiranje odgovora.
- Predefinisani obrasci – koristi unapred pripremljene obrasce odgovora koji se popunjavaju relevantnim podacima u zavisnosti od korisnikovog zahteva.
- Pronalaženje informacija – zaniva se na poređenju korisnikovog zahteva sa predefinisanim odgovorima koji se nalaze u bazi podataka.

- Generativne metode – koristi algoritme mašinskog učenja i generativne jezičke modele za kreiranje odgovora.

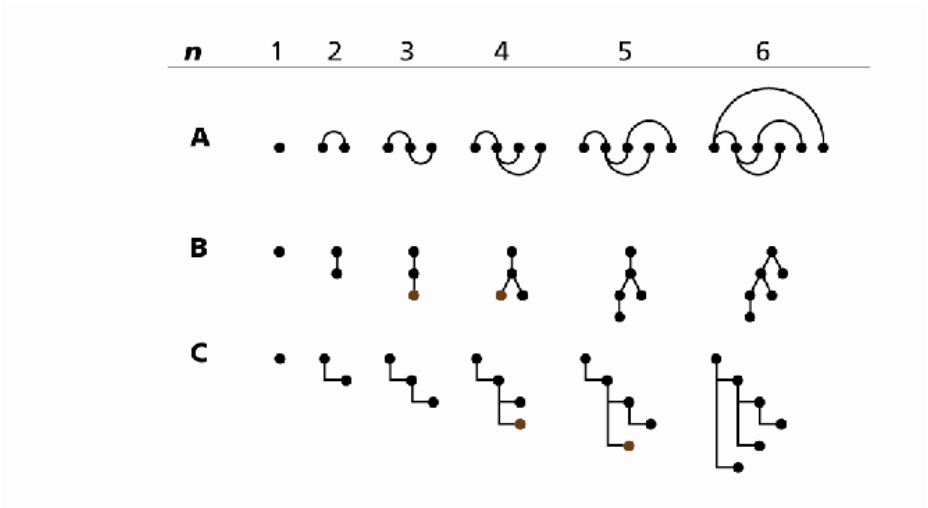
Od navedenih tehnika, skupovi pravila i predefinisani obrasci mogu biti pogodni za jednostavnije alate sa ograničenim skupom mogućih odgovora, dok su pronalaženje informacija i generativna metoda pogodnije za alate kojima je neophodan složeniji sistem odgovora prema krajnjim korisnicima. Najnovija istraživanja o čatbotovima pokazuju njihove različite primene u više oblasti, ističući dobrobiti i izazove koje donose. Jedna studija ispitivala je značaj čatbot alata u učenju stranog jezika i pokazala da oni poboljšavaju učenje unapređivanjem veština, motivacije i samopouzdanja učenika, uprkos uočenim ograničenjima u korisničkom iskustvu [42]. U korisničkoj podršci, čatbot alati zasnovani na mašinskom učenju značajno poboljšavaju efikasnost i zadovoljstvo korisnika, dok istovremeno smanjuju troškove kompanijama [3]. Generisanje potpomognuto pretragom (eng. *Retrieval-Augmented Generation, RAG*) je napredna metoda u razvoju čatbotova za pronalaženje relevantnih informacija koja kombinuje pretragu dokumenata indeksiranih u vektorskim bazama podataka sa odgovorima koje generiše LLM, čime se poboljšava tačnost i pouzdanost odgovora [162]. U zdravstvenom sektoru, Query-Based RAG (QB-RAG) sistem efikasno usklađuje korisničke upite sa unapred indeksiranim upitima iz baze podataka, čime se poboljšava tačnost saveta vezanih za zdravlje i smanjuje rizik od netačnih informacija [217]. Pored toga, RAG zasnovani čatbotovi zahtevaju pažljivo projektovanje kako bi se postigao balans između tačnosti, brzine i sigurnosti podataka [5]. Pokazano je da u obrazovanju RAG čatbot može da poboljša angažovanost studenata i pripremljenost za ispite, sa 97.1% studenata koji su prijavili pozitivna iskustva sa korišćenjem čatbota za tu namenu [200].

3.2. Vizuelizacija konverzacionih tekstova

Vizualizacija konverzacionih tekstova dodatno može pomoći u boljem razumevanju konverzacionog toka grafičkim prikazivanjem zakonitosti, trendova i odnosa između korisnika ili poruka u okviru konverzacije. U zavisnosti od karakteristika konverzacija i zadatka istraživanja, za vizuelizaciju konverzacionih tekstova može se koristiti neka od sledećih tehnika:

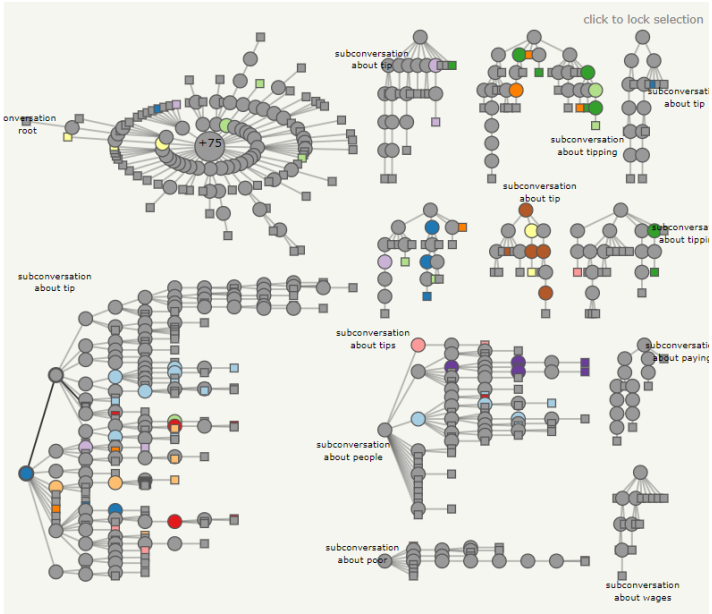
- Dijagrami stabala (eng. *tree diagrams*) - predstavljaju konverzaciju kao hijerarhijsku strukturu u obliku stabla u kojem svaki čvor predstavlja poruku ili korisnika u konverzaciji.
- Mreže grafova (eng. *network graphs*) - predstavljaju konverzaciju kao povezani graf sa čvorovima i granama u kome svaki čvor predstavlja korisnika, a grana poruku u konverzaciji.
- Toplotne mape (eng. *heatmaps*) - ova vrsta vizuelizacije je pogodna za predstavljanje frekvencije ili intenziteta delova konverzacionog toka korišćenjem različitih nijansi boja. Svaka ćelija na mapi predstavlja korisnika ili poruku, dok boja označava intenzitet tog dela konverzacije.
- Dijagrami rasejanja (eng. *scatter plots*) - predstavljaju konverzaciju kao skup tačaka u koordinatnom prostoru u kojem svaka tačka predstavlja poruku u konverzaciji. Na koordinatnim osama se predstavljaju različite karakteristike poruke.

Jedan od prvih predloženih pristupa vizuelizaciji konverzacija poruka elektronske pošte jeste pomoću nizova lukova, dijagrama stabala i tabela stabala (slika 3.2). Najveći



Slika 3.2: *Prikaz razvoja konverzacionog toka korišćenjem tri različite tehnike za vizuelizaciju konverzacionog toka - Nizovi lukova (A), Dijagrami stabala (B) i Tabele stabala (C) [95]*

broj naučnih radova opisuje razvoj alata za vizuelnu analitiku koji imaju funkcionalnost za rekonstrukciju konverzacionog toka iz nestruktuiranih tekstualnih podataka [52]. Neki od predloženih alata kombinuju NLP tehnike, kao što su modelovanje tema i analiza sentimenta, sa tehnikama vizualizacije grafovskih struktura podataka kako bi omogućili bogatiji prikaz strukture konverzacije sa dodatnim skupom atributa [83]. Na slici 3.3 je vizuelno predstavljena jedna konverzacija sa foruma *HackerNews* i pojedini njeni delovi, napravljena pomoću *Forum Explorer*[6] alata, koji poboljšava skalabilnost vizualizacije za konverzacije sa velikim brojem poruka i grana, omogućava obradu podataka u realnom vremenu i olakšava navigaciju kroz složene konverzacije na forumskim platformama [128]. Vizuelizacija predstavlja jedan od načina za analizu i razumevanje konverzacionih podataka kojim se mogu dobiti uvidi u dinamiku konverzacije, pokrivenost tema i pravilnosti koje postoje u konverzaciji i koje mogu uputiti na dalji tok istraživanja [98].



Slika 3.3: *Vizuelni prikaz konverzacija sa internet foruma u obliku mreže grafova*

[6] https://www.mcnutt.in/forum-explorer/

4. Klasifikacija kao problem mašinskog učenja

4.1. Opšte o klasifikaciji

Klasifikacija je proces pridruživanja podataka u unapred definisan skup diskretnih klasa. U mašinskom učenju, klasifikacija pripada prediktivnim metodama, odnosno metodama sa nadgledanim učenjem (eng. *supervised learning*). Ova vrsta učenja podrazumeva postojanje označenih podataka (eng. *labeled datasets*) u kojima je svakoj instanci u podacima pridružena odgovarajuća klasa kojoj ta instanca pripada. Pored pridružene klase, svaka instanca sadrži pridružene vrednosti za skup atributa koji će se koristiti kao ulazne vrednosti u klasifikaciji. Ovi vektori atributa su unapred zadati ili se, u slučaju tekstualnih podataka, korišćenjem određenog preslikavanja mogu ekstrahovati iz teksta. Proces klasifikacije podataka na najvišem nivou deli se na dve osnovne faze:

- **Faza obučavanja (učenja)** – obuhvata konstruisanje prediktivnog modela na osnovu dostupnog skupa obeleženih podataka, koji uključuje optimizaciju parametara i podešavanje hiperparametara modela.
- **Faza testiranja** – obuhvata procenu tačnosti i generalizacione sposobnosti napravljenog modela.

U fazi obučavanja modela, primenom ML algoritma nad skupom za obuku (T), pronalazi se funkcija f koja ulazne atribute preslikava u njima pridružene klase $C_i, k = 2, ...n$. Kao što je prikazano u jednačini 4.1, korišćenjem preslikavanja ϕ, svakoj tekstualnoj sekvenci $T_k, k = 2, ...m$ je pridružen vektor atributa koji predstavlja jednu tačku u d-dimenzionalnom prostoru atributa:

$$\phi(T_k) = (\phi_1(T_k), \dots \phi_d(T_k)), \phi(T_k) \in R^d, k = 1, 2, ..., m \tag{4.1}$$

Dodatno, vektor parametara koji određuje koliko svaki vektor atributa doprinosi u predviđanju klase je predstavljen u jednačini 4.2:

$$P = (P_1, P_2, \dots P_d), P \in R^d \tag{4.2}$$

Spajanjem $\phi(T)$ i P, kao što je prikazano u jednačini 4.3, dobijamo matematički izraz za traženu funkciju f:

$$f_P(T_k) = \phi(T_k) * P, k = 1, 2, ..., m \tag{4.3}$$

Da bi se obezbedila što objektivnija i nepristrasnija procena kvaliteta modela, skup obeleženih podataka neophodno je podeliti na tri međusobno nezavisna dela:

- **Skup za obučavanje** (eng. *training set*) – koristi se za učenje modela, odnosno za prilagođavanje parametara modela podacima.
- **Skup za proveru** (eng. *validation set*) – koristi se za podešavanje hiperparametara modela i praćenje njegove sposobnosti generalizacije tokom procesa učenja, sa ciljem sprečavanja prekomernog prilagođavanja modela (eng. *overfitting*).

- **Testni skup** (eng. *test set*) – koristi se za konačnu proveru performansi modela, pri čemu je ovaj skup potpuno nezavistan (eng. *out of sample*) od skupova za obučavanje i proveru.

U slučajevima kada je ukupna količina označenih podataka ograničena i kada bi statička podela na fiksne skupove dovela do nestabilne procene performansi, primenjuje se tehnika unakrsne validacije (eng. *Cross-Validation, CV*). Ova tehnika podrazumeva višestruko deljenje skupa podataka na podskupove za obučavanje i validaciju, pri čemu se postupak treniranja i evaluacije ponavlja u unapred definisanom broju iteracija (eng. *folds*), a konačna ocena performansi dobija se agregiranjem rezultata kroz sve iteracije [107]. Često korišćena varijanta CV tehnike je stratifikovana unakrsna validacija (eng. *stratified cross-validation*), kod koje se prilikom podele vodi računa o očuvanju proporcija klasa u svakom podskupu, dok se u situacijama sa veoma malim skupovima podataka koristi unakrsna validacija izostavljanjem jednog primera (eng. *leave-one-out*), u kojoj se u svakoj iteraciji za proveru izdvaja tačno jedan primer [13].

4.1.1 Vrste klasifikacije

U zavisnosti od broja klasa, postoje sledeće vrste klasifikacije:

- Binarna – kada su definisane samo dve klase ($n = 2$).
- Višeklasna – kada je definisano više od dve klase ($n > 2$).

U zavisnosti od prirode zadatka klasifikacije, ona može biti:

- Jednoznačna (eng. *single-label*) – svakoj instanci podataka je dodeljena tačno jedna klasa.
- Višeznačna (eng. *multi-label*) – instanci podataka može biti dodeljeno više od jedne klase.

4.1.2 Merenje tačnosti klasifikacije

Uspešnost klasifikacije podataka se izračunava korišćenjem različitih skupova mera koje odgovaraju različitim vrstama klasifikacionih algoritama. Jedan od osnovnih načina predstavljanja i merenja uspešnosti jeste korišćenjem matrice konfuzije koja se može primeniti na binarnu ili višeklasnu klasifikaciju (pogledati tabelu 4.1). Elementi matrice konfuzije imaju sledeća značenja:

- Tačna pozitivna (eng. *True Positive, TP*) – broj pozitivnih instanci koje su ispravno predviđene kao pozitivne.
- Tačna negativna (eng. *True Negative, TN*) – broj negativnih instanci koje su ispravno predviđene kao negativne.
- Pogrešna pozitivna (eng. *False Positive, FP*) – broj negativnih instanci koje su pogrešno predviđene kao pozitivne.
- Pogrešna negativna (eng. *False Negative, FN*) – broj pozitivnih instanci koje su pogrešno predviđene kao negativne.

Na osnovu ovih vrednosti možemo izračunati sledeće mere koje se koriste za evaluaciju modela:

Tabela 4.1: *Matrica konfuzije u binarnoj klasifikaciji*

		Predviđena	
		Pozitivna	Negativna
Stvarna	Pozitivna	TP	FN
	Negativna	FP	TN

- Preciznost (eng. *Precision*, *Prec*) – meri odnos tačno predviđenih pozitivnih instanci i svih instanci koje su predviđene kao pozitivne:

$$Prec = \frac{TP}{TP + FP} \tag{4.4}$$

- Odziv (eng. *Recall*, *Rec*) – meri odnos tačno predviđenih pozitivnih instanci i svih instanci koje su stvarno pozitivne:

$$Rec = \frac{TP}{TP + FN} \tag{4.5}$$

- F_1-mera (eng. F_1-*measure*, F_1) – jeste harmonijska sredina preciznosti i odziva:

$$F_1 = 2 * \frac{Prec * Rec}{Prec + Rec} \tag{4.6}$$

  i predstavlja specijalan slučaj F_β mere za $\beta = 1$ koja je data sledećom jednačinom:

$$F_\beta = (1 + \beta^2) * \frac{Prec * Rec}{(\beta^2 * Prec) + Rec} \tag{4.7}$$

- Tačnost (eng. *Accuracy*, *Acc*) – meri udeo tačno klasifikovanih instanci u ukupnom broju instanci:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.8}$$

- Specifičnost (eng. *Specificity*, *S*):

$$S = \frac{TN}{TN + FP} \tag{4.9}$$

U odnosu na karakteristike klasa koje se koriste u klasifikaciji, što je od naročitog značaja kod višeklasne i višeznačne klasifikacije, najčešće korišćene Prec, Rec i F_1 mere se mogu izračunati kao:

- Makro prosek – mere se izračunavaju za svaku klasu pojedinačno, a zatim se izračunava srednja vrednost dobijenih mera po klasama. U tom kontekstu, uvodimo oznake:
  - Makro preciznost (eng. *Macro Precision*, $Prec^{Ma}$);
  - Makro odziv (eng. *Macro Recall*, Rec^{Ma});
  - Makro F_1-mera (eng. *Macro* F_1-*measure*, F_1^{Ma}).

- Mikro prosek – TP, FP, TN i FN se računaju za sve instance, a zatim se na osnovu njih izračunavaju konačne mere prema datim formulama (pogledati jednačine 4.4, 4.5 i 4.6). Odgovarajuće oznake u ovom kontekstu su:
  - Mikro preciznost (eng. *Micro Precision*, $Prec^{Ma}$);
  - Mikro odziv (eng. *Micro Recall*, Rec^{Mi});
  - Mikro F_1-mera (eng. *Micro F_1-measure*, F_1^{Mi}).

Izbor odgovarajuće mere za evaluaciju klasifikacionog modela zavisi od prirode problema i ciljeva analize. Acc je pogodna kada su podaci uravnoteženi i sve greške podjednako važne. Prec je ključna u situacijama gde je minimizacija lažno pozitivnih predviđanja od presudnog značaja, kao što je detekcija prevara. S druge strane, Rec postaje prioritet u problemima gde je važno otkriti sve pozitivne primere, čak i po cenu većeg broja lažno pozitivnih, kao u medicinskoj dijagnostici (pogledati tabelu 4.2).

Tabela 4.2: *Poređenje karakteristika Prec i Rec evaluacionih mera*

Osobina	Prec	Rec
Fokus	Koliko su tačne pozitivno predviđene instance	Koliko pozitivnih instanci je pravilno otkriveno
Rizik	Visok udeo pogrešno pozitivnih instanci (FP)	Visok udeo pogrešno negativnih instanci (FN)
Kada je važan?	Kada su pogrešno pozitivne instance skupe ili štetne	Kada je važno ne propustiti nijednu pozitivnu instancu

Kada je potrebno postići balans između Prec i Rec, posebno na nebalansiranim podacima, F_1 mera pruža optimalnu evaluaciju. Kod nebalansiranih skupova podataka, odnosno podataka kod kojih je učestalost pojavljivanja jedne klase zastupljenija od drugih, potrebno je koristiti mere koje će na adekvatan način prikazati tačnost klasifikacije i prema manje zastupljenim klasama. Za tu namenu najčešće se koriste:

- Balansirana tačnost (eng. *Balanced Accuracy*, Acc_{Bal}) – računa prosečnu vrednost između odnosa tačno klasifikovanih negativnih instanci (S) i odnosa tačno klasifikovanih pozitivnih instanci (Rec):

$$Acc_{Bal} = \frac{S + Rec}{2} \tag{4.10}$$

- F_1 na manje zastupljenoj klasi.
- Težinska F_1-mera (eng. *Weighted* F_1*-measure*, F_1^w) – uzima u obzir F_1 mere za sve klase (F_1^i), pri čemu im dodeljuje težine na osnovu relativne frekvencije svake klase u skupu podataka (w_i):

$$F_1^w = \frac{\sum_{i=1}^{C} w_i * F_1^i}{\sum_{i=1}^{C} w_i} \tag{4.11}$$

- Računanjem makro proseka – čime se svim klasama dodeljuje podjednaka važnost.

U slučaju višeznačne klasifikacije, kada se jednoj instanci dodeljuje više klasa, mere bi trebalo da obuhvate slučajeve kada su predviđanja potpuno tačna, delimično tačna ili potpuno netačna. U tom cilju razvijen je čitav skup mera za merenje kvaliteta klasifikacije podataka u višeznačnom kontekstu [183]. Za višeznačni klasifikator $Z_i = f(x_i) \in \{0,1\}^k$ i

višeznačni skup od n označenih instanci u $k = |L|$ različitih oznaka $(x_i, Y_i), 1 \leq i \leq n, (x_i \in X, Y_i \in Y = \{0,1\}^k)$, definisane su dve osnovne grupe mera za merenje kvaliteta višeznačne klasifikacije:

- **Mere na nivou instanci** – jeste pristup u kome se računa prosečna razlika između predviđenih i tačnih klasa na nivou svake instance, a zatim se računa prosek na nivou svih instanci u testnom skupu podataka. Ovoj grupi mera pripadaju:
  - Egzaktna tačnost (eng. *Exact Match Ratio, EMR*):
$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Z_i) \tag{4.12}$$
  - Tačnost:
$$Acc = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{4.13}$$
  - Preciznost:
$$Prec = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Z_i|} \tag{4.14}$$
  - Odziv:
$$Rec = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Y_i|} \tag{4.15}$$
  - F_1 mera:
$$F_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|} \tag{4.16}$$
  - Hamingov gubitak (eng. *Hamming Loss, HL*):
$$HL = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L} I\left(Y_i^j \neq Z_i^j\right) \tag{4.17}$$
  - Hamingov skor (eng. *Hamming Score, HS*):
$$HS = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L} I\left(Y_i^j = Z_i^j\right) \tag{4.18}$$

- **Mere na nivou klasa:**
  - **Makro prosek** – mere se izračunavaju za svaku klasu, a zatim se računa srednja vrednost, kao što je prikazano sledećim jednačinama:
$$Prec_\lambda^{Ma} = \frac{\sum_{i=1}^{n} Y_i^\lambda Z_i^\lambda}{\sum_{i=1}^{n} Z_i^\lambda}, Prec^{Ma} = \frac{1}{k}\sum_{i=1}^{k} P_\lambda^{Ma} \tag{4.19}$$
$$Rec_\lambda^{Ma} = \frac{\sum_{i=1}^{n} Y_i^\lambda Z_i^\lambda}{\sum_{i=1}^{n} Y_i^\lambda}, Rec^{Ma} = \frac{1}{k}\sum_{i=1}^{k} R_\lambda^{Ma} \tag{4.20}$$
$$F_{1,\lambda}^{Ma} = \frac{2\sum_{i=1}^{n} Y_i^\lambda Z_i^\lambda}{\sum_{i=1}^{n} Y_i^\lambda + \sum_{i=1}^{n} Z_i^\lambda}, F_1^{Ma} = \frac{1}{k}\sum_{i=1}^{k} F_{1,\lambda}^{Ma} \tag{4.21}$$

– **Mikro prosek** – mere se izračunavaju za sve klase istovremeno na celom skupu podataka:

$$Prec^{Mi} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^j Z_i^j}{\sum_{j=1}^{k}\sum_{i=1}^{n} Z_i^j} \tag{4.22}$$

$$Rec^{Mi} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^j Z_i^j}{\sum_{j=1}^{k}\sum_{i=1}^{n} Z_i^j} \tag{4.23}$$

$$F_1^{Mi} = \frac{2\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^j Z_i^j}{\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^j + \sum_{j=1}^{k}\sum_{i=1}^{n} Z_i^j} \tag{4.24}$$

gde su Y_i^λ i Z_i^λ definisane na sledeći način:

$$Y_i^\lambda = \begin{cases} 1 & ukoliko\ \ x_i\ pripada\ klasi\ \lambda, \\ 0 & u\ suprotnom \end{cases} \tag{4.25}$$

$$Z_i^\lambda = \begin{cases} 1 & ukoliko\ se\ predvia\ da\ x_i\ pripada\ klasi\ \lambda, \\ 0 & u\ suprotnom \end{cases} \tag{4.26}$$

Ukoliko je ispunjeno da je $k = 1$ zadatak se svodi na višeklasnu, a ukoliko je dodatno i $n = 2$ na binarnu klasifikaciju.

4.1.3 Pronalaženje značajnih i korelisanih atributa

Jedan od ključnih zadataka u mašinskom učenju jeste određivanje ulaznih atributa koji najviše doprinose tačnosti klasifikacije podataka. Identifikacija značajnih atributa iz celokupnog skupa dostupnih atributa je od suštinskog značaja za razumevanje zadatka i dostupnih podataka, kao i za poboljšanje performansi i interpretaciju izgrađenog modela. Pristupi za utvrđivanje značaja atributa mogu se podeliti na:

- **Pristup jedne promenljive** (eng. *univariate*) – u ovom pristupu značaj svakog atributa se procenjuje pojedinačno, nezavisno od drugih atributa. Prednost ovog pristupa jeste što su njegove tehnike obično jednostavne i intuitivne, ali je njegov glavni nedostatak zanemarivanje mogućih korelacija koje mogu postojati među ulaznim atributima [37]. Neke od tehnika ovog pristupa su:
  - **F-statistika (eng. *F-statistic*, F_s)** – statistička mera zasnovana na analizi varijanse (eng. *Analysis of Variance*, *ANOVA*), koja se definiše se kao odnos varijabilnosti srednjih vrednosti atributa između (eng. *Mean Square Between*, *MSB*) i unutar (eng. *Mean Square Within*, *MSW*) klasa. Ova mera se računa na osnovu sledećih jednačina:

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{k-1} \tag{4.27}$$

$$MSW = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{N-k} \tag{4.28}$$

$$F_s = \frac{MSB}{MSW} \tag{4.29}$$

pri čemu je:

* X_{ij} vrednost atributa za instancu j u klasi i,
* $\bar{X}_i$ srednja vrednost atributa u klasi i,
* $\bar{X}$ srednja vrednost atributa u celokupnom skupu podataka,
* n_i broj instanci koje pripadaju klasi i,
* N ukupan broj instanci u skupu podataka,
* k broj različitih klasa u skupu podataka.

– **Maksimalni koeficijent informativnosti** (eng. *Maximal Information Coefficient*, *MIC*) – mera koja se koristi za kvantifikaciju statističke zavisnosti između dve promenljive [167].

– **Vrednost prediktivne moći atributa** – mera koja uključuje izgradnju modela sa posmatranim ulaznim atributom i proveru njegove sposobnosti predviđanja u takvom scenariju, čime se mogu identifikovati potencijalne nelinearne zavisnosti između ulaznih i izlaznih promenljivih.

- **Pristup više promenljivih** (eng. *multivariate*) – u ovom pristupu značaj atributa se procenjuje uzimajući u obzir prisustvo i zavisnost sa drugim atributima u skupu podataka. Razvijene tehnike utvrđivanja značajnosti atributa koje pripadaju ovom pristupu mogu biti:

  – **Zavisne od algoritma** – značaj atributa se utvđuje u toku obučavanja modela na osnovu pravilnosti koje su uočene u toku obuke. Prednost ovog pristupa je da se značajni atributi mogu efektivno pronaći korišćenjem ugrađenih funkcija u implementacijama nekih algoritama. Nedostaci uključuju zavisnost od algoritma i skupa podataka koji se koristi za obučavanje modela. U algoritmima zasnovanim na stablima odlučivanja, značajni atributi se mogu pronaći među onima koji redukuju Gini indeks ili entropiju, pojavljuju se u većem broju grananja ili prema broju instanci koje su obuhvaćene testiranjem pomoću posmatranog atributa. Kod linearnih modela značajni atributi mogu pronaći među onima sa većim vrednostima težinskih koeficijenata i manjim p-vrednostima statističke mere.

  – **Nezavisne od algoritma** – značaj atributa se utvrđuje testiranjem izgrađenog modela nad novim i nezavisnim skupom podataka nakon obučavanja modela. Prednost ovog pristupa je nezavisnost od modela i skupa podataka nad kojim je model izgrađen, što ih čini primenjivim nad velikim brojem algoritama [168]. Međutim, metode iz ovog pristupa mogu biti spore za izvršavanje jer mogu zahtevati veliki broj iteracija ili primenu složenih algoritama za pronalaženje značajnih atributa. Neke od najpoznatijih tehnika ovog pristupa su:

    * **Permutacija vrednosti atributa** – u kojoj se vrednosti jednog atributa se nasumično permutuju. Smanjenje performansi modela nakon primenjene permutacije ukazuje na stepen važnosti datog atributa u procesu izgradnje modela.
    * **Isključivanje atributa** – podrazumeva potpuno isključivanje jednog atributa iz skupa ulaznih atributa prilikom izgradnje modela. Značaj atributa se procenjuje na osnovu stepena promene u performansama modela, odnosno merenjem razlike u performansama pre i nakon isključivanja tog atributa [57].
    * **SHAP** (eng. *SHapley Additive exPlanations*) algoritam – procenjuje prosečan doprinos svakog atributa u svim mogućim permutacijama dostupnih atribu-

ta. SHAP omogućava preciznu interpretaciju modela, dajući uvid u to kako svaki atribut doprinosi konačnoj odluci modela [125].

Različite metode selekcije atributa ne daju uvek isti skup značajnih atributa, jer njihovi rezultati zavise od prirode podataka, vrste modela, koraka predprocesiranja i specifičnih karakteristika svake metode. Da bi se odabrao najbolji skup atributa, preporučuje se korišćenje CV metode za evaluaciju performansi modela sa različitim skupovima atributa, kombinovanje rezultata različitih metoda selekcije, rangiranje atributa prema njihovom značaju u različitim metodama, kao i konsultovanje domenskih eksperata za zadatak koji se rešava. Takođe je važno proceniti stabilnost skupa značajnih atributa pod različitim uzorcima podataka ili različitim podešavanjima hiperparametara modela, u cilju utvrđivanja njihove pouzdanosti.

Osim toga, statističke analize mogu se koristiti za identifikaciju korelacija između atributa i merenje jačine njihovih asocijacija. Jačina asocijacije, izražena koeficijentom korelacije, može imati vrednosti između -1 i +1. Vrednosti na krajevima ovog numeričkog ranga (± 1) označavaju savršen stepen korelacije, dok vrednosti bliže 0 označavaju slabiji stepen korelacije. Znak u vrednosti koeficijenta korelacije označava usmerenje korelacije (pozitivno, negativno). U ovom radu će biti korišćene sledeće statističke mere za merenje stepena korelacije između atributa:

- **Pirsonov koeficijent korelacije (eng. *Pearson correlation coefficient*, r)** – statistička mera za utvrđivanje linearnih zavisnosti između numeričkih atributa (pogledati jednačinu 4.30). Ovaj pristup zahteva da vrednosti atributa zadovoljavaju uslov normalne raspodele i pretpostavlja postojanje linearne zavisnosti između atributa.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4.30}$$

  gde su X_i i Y_i vrednosti atributa X i Y respektivno za instancu i, $\bar{X}$ i $\bar{Y}$ su srednje vrednosti ovih atributa, a n je broj instanci.

- **Spirmanov koeficijent korelacije (eng. *Spearman correlation coefficient*, ρ)** – neparametarska statistička mera za utvrđivanje monotone zavisnosti između numeričkih atributa (pogledati jednačinu 4.31):

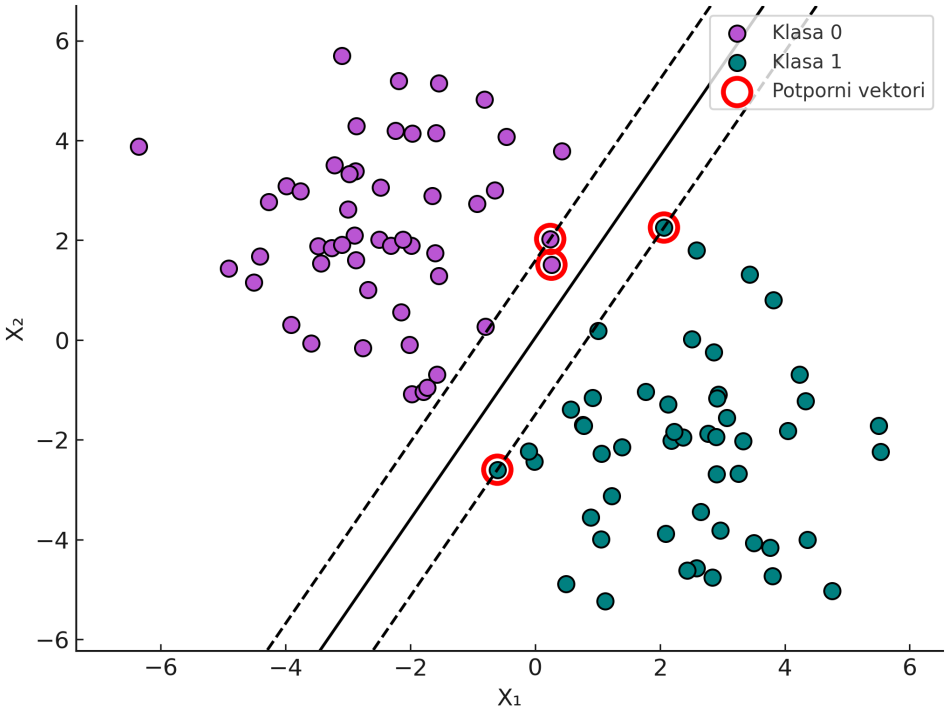
$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{4.31}$$

  gde je d_i razlika između redosleda za svaki par instanci, a n broj instanci. Ovaj pristup zahteva da vrednosti atributa zadovoljavaju uslov uređenosti (redosleda).

4.2. Tradicionalni prediktivni algoritmi mašinskog učenja

Tradicionalni algoritmi mašinskog učenja predstavljaju osnovu razvoja savremenih algoritama i zasnivaju se na statističkim principima za analizu podataka. Među najpoznatijim prediktivnim algoritmima su logistička regresija (eng. *Logistic Regression, LR*), stabla odlučivanja ili naivni Bajesov algoritam. Ovi algoritmi se oslanjaju na ručno i pažljivo odabrane nezavisne atribute (eng. *feature engineering*) koje omogućavaju modelu da precizno analizira podatke. Tradicionalni prediktivni algoritmi kao što su slučajne šume (eng. *Random Forests, RF*) i metoda potpornih vektora (eng. *Support Vector Machine, SVM*) se često

koriste za zadatke klasifikacije i regresije, jer donose uravnoteženost između tačnosti i interpretabilnosti izgrađenog modela. Metoda potpornih vektora predstavlja nadgledani algoritam učenja koji se koristi za rešavanje problema klasifikacije i regresije. Osnovni princip SVM algoritma jeste pronalaženje optimalne granične hiperravni (jednačina 4.32) koja maksimalizuje marginu između klasa (jednačine 4.33), odnosno rastojanje do najbližih instanci svake klase, koje se nazivaju potporni vektori (pogledati sliku 4.1). U slučajevima kada podaci nisu linearno separabilni, algoritam SVM koristi funkcije preslikavanja koje omogućavaju projektovanje podataka u višedimenzionalni prostor u kome je moguće pronaći linearnu separabilnu hiperravan.



Slika 4.1: *Granična hiperravan, margine i potporni vektori u SVM algoritmu za klasifikaciju podataka*

$$w^T x + b = 0 \tag{4.32}$$

$$\begin{aligned} w^T x_i + b &\geq 1, \quad \text{za} \quad y_i = +1 \\ w^T x_i + b &\leq -1, \quad \text{za} \quad y_i = -1 \\ \min_{w^T, b} \frac{1}{2} \|w^T\|^2 \quad za \quad y_i(w^T x_i + b) &\geq 1, \quad \forall i \end{aligned} \tag{4.33}$$

Iako tradicionalni algoritmi nisu uvek dovoljno moćni za pronalaženje složenih zavisnosti u podacima, kao što su to u stanju duboke neuronske mreže (pogledati odeljak 4.3), ovi algoritmi i dalje ostaju ključni za manje skupove podataka i zadatke u kojima je važna interpretabilnost modela.

4.2.1 Optimizovane linearne metode

Optimizovane linearne metode koriste stohastički gradijentni spust (eng. *Stochastic Gradient Descent, SGD*) za treniranje linearnih modela kao što su SVM i LR. Ove metode

su efikasne za velike skupove podataka jer, za razliku od tehnike gradijentni spust (eng. *Gradient Descent, GD*) koja u svakom koraku računa gradijent funkcije greške nad celim skupom podataka, SGD koristi samo jedan primer ili mali skup primera u svakom koraku za ažuriranje težina. Zahvaljujući ovakvom pristupu, značajno se smanjenju računarski zahtevi, čime se postiže brža konvergencija u procesu optimizacije. Tokom obuke, težine modela w se iterativno ažuriraju u smeru gradijenta funkcije greške L, definisane specifičnom funkcijom za svaki algoritam. U funkciji linearnog modela datog opštom jednačinom:

$$f_w(x) = w^T x + b \tag{4.34}$$

cilj da minimizuje funkciju greške klasifikacije koja je definisana na sledeći način:

$$E(w, b) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f_w(x_i)) + \alpha R(w) \tag{4.35}$$

gde je L funkcija greške koja meri uspešnost treniranja, $R(w)$ je regularizacioni parametar koji kontroliše složenost modela i α je nenegativan hiperparametar koji kontroliše jačinu regularizacije. Jedan od značajnih primera primene SGD metode jeste optimizacija funkcije greške kod linearnog SVM klasifikatora. U ovom kontekstu, SGD se koristi za minimizaciju funkcije *hinge* sa dodatkom L2 regularizacije, koja obezbeđuje maksimalno razdvajanje klasa i data je sledećom jednačinom:

$$L(y_i, f(x_i)) = \max(0, 1 - y_i f(x_i)) \tag{4.36}$$

Optimizacija ove funkcije greške obezbeđuje da model pronalazi optimalnu graničnu hiperravan koja maksimizuje marginu između klasa (jednačine 4.33), čime se smanjuje ukupna greška klasifikacije SVM algoritmom.

Tabela 4.3: *Pregled algoritama podržanih u SGDClassifier sa odgovarajućim funkcijama greške*

Algoritam	**Funkcija greške**	**Opis**	**Formula**
Linearni SVM	`hinge`	Maksimizuje marginu između klasa	$\max(0, 1 - y_i f(x_i))$
LR	`log`	Minimizuje log grešku za binarnu klasifikaciju	$\log(1 + e^{-y_i f(x_i)})$
Modifikovani Huber	`modified_huber`	Kombinacija funkcije *hinge* i kvadratne greške	$\begin{cases} \max(0, 1 - y_i f(x_i))^2, & y_i f(x_i) > 1 \\ -4 y_i f(x_i), & \textit{u suprotnom} \end{cases}$
Perceptron	`perceptron`	Klasičan perceptron za binarnu klasifikaciju	$\max(0, -y_i f(x_i))$
Kvadratna greška	`squared_loss`	Minimizuje kvadratnu grešku između predviđanja i stvarnih vrednosti	$\frac{1}{2}(y_i - f(x_i))^2$
Huber	`huber`	Kombinuje linearni i kvadratni pristup grešci	$\begin{cases} \frac{1}{2}(y_i - f(x_i))^2, & \lvert y_i - f(x_i)\rvert \le \delta \\ \delta \cdot \lvert y_i - f(x_i)\rvert - \frac{1}{2}\delta^2, & \textit{u suprotnom} \end{cases}$

Klasa *SGDClassifier*[7] u okviru *scikit-learn* Pajton biblioteke predstavlja uspešnu implementaciju SGD optimizacije nad linearnim klasifikacionim algoritmima izborom odgovarajuće funkcije greške (pogledati tabelu 4.3), što omogućava fleksibilnost u treniranju linearnih modela za različite zadatke klasifikacije i skupove ulaznih podataka. Pored funkcije

[7] https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.SGDClassifier.html

greške, *SGDClassifier* regularizaciju kontroliše pomoću hiperparametra stope učenja (eng. *learning rate*), koeficijenta α, kao i različitih metoda normalizacije (*L1*, *L2*, i *ElasticNet*), čime se smanjuje rizik od prekomernog prilagođavanja i kontroliše složenost modela.

4.2.2 Ansambl metode

Ansambl metode (eng. *ensemble methods*) jesu grupa algoritama koji kombinuju rezultate više osnovnih modela u cilju poboljšanja performansi pojedinačnih modela kod rešavanja klasifikacionih i regresionih zadataka [78]. Ovi algoritmi omogućavaju redukciju varijanse (eng. *variance*), pristrasnosti (eng. *bias*) i poboljšanje ukupne prediktivne stabilnosti modela, u zavisnosti od načina na koji se osnovni modeli treniraju i kombinuju. Prema načinu rada, ovi algoritmi su svrstani u sledeće tehnike:

- **Paralelna agregacija** (eng. *bootstrap aggregating*, *bagging*) – je ansambl tehnika koja ima za cilj smanjenje varijanse modela treniranjem više nezavisnih modela nad različitim podskupovima podataka. Podskupovi se generišu nasumičnim uzorkovanjem nad primerima iz originalnog skupa uz primenjeni princip vraćanja, što omogućava da se isti primer može se više puta pojaviti u jednom podskupu. Svi modeli u ansamblu se treniraju paralelno, a njihova predviđanja se kombinuju glasanjem (kod klasifikacije) ili uprosečavanjem (kod regresije). Jedan od primera algoritama koji pripadaju ovoj tehnici jeste algoritam RF, koji je u širokoj upotrebi zbog uspešne primene na velikom broju raznovrsnih zadataka. Algoritam RF je zasnovan na stablima odlučivanja, a pojedinačna stabla izgrađuje nad podskupovima inicijalnog skupa za obučavanje i različitim skupovima atributa, pri čemu se u svakom koraku odlučivanja vrši izbor optimalnog atributa za grananje. Ekstremno nasumična stabla (eng. *Extremely Randomized Trees*, *ERT*) predstavlja drugi značajan algoritam iz ove tehnike, pri čemu se najbolji atribut za grananje u svakom čvoru bira iz nasumičnog podskupa svih atributa u toku izgradnje stabla. Algoritam ERT, takođe, gradi pojedinačna stabla koristeći celokupni skup za obuku, čime se eliminiše mogućnost stvaranja šuma prilikom kreiranja podskupova. Uključivanjem slučajnosti u izbor atributa, algoritam ERT pravi skup stabala odlučivanja čija struktura nije zavisna od strukture samih atributa [64].
- **Sekvencijalno pojačavanje** (eng. *boosting*) – je sekvencijalna ansambl tehnika koja kombinuje niz slabih modela tako što svaki naredni model uči iz grešaka prethodnih. Za razliku od tehnike paralelne agregacije u kojoj se modeli izgrađuju nezavisno, u tehnici sekvencijalnog pojačavanja modeli zavise jedan od drugog, odnosno svaki model pokušava da popravi greške prethodnih modela u nizu. Tahnika sekvencijalnog pojačavanja smanjuje pristrasnost i često daje izuzetno precizne modele, ali modeli mogu biti podložni preteranom prilagođavanju. Najpoznatiji algoritmi koji pripadaju ovoj grupi su *AdaBoost*, *Gradient Boosting*, *XGBoost*, *LightGBM* i *CatBoost* algoritam.
- **Slaganje** (eng. *stacking*) – je tehnika koja kombinuje predviđanja više različitih i heterogenih modela u višeslojnu arhitekturu. Prvi sloj sadrži više osnovnih modela koji se obučavaju nezavisno, dok njihov izlaz služi kao ulaz za meta-model koji se obučava u višem sloju. Cilj tehnike slaganja je da meta-model nauči kako da kombinuje predviđanja osnovnih modela da bi poboljšao ukupnu tačnost predviđanja.
- **Glasanje** (eng. *voting*) – predstavlja najjednostavniju tehniku ansambl metoda, u kojoj se izlazi više nezavisno izgrađenih modela koriste za dobijanje konačnog predviđanja. Kod klasifikacionih zadataka se može koristiti „tvrdo glasanje“, kada se odabira većinska klasa, ili nasuprot tome, „meko glasanje“, kada se klasa odabira na osnovu proseka dobijenih verovatnoća za svaku klasu. U slučaju regresionih metoda, najčešće

se primenjuje aritmetički prosek dobijenih vrednosti u osnovnim modelima.

4.3. Algoritmi dubokog učenja

Neuronske mreže predstavljaju temelj dubokog mašinskog učenja (eng. *Deep Learning, DL*) i inspirisane su načinom na koji mozak obrađuje informacije. Neuronska mreža se sastoji od više slojeva povezanih neurona, gde svaki neuron obrađuje informacije primljene sa svojih ulaza i prosleđuje ih kroz mrežu. Potpuno povezane neuronske mreže (eng. *Fully Connected Neural Networks, FCNN*) jesu najrasprostranjeniji model neuronskih mreža. Mreža FCNN sastoji se od slojeva neurona, a svaki neuron u jednom sloju prima ulazne vrednosti od svih neurona iz prethodnog sloja i prosleđuje svoje izlaze svim neuronima u narednom sloju, čime se obezbeđuje potpuna međusobna povezanost između susednih slojeva neurona [90]. mreže sa propagacijom unapred (eng. *Feed Forward Neural Networks, FFN*) su specijalna vrsta FCNN neuronskih mreža, u kojima podaci teku samo u jednom smeru, od ulaznog sloja preko skrivenih slojeva do izlaznog sloja, bez povratnih ili cikličnih veza. Težine predstavljaju parametre koji modeluju veze između neurona u različitim slojevima mreže (pogledati sliku 4.2). Težine se inicijalno postavljaju na proizvoljne vrednosti, koje se tokom procesa obučavanja modela postepeno menjaju prema optimalnim vrednostima.

Tokom obučavanja, mreža koristi algoritam propagacije greške unazad zajedno sa metodom optimizacije (na primer, GD ili SGD) da bi prilagodila težine. Glavni cilj obuke je da se smanji razlika između stvarnih izlaza i onih koje model predviđa, prilagođavanjem težina tako da mreža bolje odgovara ulaznim podacima [22]. Pored težina, ključna komponenta FCNN mreže je aktivaciona funkcija, koja omogućava mreži da modelira nelinearne odnose i na taj način uči složene relacije koje mogu postojati između atributa u ulaznim podacima (pogledati sliku 4.2).

Najčešće korišćene aktivacione funkcije u neuronskim mrežama, prikazane grafički na slici 4.3, su:

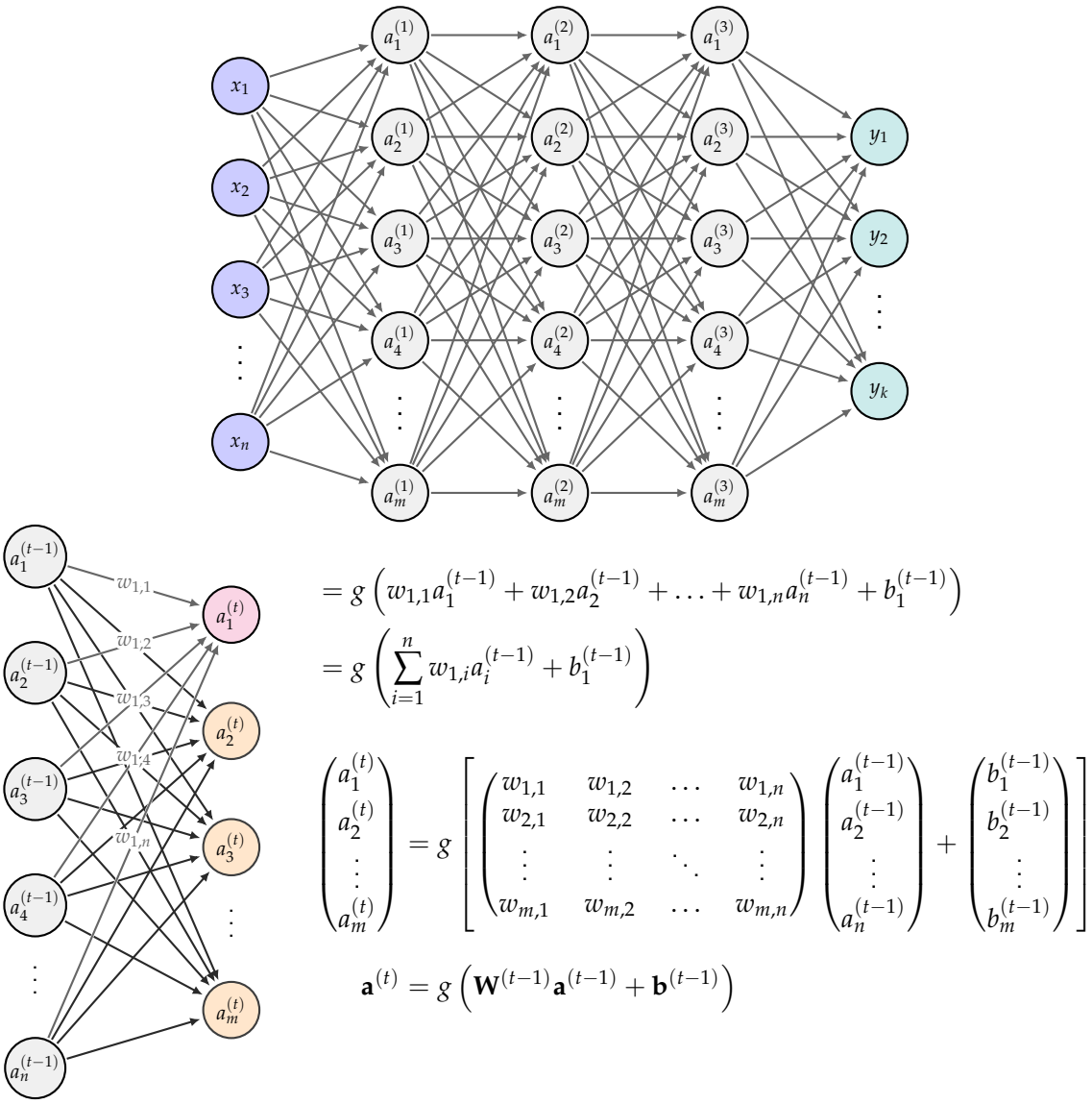
- Funkcija ispravljene linearne jedinice (eng. *Rectified Linear Unit, ReLU*);
- Funkcija hiperboličkog tangensa (eng. *Hyperbolic Tangent, tanh*);
- Sigmoidna funkcija (eng. *Sigmoid*, σ);
- Funkcija mekog maksimuma (eng. *Softmax*, $softmax$).

Ove funkcije se prema svojim karakteristikama koriste u različitim slojevima neuronske mreže i za razrešavanje drugačijih zadataka. Tako je, na primer, *ReLU* funkcija značajna zbog jednostavne primene i efikasnosti u učenju, posebno u unutrašnjim slojevima mreže, jer smanjuje problem nestajućih (eng. *vanishing*) i eksplodirajućih (eng. *exploding*) gradijenata u dubljim slojevima [22], dok se σ i $softmax$ najčešće koriste u izlaznom sloju neuronske mreže jer omogućavaju modelu da izračuna verovatnoću izlaza za svaku klasu kod binarne i višeklasne klasifikacije, u tom redosledu [90].

4.3.1 Rekurentne neuronske mreže

Rekurentne neuronske mreže (eng. *Reccurent Neural Networks, RNN*) su posebna arhitektura neuronskih mreža dizajnirana za rad sa sekvencijalnim podacima. Za razliku od tradicionalnih neuronskih mreža, koje tretiraju svaku ulaznu tačku nezavisno, RNN koriste povratne veze koje omogućavaju modelu da zadrži informacije iz prethodnih koraka

[8] Izvor: https://tikz.net/neural_networks/



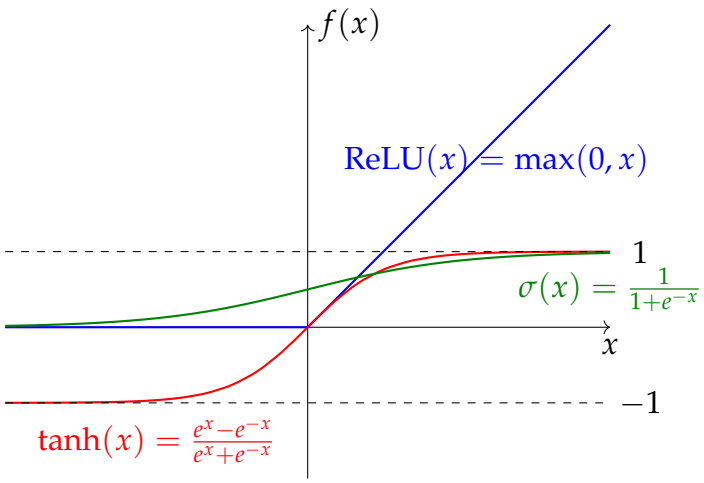
Slika 4.2: *Duboka nuronska mreža (DNN) i osnovni princip rada jednog neurona neuronske mreže. Preuređena slika originalne slike*[8]

u sekvenci, što je ključno za zadatke kao što su obrada tekstualnih podataka, vremenskih serija ili prepoznavanje govora. Međutim, standardna RNN često ima problema sa dugoročnim pamćenjem, odnosno problem nestajućeg gradijenta, gde vrednost gradijenta tokom vremena eksponencijalno opada u toku povratne propagacije [22]. Kako bi rešili ove probleme, istraživači su razvili varijante RNN, koje koriste dodatne kontrolne mehanizme za čuvanje i upravljanje informacijama kroz duži vremenski period. U RNN mreži, u svakom koraku sekvence (t) mreža prihvata ulazni vektor x_t i ažurira skriveno stanje h_t na osnovu prethodnog stanja h_{t-1} i trenutnog ulaza x_t. Skriveno stanje h_t sadrži informacije o prethodnim koracima, omogućavajući mreži da zadrži memoriju tokom obrade sekvence (jednačina 4.37)

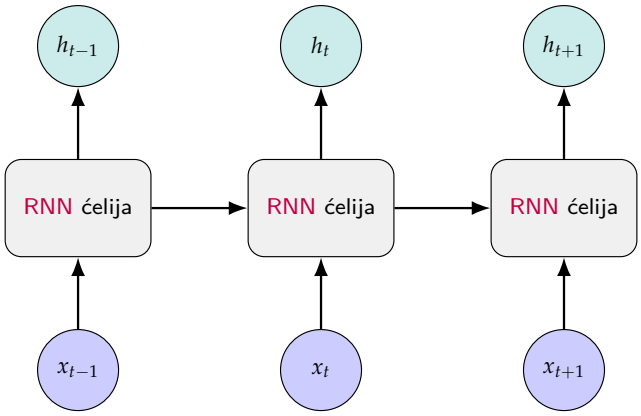
$$h_t = f(x_t, h_{t-1}; \theta) = \tanh(W_h h_{t-1} + W_x x_t + b_h) \tag{4.37}$$

gde su $x_t \in \mathbb{R}^{|V|}$ ulaz u koraku t, $h_t \in \mathbb{R}^d$ skriveno stanje u koraku t, i $\theta = \{W_x, W_h, b_h\}$ težinske matrice i f nelinearna aktivaciona funkcija (tanh, *ReLU*).

Konačno skriveno stanje h_t se u klasifikaciji teksta može koristiti za izračunavanje predviđanja za datu klasu, kako je prikazano sledećom jednačinom:



Slika 4.3: *Najčešće korišćene aktivacione funkcije: ReLU, σ i* tanh



Slika 4.4: *Razvijena RNN mreža*

$$\hat{y} = softmax(W_y h_T + b_y), W_y \in \mathbb{R}^{2\times d}, b_y \in \mathbb{R}^d \tag{4.38}$$

gde je $\hat{y}$ vektor verovatnoća za svaku klasu, W_y je težinska matrica za izlazni sloj i b_y je slobodni vektor. Funkcija $softmax$ omogućava da zbir izlaza bude jednak jedinici, pri čemu se ovi izrazi mogu posmatrati kao verovatnoće pripadnosti svakoj od klasa.

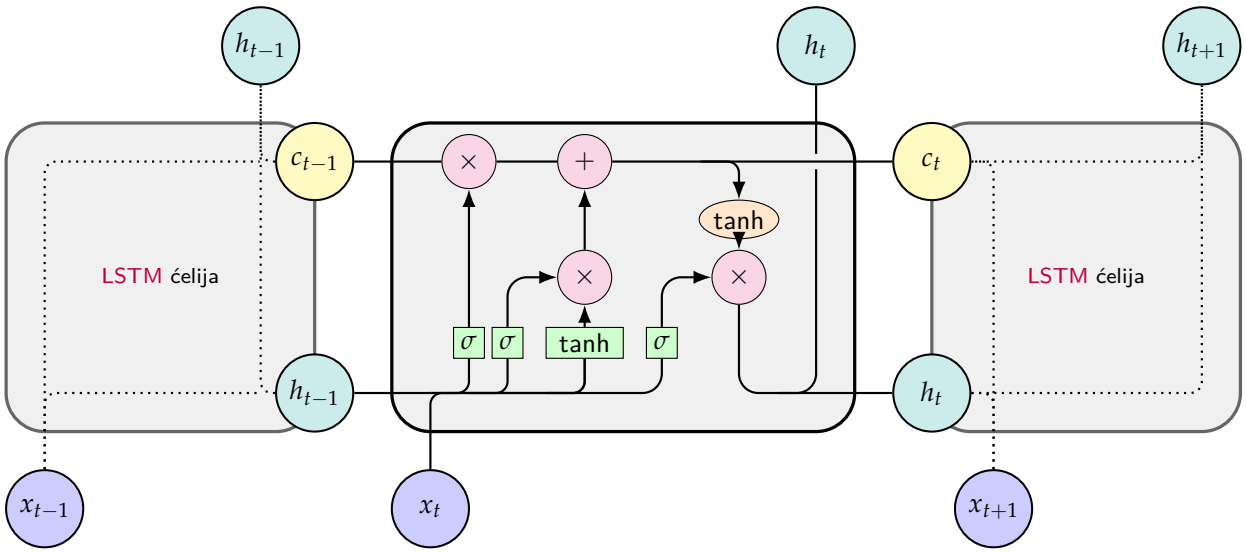
$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{4.39}$$

RNN mreže sa dugom kratkoročnom memorijom

RNN mreže sa dugom kretkoročnom memorijom (eng. *Long Short-Term Memory, LSTM*) uvode složeniju unutrašnju strukturu sa memorijskim ćelijama koje omogućavaju pamćenje obrađenih informacija u kraćem ili dužem vremenskom periodu [69]. Iako LSTM ćelije mogu varirati po načinu povezivanja i aktivacionih funkcija koje se koriste, sve imaju eksplicitne memorijske ćelije za čuvanje informacija. U okviru LSTM mreže, sadržaj memorijske ćelije se može ažurirati ili proslediti sledećem koraku u iteraciji (slika 4.5).

Arhitektura LSTM ćelije uvodi sledeći sistem kapija za kontrolisanje toka informacija kroz mrežu:

- **Ulazna kapija** (eng. *input gate, i*) — odlučuje da li će memorijska ćelija biti ažurirana.



Slika 4.5: *Razvijena LSTM mreža sa prikazom strukture LSTM ćelije i prenosa informacija između ćelija*

Takođe, kontroliše koliko će informacija trenutna memorijska ćelija primiti od potencijalno nove memorijske ćelije:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{4.40}$$

- **Kapija zaboravljanja** (eng. *forget gate*, f) – kontroliše koliko informacija će memorijska ćelija primiti od memorijske ćelije iz prethodnog koraka:

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{4.41}$$

- **Izlazna kapija** (eng. *output gate*, o) -– kontroliše vrednost sledećeg skrivenog stanja:

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{4.42}$$

pri čemu se vrednosti memorijskih ćelija i skrivenih stanja izračunavaju na sledeći način:

$$\tilde{c_t} = \tanh(w_c[h_{t-1}, x_t] + b_c) \tag{4.43}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t} \tag{4.44}$$

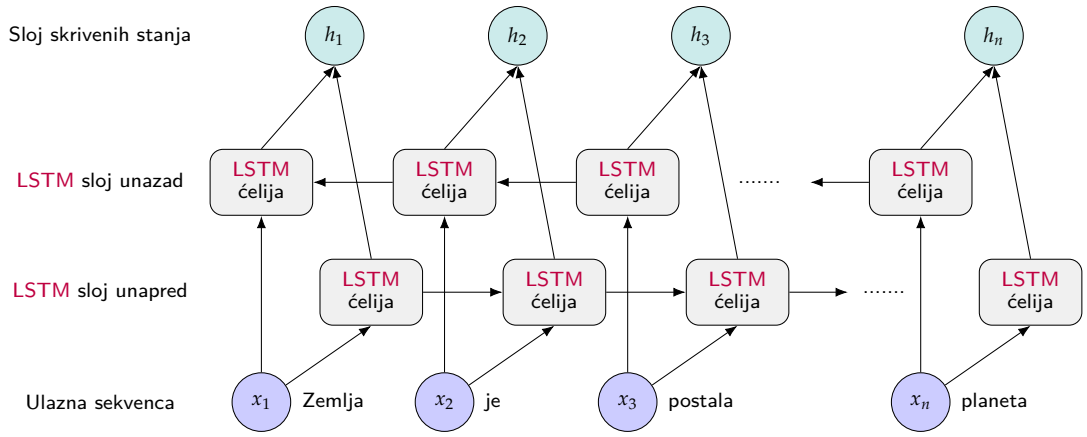
$$h_t = o_t * \tanh(c_t) \tag{4.45}$$

Jedan od nedostataka standardne LSTM mreže jeste da tokom obrade sekvenci koristi samo informacije iz prethodnih vremenskih koraka. Buduće kontekstualne informacije, koje obuhvataju delove sekvence koje nastupaju nakon trenutnog elementa, kao što su naredne reči u rečenici ili sledeće vrednosti u vremenskoj seriji, ostaju nedostupne iako mogu biti značajne za preciznije modelovanje podataka. Iz tog razloga razvijena su različita unapređenja ove arhitekture.

Dvosmerne LSTM mreže

Dvosmerna LSTM mreža (eng. *Bidirectional LSTM, BiLSTM*) koristi prethodne i buduće kontekste tako što tekstualnu sekvencu obrađuje u oba smera [176]. Dvoslojnom i

dvosmernom arhitekturom BiLSTM mreža je u mogućnosti da modeluje sekvencijalne zavisnosti koje postoje između reči i fraza u tekstualnoj sekvenci i stoga je pronašla svoju primenu u analizi tekstualnih sadržaja. LSTM slojevi unapred i unazad u okviru BiLSTM arhitekture poseduju sopstvene skupove težina, koje se tokom procesa treniranja nezavisno optimizuju. Skrivena stanja iz oba smera se kombinuju nekom od metoda agregacije, kao što su sumiranje, množenje ili dopisivanje, čime mreža istovremeno koristi informacije iz prošlih i budućih vremenskih koraka i time obezbeđuje robusnije modelovanje ulaznih sekvenci (slika 4.6).



Slika 4.6: *Obrada ulazne tekstualne sekvence na srpskom jeziku pomoću BiLSTM mreže*

Na primeru rečenice iz srpskog jezika koja počinje tekstom:

„Zemlja je postala...“

standardna LSTM mreža iz početnog konteksta rečenice ne bi mogla da utvrdi pravo značenje reči „zemlja“ koja u srpskom jeziku ima višestruka značenja (homonim), dok bi BiLSTM mreža njeno značenje mogla da utvrdi iz nastupajućeg konteksta (dodatni LSTM sloj unazad) dodeljujući joj različite vrednosti u sledećim primerima:

„Zemlja je postala rastresita nakon primene agrotehničkih mera.“

„Zemlja je postala član međunarodne organizacije za zaštitu životne sredine.“

„Zemlja je postala prenaseljena planeta.“

Ovim primerom jasno se ilustruje ograničenje standardne LSTM mreže i prednost BiLSTM arhitekture, koja zahvaljujući uvažavanju budućih kontekstualnih informacija omogućava preciznije razumevanje značenja višeznačnih reči.

4.3.2 Mehanizam pažnje

Mehanizam pažnje (eng. *Attention mechanism, Att*) je tehnika koja omogućava neuronskim mrežama da se fokusiraju na određene delove ulaznih podataka tokom procesiranja. Ovo je posebno korisno kod obrade sekvencijalnih podataka, gde su neki delovi sekvence relevantniji od drugih za rešavanje određenog zadatka. Mehanizmi pažnje se mogu ograničiti samo na određene delove ulaznih podataka, celokupnu sekvencu ili se mogu iterativno izgrađivati prateći kontekst sekvence [17]. Mehanizmi pažnje rešavaju izazov velikih količina podataka, odnosno dugačkih tekstualnih sekvenci, tako što neuronskim mrežama u koje se ugrađuju omogućavaju da se selektivno fokusiraju na određene delove ulaznih podataka, čime se smanjuje računsko opterećenje i poboljšava tačnost predviđanja. Koncept

pažnje ima svoje korene u kognitivnoj psihologiji, koja označava selektivno obrađivanje informacija od strane ljudskog mozga. U dubokom učenju, mehanizmi pažnje oponašaju ovaj proces dodeljivanjem težine različitim delovima ulaznih podataka na osnovu njihove relevantnosti za zadatak koji se rešava.

U DL modelima, kao što su RNN, LSTM ili BiLSTM, model obrađuje čitavu ulaznu sekvencu u obliku kontekstualnog vektora fiksne dužine, koji se koristi za generisanje izlaznog rezultata. Ovaj pristup, međutim, ima značajno ograničenje jer kontekstualni vektor fiksne veličine možda neće efikasno uhvatiti sve relevantne informacije iz dugih ulaznih sekvenci, što može dovesti do suboptimalnih performansi. Mehanizmi pažnje rešavaju ovo ograničenje dozvoljavajući modelu da kreira drugačiji vektor konteksta za svaki izlazni element (jednačine 4.46 i 4.47). Ovo se postiže izračunavanjem skupa težina koje određuju važnost svakog ulaznog elementa u odnosu na trenutni izlazni element koji se generiše (jednačina 4.48). Težine pažnje se zatim koriste za kreiranje težinske sume ulaznih elemenata, koja služi kao vektor konteksta koji se obrađuje u narednim koracima (jednačina 4.49).

$$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}] \tag{4.46}$$

$$e_i = tanh(W_h h_i + b_h), e_i \in [-1, 1] \tag{4.47}$$

$$w_i = \frac{exp(e_i)}{\sum_{i=1}^{N} exp(e_i)}, \sum_{i=1}^{N} w_i = 1 \tag{4.48}$$

$$r = \sum_{i=1}^{N} w_i h_i, r \in R^{2N} \tag{4.49}$$

gde su W_h i b_h matrica težina i vektor slobodnih koeficijenata u nivou pažnje.

U transformer arhitekturi neuronskih mreža (pogledati naredni odeljak 4.3.3), mehanizam samopažnje (eng. *self-attention*) omogućava modelu da izračuna odnose među rečima unutar iste sekvence tako što se svaka reč kombinuje sa svim ostalim rečima za prikupljanje relevantnih informacija. Ova funkcionalnost se postiže generisanjem tri vektora za svaku reč koji se dobijaju množenjem ulaznog vektora sa naučenim matricama težine, i to su vektori:

- upita (eng. *Query*, Q),
- ključeva (eng. *Key*, K),
- vrednosti (eng. *Value*, V).

Relevantnost odnosa između dve reči (pažnja) se izračunava korišćenjem sledeće formule:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V \tag{4.50}$$

gde su Q, K i V vektori upita, ključeva i vrednosti, a d_k dimenzija vektora ključeva.

Ovaj proces omogućava modelu da prepozna kontekstualne reprezentacije reči koje uzimaju u obzir njihovu povezanost sa drugim rečima unutar tekstualne sekvence. Za dodatno unapređenje performansi, transformer arhitektura uvodi mehanizam višestruke pažnje (eng. *multi-head attention*) u kojem se pažnja deli na više paralelnih mehanizama pažnje. Svaki mehanizam pažnje koristi nezavisne težinske matrice za vektore upita (Q),

ključeva (K) i vrednosti (V) za jedan aspekt zavisnosti među rečima, čime se modelu omogućava da nauči različite odnose koji postoje među rečima. Nakon što svaki mehanizam pažnje izračuna svoj izlaz Att_i, rezultati se spajaju i projektuju u isti dimenzionalni prostor:

$$\text{Att}_i = \text{Att}(QW_i^Q, KW_i^K, VW_i^V), \tag{4.51}$$

$$\text{Multi-Att}(Q, K, V) = Concat(\text{Att}_1, \text{Att}_2, \ldots, \text{Att}_h)W^O \tag{4.52}$$

gde su W_i^Q, W_i^K, W_i^V težinske matrice specifične za svaki mehanizam pažnje, a W^O matrica projekcije za konačnu transformaciju. Kombinovanjem rezultata iz više mehanizama pažnje, model može istovremeno da analizira zavisnosti iz više aspekata teksta, što je ključna prednost modela transformer arhitekture u NLP zadacima.

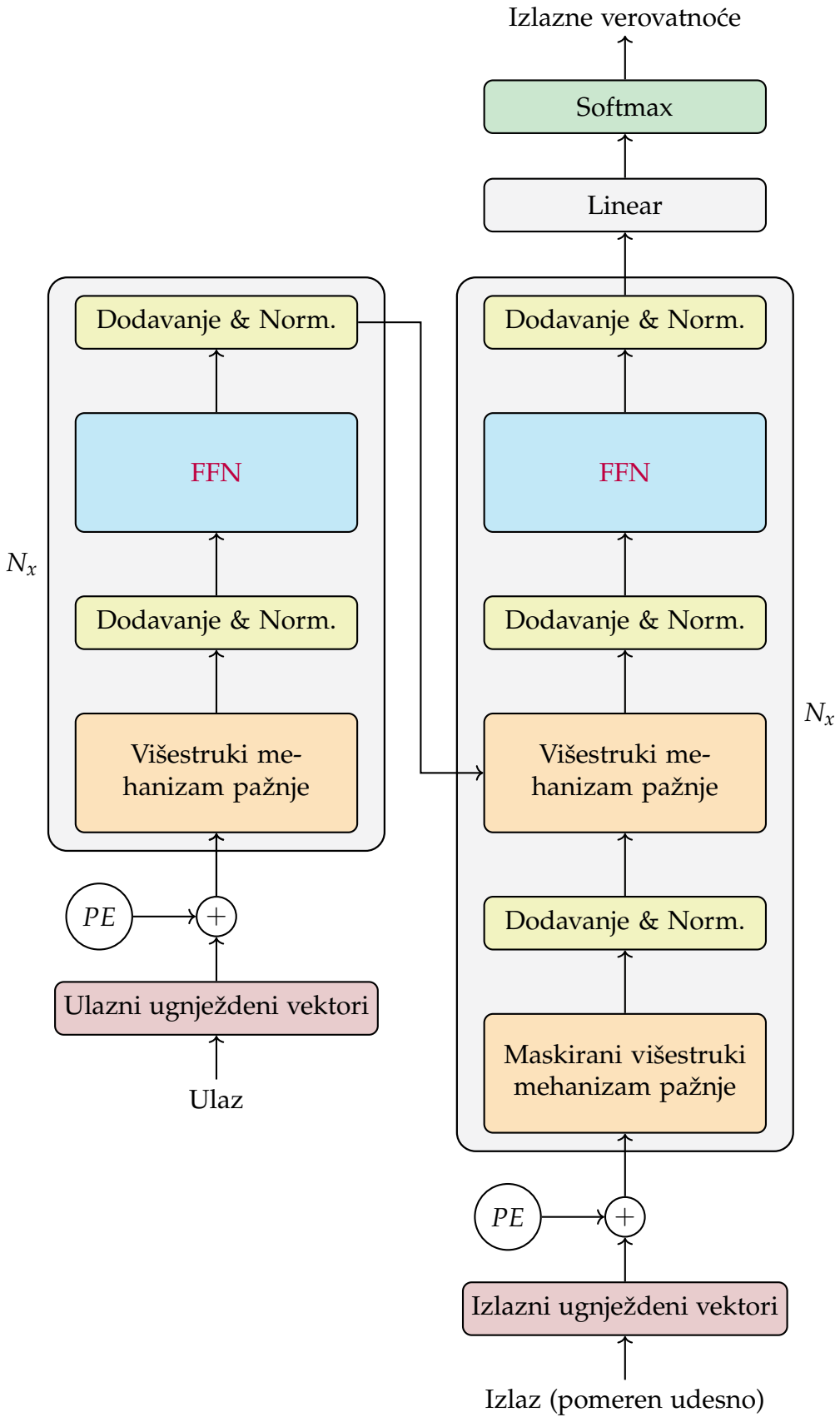
4.3.3 Transformeri

Transformeri predstavljaju jednu vrstu arhitekture algoritama dubokog učenja čije glavne komponente čine enkoder i dekoder, koji se sastoje iz više slojeva (pogledati sliku 4.7). Proces učenja počinje konvertovanjem ulaznih reči u ugnježdene vektorske reprezentacije reči (pogledati poglavlje 5), koje se zatim dopunjuju informacijom o redosledu reči kroz poziciono enkodiranje. Ovi vektori prolaze kroz enkoder komponentu transformera, u kojem višestruki mehanizam pažnje omogućava višedimenzionalnu analizu odnosa između reči u sekvenci, dok FFN sloj dodatno obrađuje prikupljene informacije iz prethodnih slojeva. Mehanizam samopažnje unutar enkodera omogućava modelu da se fokusira na relevantne delove ulaza i uhvati odnose između različitih reči i fraza u tekstualnoj sekvenci, dok slojevi za dodavanje i normalizaciju stabilizuju proces učenja i omogućavaju lakšu konvergenciju modela. Dekoder, slične strukture, koristi maskirani višestruki mehanizam pažnje kako bi se sprečilo „gledanje unapred“ tokom generisanja teksta, dok višestruki mehanizam pažnje povezuje trenutni izlaz dekodera sa izlazima enkodera kako bi se integrisao kontekst ulazne sekvence u procesu zaključivanja. Na kraju, linearni sloj sa $softmax$ aktivacionom funkcijom transformiše izlaz iz dekoder komponente u verovatnoće pojavljivanja reči iz rečnika, čime se omogućava generisanje najverovatnije nastupajuće reči u izlaznoj sekvenci [206].

Jedan od prvih modela izgrađenih na arhitekturi transformera jesu dvosmerne enkoderske reprezentacije iz transformera (eng. *Bidirectional Encoder Representations from Transformers, BERT*) koji je napravio značajan iskorak u vektorskoj reprezentaciji reči koja je zavisna od konteksta [94]. BERT uključuje dvosmernost u rešavanju zadatka predviđanje skrivenog tokena (eng. *Masked Language Model, MLM*) sa ciljem da predvidi pojavljivanje reči na određenoj maskiranoj poziciji, označenoj sa [MASK], koristeći samo kontekst u kojem se ta reč pojavljuje. Preciznije, iz datog skupa tokena ulazne sekvence $x = [x_1, x_2, ..., x_N]$, dužine N, maskira se određeni skup tokena $y = [y_1, y_2, ..., y_T]$, gde je $T < N$, sa ciljem da se maksimizuje verovatnoća pojavljivanja pojedinačnih tokena na njima odgovarajućim pozicijama koja je data sledećom jednačinom:

$$\max_{\theta} log p_\theta(y|x) = \sum_{t=1}^{T} log p_\theta(y_t|x) \tag{4.53}$$

Drugi zadatak koji BERT rešava jeste zadatak predviđanje nastupajuće rečenice (eng. *Next Sentence Prediction, NSP*) koji se obučava na parovima neposrednih rečenica u tekstu. BERT koristi samo enkoderski deo transformer arhitekture sa dvosmernim pristupom, što ga čini idealnim za zadatke kao što su razumevanje konteksta, klasifikacija i ekstrakcija informacija. Sa druge strane, enkoder-dekoder arhitektura koristi enkoder za pretvaranje ulaznih



Slika 4.7: *Arhitektura transformera koja uključuje enkoder, dekoder i višestruki mehanizam pažnje. Preuređena slika na osnovu originalne slike iz rada [206]*

podataka u kontekstualne reprezentacije i dekoder za generisanje transformisanog izlaza, što omogućava primene poput prevođenja, sažimanja teksta i generisanja teksta. Specijalno, autoenkoder transformer arhitektura koristi samo enkoderski blok za kompresiju i rekonstrukciju ulaznih podataka, u cilju generisanja latentnih reprezentacija koje sažimaju ključne informacije ulazne sekvence. Za razliku od BERT modela, grupa modela označenih kao generativni prethodno trenirani transformer (eng. *Generative Pre-Trained Transformer, GPT*), koristi samo dekoder deo arhitekture i jednosmerno obrađuje tekst, u cilju generisanja sledeće reči u sekvenci, zbog čega je odličan za kreiranje sadržaja i zadatke poput rezimiranja i prevođenja teksta (pogledati tabelu 4.4). Čet generativni prethodno trenirani transformer (eng. *Chat Generative Pre-Trained Transformer, Čet-GPT*), koji se zasniva na GPT modelu (kao što su GPT-3 ili GPT-4), je specijalno fino podešen model za vođenje dijaloga sa korisnicima. Dok je BERT fokusiran na razumevanje teksta, GPT i Čet-GPT su primarno dizajnirani za generisanje teksta, pri čemu je Čet-GPT dodatno usmeren na pružanje relevantnih i kontekstualno tačnih odgovora u konverzacijama sa korisnicima (pogledati tabelu 4.6).

Tabela 4.4: *Poređenje transformer arhitektura zasnovanih na enkoderu, dekoderu i autoenkoderu*

Karakteristika	**Enkoder**	**Dekoder**	**Autoenkoder**	**Enkoder-Dekoder**
Ulaz	Podaci (tekst, slika)	Trenutni izlaz + kontekst	Podaci (tekst, slika)	Podaci (tekst, slika, audio)
Izlaz	Kontekstualne reprezentacije	Generisani niz	Rekonstruisani ulaz	Transformisani niz
Glavna primena	Analiza teksta, prepoznavanje	Generisanje teksta	Rekonstrukcija, kompresija	Prevođenje, sažimanje, generisanje teksta i slika
Primer modela	BERT	GPT	Autoenkoder	T5, BART, DALL-E

Izmenama osnovne transformer arhitekture razvijene su brojne varijacije transformer modela, nastalih sa ciljem unapređenja performansi, podrške jezicima i opsega zadataka koje rešava osnovni model (BERT). Neke od najpoznatijih takvih varijacija su:

- Višejezični BERT (eng. *Multilingual BERT, mBERT*) – varijanta BERT modela obučen nad skupu Vikipedija podataka iz 104 različita jezika sa ciljem da model nauči reprezenzacije koje su istovremeno primenljive na više različitih jezika.
- Robusno optimizovani BERT (eng. *Robustly Optimized BERT, RoBERTa*) – unapređuje performanse BERT modela korišćenjem dodatnih skupova podataka (CommonCrawl News), dužim obučavanjem i većim skupom podataka koji se koristi u svakoj iteraciji u toku obuke. U ovoj arhitekturi modela je isključen NSP zadatak, a maskirani tokeni se dinamički menjaju u podacima u toku obučavanja modela [120].
- Unakrsno-jezički model (eng. *Cross-lingual Language Model, XLM*) – koristi Predviđanje skrivenog tokena prevođenjem (eng. *Translation Language Model, TLM*) koji proširuje MLM model korišćenjem paralelnih rečenica iz različitih jezika. Tokeni se maskiraju metodom slučajnog izbora u izvornoj i izlaznoj rečenici, sa ciljem da se predvide nedostajući tokeni u svakom od jezika koji se koriste za obučavanje modela [45].
- XLM-R – proširuje XLM na RoBERTa arhitekturu, optimizovanu za rad sa 100 jezika. Ovaj model predstavlja unakrsno-jezičku (eng. *cross-lingual*) arhitekturu koja je omogućena obučavanjem nad velikim skupovima podataka, deljenim rečnikom između jezika i korišćenjem TLM pristupa u maskiranju tokena.

- GPT-2/3/4 – jesu modeli transformer arhitekture zasnovani na dekoder bloku, izgrađeni sa ciljem rešavanja NSP zadatka. Ove modele karakteriše veliki broj parametara i obuka na ogromnim skupovima podataka (570 GB teksta - korpusi *Common Crawl* i *Books*, tekstovi preuzeti sa Vikipedija portala, izvorni programski kod sa *GitHub* platforme, Redit poruke, tekstovi sa internet foruma i drugi) što im omogućava razumevanje konteksta i generisanje koherentnog teksta. Verzije modela GPT-3 i GPT-4 donose poboljšanja u kapacitetu, efikasnosti i sposobnosti rešavanja kompleksnih zadataka, koji uključuju prevođenje, kodiranje, logičko rezonovanje i kreativno pisanje.
- Veliki jezički Meta AI modeli (eng. *Large Language Models Meta AI, LLaMA*)[9] – jesu kolekcija LLM transformer arhitekture, koju je razvila Meta (Facebook) kompanija, sa modelima dostupnim u različitim verzijama i veličinama. Modeli su obično označenim brojem parametara (1B, 3B, 7B i drugi), što omogućava fleksibilnost u odabiru modela zavisno od dostupnih računarskih resursa. Slično kao i kod GPT modela, ovi modeli imaju sposobnost za efikasno obavljanje različitih NLP zadataka, kao što su generisanja teksta, prevođenje, sumarizacija i odgovaranje na pitanja.

Pojava transformer arhitekture modela uvela je i inovativne metode tokenizacije teksta kao što su kodiranje po parovima bajtova (eng. *Byte-Pair Encoding*, *BPE*), delovima reči (eng. *WordPiece*, *WP*) ili delovima rečenice (eng. *SentencePiece*, *SP*). BPE spaja najčešće parove karaktera u stabilne tokene, omogućavajući modelu fleksibilnost kod reči van rečnika; često se koristi u modelima poput RoBERTa i GPT-3. WP koristi sličan princip, ali spaja segmente na osnovu verovatnoće koja optimizuje predviđanje reči, što ga čini korisnim za XLM i druge modele sa potrebom za finom segmentacijom. SP tretira razmake kao deo tokena, što omogućava segmentaciju jezika bez razdvajanja na reči i čini ga korisnim za jezike bez jasnih razmaka između reči, kao što je japanski; koristi se u modelima kao što je XLM. Ove metode omogućavaju modelima prilagodljivost u segmentaciji i generalizaciju na različite jezike i rečnike. U tabeli 4.5 predstavljeni su neki od najznačajnijih modela BERT arhitekture koji pružaju podršku za srpski jezik ili su indirektno uticali na razvoj jednojezičkih modela za srpski jezik. Oznake u tabeli predstavljaju: #J broj podržanih jezika, T vrsta tokenizatora, L broj slojeva modela, H dimenizija skrivenog sloja, H_{FF} dimenizija FFN sloja, A broj ponavljanja u višestukom mehanizmu pažnje, V veličina rečnika i #Param ukupan broj parametara modela.

Tabela 4.5: *Karakteristike nekih od modela BERT arhitekture*

Model	#J	Jezici	T	L	H	H_{FF}	A	V	#Param
BERT	1	engleski	WP	12	768	3072	12	30k	110M
RoBERTa	1	engleski	BPE	12	768	3072	12	50k	125M
mBERT	100	višejezični (100+ jezika)	WP	12	768	3072	12	120k	172M
XLM	15	višejezični (15 jezika)	BPE	12	1024	4096	8	95k	250M
XLM-R_{Base}	100	višejezični (100 jezika)	SP	12	768	3072	12	250k	270M
XLM-R_{Large}	100	višejezični (100 jezika)	SP	24	1024	4096	16	250k	550M
BERTić	1	srpski	BPE	12	768	3072	12	32k	110M
Jerteh-81	4	balkanski jezici	BPE	12	768	3072	12	50k	81M
Jerteh-355	4	balkanski jezici	BPE	24	2048	8192	16	50k	355M

U modelima dubokog učenja neophodno je da obučeni model podržava dati jezik, odnosno da je obučen nad tekstualnim podacima iz datog jezika. Ukoliko je model na-

[9] https://www.llama.com

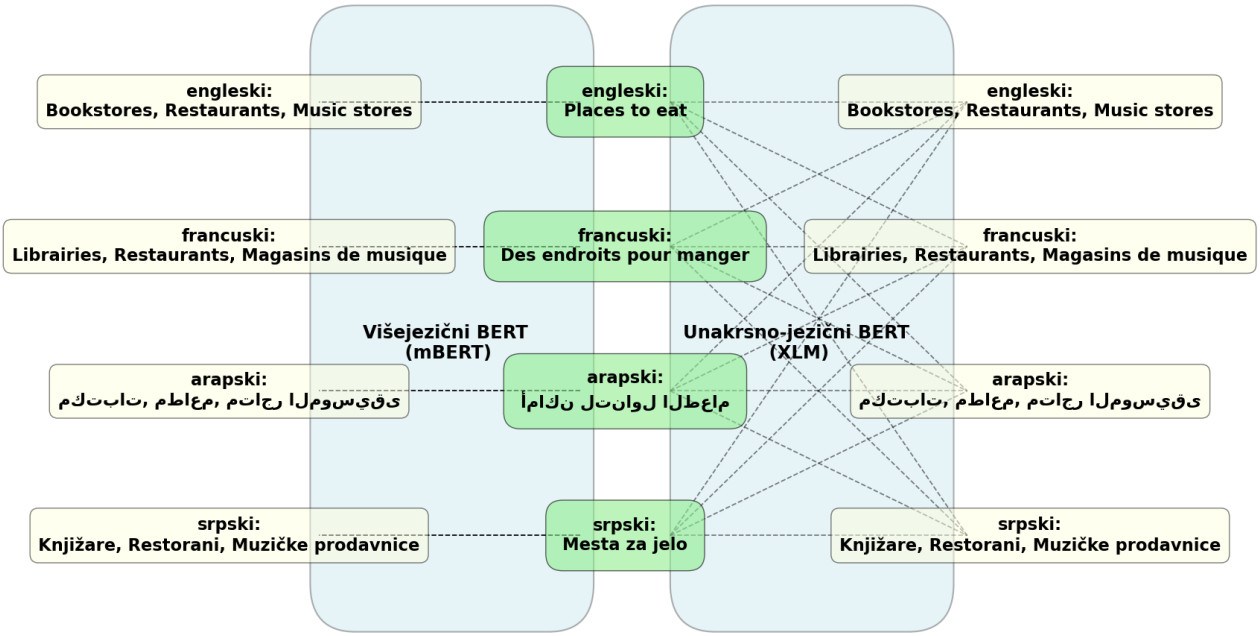
Tabela 4.6: *Karakteristike nekih velikih jezičkih modela GPT arhitekture*

Model	Proizvođač	Otvoren	#Params	Doobuč.	Višemodal.	#J	Jezici
GPT-3	OpenAI	Ne	175B	Ne	Ne	1	engleski
GPT-3.5 Turbo	OpenAI	Ne	175B	Da	Ne	1	engleski
GPT-4	OpenAI	Ne	1.76T	Ne	Da	26+	višejezični (26+ jezika)
GPT-4o	OpenAI	Ne	350B	Da	Da	26+	višejezični (26+ jezika)
GPT-4o Mini	OpenAI	Ne	3B	Da	Ne	26+	višejezični (26+ jezika)
LLaMA	Meta AI	Da	65B	Da	Ne	8+	višejezični (8+ jezika)
LLaMA 2	Meta AI	Da	7B, 13B, 70B	Da	Ne	8+	višejezični (8+ jezika)
LLaMA 3.1	Meta AI	Da	8B, 70B, 450B	Da	Ne	8+	višejezični (8+ jezika)
LLaMA 3.2	Meta AI	Da	1B, 3B, 11B, 90B	Da	Da	8+	višejezični (8+ jezika)
Gemini	Google	Ne	175B, 1.76T	Da	Da	50+	višejezični (50+ jezika)
Mistral 7B	Mistral AI	Da	7.3B, 46.7B	Da	Ne	20+	višejezični (20+ jezika)
Falcon 40B	Falcon LLM	Da	7B, 40B, 180B	Da	Ne	20+	višejezični (20+ jezika)
Claude	Anthropic	Ne	52B, 70B	Da	Ne	1	engleski
Vrabac	Jerteh	Da	110M	Da	Ne	1	srpski
Orao	Jerteh	Da	355M	Da	Ne	4	balkanski jezici
Jugo-GPT	RunAI	Da	1.2B	Da	Ne	4	balkanski jezici

pravljen isključivo nad podacima datog jezika, za njega kažemo da je jednojezični (eng. *mono-lingual*). Ovakvi modeli bi trebalo da pokazuju najbolje performanse jer su izgrađeni nad velikim korpusima tekstualnih podataka jednog jezika te su stoga naučeni da modeluju morfologiju i sintaksu tog jezika, isključujući bojazan za potencijalnim mešanjem iste forme reči koje u različitim jezicima mogu imati drugačija značenja. Međutim, veliki korpusi označenih tekstualnih podataka su često nedostupni i to je najveći problem u jezicima koji su manje zastupljeni (eng. *low-resource languages*) u odnosu na engleski jezik. Iz ovog razloga su razvijeni višejezični modeli koji su obučeni nad podacima iz više jezika. Transformer modeli se mogu doobučavati (eng. *fine-tuning*) za rešavanje specifičnih zadataka u podržanim jezicima. Standardno doobučavanje uključuje obuku celog modela na ciljnom skupu podataka za specifičan zadatak, što je uobičajeno kod jednostavnih klasifikacionih i regresionih problema. Kod pristupa koji koristi transferno učenje, slojevi bliže početku modela se zamrzavaju, dok se samo završni slojevi obučavaju. Ova tehnika je naročito korisna kada su računarski resursi ograničeni, dostupni podaci male veličine ili kada je pretrenirani model specifičan za određeni domen. Doobučavanje zasnovano na instrukcijama, koje je karakteristično za generativne modele, podrazumeva upotrebu tekstualnih instrukcija za rešavanje zadataka bez ponovnog procesa obučavanja modela.

Glavni nedostatak višejezičnih modela je da oni prave odvojene vektorske prostore za svaki podržani jezik što rezultuje time da ne postoji međusobno „razumevanje“ među jezicima, odnosno da se vektori semantički sličnih reči iz različitih jezika u vektorskom prostoru reči izgrađenih modela nalaze udaljeni jedni od drugih (mBERT). U cilju rešavanja uočenog problema, razvijeni su unakrsno-jezički modeli (XLM) koji na efikasan način rešavaju problem međusobnog „razumevanja“ među jezicima u modelima izgrađenim nad ovakvom arhitekturom (slika 4.8).

Transformer modeli, zahvaljujući svojoj sposobnosti da prepoznaju značenje reči na osnovu konteksta iz velikih korpusa tekstualnih podataka nad kojima su obučeni, imaju primenu na raznovrsne zadatke u oblasti računarske lingvistike [132]. Na primeru tekstualne sekvence u srpskom jeziku *„Student je završio projekat iz mašinskog učenja koristeći neuronske mreže.“*, modeli zasnovani na transformer arhitekturi mogu rešavati zadatke kao što su:



Slika 4.8: *Razlike u razumevanju između jezika u mBERT i XLM arhitekturama*

- Klasifikacija teksta – koristi se enkoder arhitektura, u kojoj se ulaz *„Student je završio projekat iz mašinskog učenja...“* može klasifikovati kao tema *„Obrazovanje i nauka“*, ili se tekstualna sekvenca može analizirati za prepoznavanje sentimenta (pozitivan, negativan ili neutralan).
- Prevođenje teksta – koristi se enkoder-dekoder arhitektura u kojoj ulaz na srpskom jeziku *„Student je završio projekat...“* može dati izlaz na engleskom jeziku *„The student completed a project...“*.
- Sažimanje teksta – koristi se enkoder-dekoder arhitektura u kojoj ulaz *„Student je završio projekat iz mašinskog učenja koristeći neuronske mreže.“* postaje *„Student je završio ML projekat.“*.
- Generisanje teksta – uz pomoć dekoder arhitekture, ulaz *„Student je završio...“* može generisati nastavak *„...projekat koristeći neuronske mreže“*.
- Rekonstrukcija teksta – uz pomoć autoenkoder arhitekture, ulaz *„Student je završio projekat...“* je rekonstruisan u izlaznu sekvencu *„Student je završio projekat...“* sa ciljem što manjeg gubitka informacija. Autoenkoder model uči strukturu rečenica, prepoznaje zavisnosti među rečima (subjekat, objekat, predikat) i postavlja temelj za složenije zadatke poput prevođenja, sažimanja ili generisanja teksta.

5. Predstavljanje teksta

5.1. Prevođenje teksta u vektorski numerički prostor

Tehnike predstavljanja teksta imaju za cilj da konvertuju tekstualne podatke u numeričke vektore koje algoritmi mašinskog učenja mogu da obrađuju. Razvijeno je više različitih tehnika za predstavljanje teksta u vektorskom obliku, od kojih su najznačajnije:

- **Vreća reči (eng. *Bag-of-Words, BoW*)** – je jednostavna i široko korišćena tehnika koja tekst predstavlja kao skup reči sa pripadajućim frekvencijama ili binarnim indikatorima pojavljivanja. Glavno ograničenje ovog pristupa je zanemarivanje redosleda i međuzavisnosti reči u tekstu.
- **Frekvencija termina - inverzna frekvencija dokumenata (eng. *Term frequency-Inverse document frequency, Tf-Idf*)** – je proširenje BoW tehnike kojom se svakoj reči, odnosno terminu, dodeljuju težine na osnovu njene učestalosti u tekstualnom dokumentu i celokupnom korpusu [172]. Tehnika Tf-Idf kombinuje dve komponente: frekvenciju pojavljivanja termina u okviru pojedinačnog dokumenta (eng. *Term Frequency*, *Tf*) i inverznu frekvenciju dokumenta (eng. *Inverse Document Frequency*, *Idf*), koja meri u koliko se dokumenata iz kolekcije određeni termin pojavljuje [187]. Osnovna ideja iza ove tehnike jeste da se veća težina dodeljuje terminima koji su učestali u konkretnom dokumentu, ali se retko pojavljuju u ostalim dokumentima korpusa, čime se postiže bolja diskriminativnost. Na ovaj način, Tf-Idf omogućava identifikovanje termina koji su relevantni za sadržaj određenog dokumenta, ali istovremeno doprinose razlikovanju tog dokumenta od drugih dokumenata u korpusu. Težine termina u Tf-Idf vektorskoj reprezentaciji dokumenta se izračunavaju korišćenjem formule predstavljene jednačinom 5.1:

$$\begin{aligned} df_{t,D} &= |\{d \in D : t \in d\}| \\ Idf_{t,D} &= \log\left(\frac{N}{df_{t,D}}\right) \\ Tf\text{-}Idf_{t,d,D} &= Tf_{t,d} \times Idf_{t,D} \end{aligned} \tag{5.1}$$

  gde je:

  - D skup dokumenata u korpusu,
  - $N = |D|$ broj dokumenata u korpusu,
  - $df_{t,D}$ je frekvencija dokumenata u korpusu u kojima se pojavljuje termin t,
  - $Idf_{t,D}$ inverzna frekvencija dokumenata u korpusu koji sadrže termin t,
  - $Tf_{t,d}$ je frekvencija termina t u dokumentu d.

  Više vrednosti Tf-Idf težina ukazuju da je reč karakteristična za dokument, dok niže ukazuju na njenu manju važnost za dokument, odnosno da reč nije diskriminatorna za dokument ili je karakteristična za više dokumenata (ceo korpus).
- **Ugnježdeni vektori reči (eng. *word embeddings, Embd*)** – jesu numerički vektori u neprekidnom vektorskom prostoru koji čuvaju semantičke i sintaksičke sličnosti između

reči. Ovi vektorski prostori jesu rezultat obučavanja modela dubokog učenja nad velikim korpusima tekstualnih podataka sa ciljem da se nauče kontekstualni odnosi između reči. Postoje dve glavne vrste ugnježdenih vektora [210]:

- **Statički vektori** – imaju fiksne vrednosti dodeljene svakoj reči koje se nakon njihovog određivanja, odnosno obučavanja modela, ne menjaju. Statički vektori su trenirani na velikim korpusima teksta iz kojih prikupljaju informacije o zavisnostima koje postoje među rečima i njihovom zajedničkom pojavljivanju. To znači da u nekom smislu predstavljaju reč na osnovu svih konteksta u kojima se pojavila tokom treniranja, kao i da će reči iz sličnih konteksta imati sličnije vektorske reprezentacije. Iako statički vektori ne menjaju svoju reprezentaciju u zavisnosti od trenutnog konteksta unutar rečenice, oni prepoznaju prosečnu kontekstualnu upotrebu reči kroz čitav korpus. Na primer, vektor za reč *kraj* će biti težinska sredina između svih značenja te reči (predlog u značenju *pored*, imenica u značenjima *deo grada* i *završetak*) koja su viđena tokom treniranja. Napoznatiji primeri izgrađenih modela statičkih vektorskih reprezenatacija jesu *Word2Vec* [129], *Glove* [151] i *FastText* [26].
- **Dinamički vektori** – se menjaju u zavisnosti od konteksta u kojem se reč pojavljuje [94]. Ovi vektori se generišu u realnom vremenu iz skrivenih stanja poslednjeg sloja modela transformer arhitekture za svaku reč u ulaznoj sekvenci. Skrivena stanja su visokodimenzionalni vektori koji sadrže kontekstualne informacije o svakoj reči (odnosno tokenu), naučene tokom obučavanja modela transformer arhitekture. Pored ugnježdenih vektora reči, iz modela transformer arhitekture je moguće kreirati i ugnježdene vektore tekstualnih sekvenci računanjem srednje vrednosti vektora pojedinačnih reči koje se pojavljuju u sekvenci, korišćenjem vektorske reprezentacije specijalnog [CLS] tokena oznake kraja sekvence ili specijalizovanih arhitektura napravljenih za tu namenu kao što su *Sentence-BERT (SBERT)* [166] i *Universal Sentence Encoder (USE)* [219].

5.2. Tehnike normalizacije teksta

Tehnike normalizacije teksta se koriste za standardizaciju i transformaciju tekstualnih podataka u konzistentan format. Neke uobičajene tehnike normalizacije teksta jesu:

- **Standardizacija sadržaja** – uključuje transliteraciju teksta, uklanjanje razmaka i višestrukih praznina, segmentaciju teksta, popravku kodiranja karaktera, restauraciju dijakritika, uklanjanje metapodataka, standardizaciju vremenskih formata i korekciju drugih nedoslednosti u tekstualnom sadržaju. Standardizacija pokekad podrazumeva i ujednačavanje veličine tekstualnih karaktera (velika ili mala slova), koja u drugim slučajevima korišćenja, može prouzrokovati pravopisne greške, te je njeno korišćenje delimično zavisno od zadatka koji se rešava.
- **Provera i ispravljanje pravopisa** (eng. *spell-check*) – podrazumeva identifikovanje i ispravljanje pravopisnih grešaka prema sintaktičkim pravilima jezika na kojem je tekst napisan.
- **Određivanje vrste reči (eng. *Part of Speech Tagging, PoS-Tagging*)** – predstavlja proces dodeljivanja gramatičkih kategorija, kao što su imenica, glagol, pridev, prilog i druge, svakoj reči u tekstu na osnovu njenog značenja i konteksta. Najčešće korišćene šeme za obeležavanje vrste reči (eng. *Part of Speech, PoS*), prilagođene različitim jezicima, uključuju *Penn Treebank* šemu, koja koristi 36 osnovnih kategorija (na primer, NN

za imenicu, VB za glagol, JJ za pridev, RB za prilog) [127], kao i *Universal Dependencies* šemu koja definiše 17 univerzalnih kategorija (na primer, NOUN za imenicu, VERB za glagol, ADJ za pridev, ADV za prilog) [142].

- **Morfološka normalizacija** – je transformacija flektivnih oblika reči na zajednički osnovni oblik [96]. Prema načinu na koji se sprovodi ova tehnika, razlikujemo dve osnovne metode:
  - **Stemovanje** (eng. *stemming*) – je redukovanje flektivnih oblika reči uklanjanjem sufiksa, odnosno odsecanjem krajeva reči, čime nastaje koreni oblik reči (eng. *stem*). Nedostatak ove metode je da koreni oblik nije uvek validna reč ili ima drugačije značenje od izvorne reči u datom jeziku.
  - **Lematizacija** (eng. *lemmatization*) – je proces kojim se reči svode na njihov osnovni oblik, odnosno lemu reči (eng. *lemma*), koja je definisana u rečniku. Za razliku od stemovanja, koje koristi statistička pravila za mehaničko odsecanje krajeva reči i pri tome uglavnom zanemaruje lingvističke karakteristike, lematizacija se oslanja na morfološke rečnike i primenjuje morfološku analizu kako bi identifikovala gramatički ispravan osnovni oblik svake reči. [191]. Za primenu lematizacije neophodno je prethodno odrediti PoS obeležje reči koja se želi lematizovati. Princip PoS obeležavanja i postupka lematizacije prikazan je na sledećem primeru tekstualne sekvence na srpskom jeziku:
    „Kako ovog jutra putujete do posla?“
    („*Kako*“, „*kako*“, SCONJ), („*ovog*“, „*ovaj*“, DET), („*jutra*“, „*jutro*“, NOUN), („*putujete*“, „*putovati*“, VERB), („*do*“, „*do*“, ADP), („*posla*“, „*posao*“, NOUN), („*?*“, „*?*“, PUNCT)

- **Uklanjanje stop reči** (eng. *stop words*) – je proces uklanjanja gramatički funkcionalnih reči koje nemaju značaja za razumevanje sadržaja teksta, koje karakteriše značajno veća učestalost pojavljivanja u odnosu na preostale (nefunkcionalne) reči u tekstu. Ove reči obično pripadaju vrstama reči član, predlog, zamenica, veznik i druge, a neki od primera su engleske reči „*the*“, „*and*“ i „*in*“, ili srpske reči „*na*“, „*od*“ i „*u*“ [180].

- **Uklanjanje šuma iz teksta** – predstavlja proces eliminacije tokena i oznaka koje sadrže specijalne karaktere ili poseduju specijalna značenja, kao što su interpunkcijski znakovi, brojevi i druge simboličke oznake koje nisu od značaja za semantičku analizu tekstualnog sadržaja. Iako takvi tokeni mogu nositi određeno značenje, kao što su internet adrese ili domenski specifične oznake, njihovo prisustvo često može negativno uticati na tačnost modela. Posebnu kategoriju ovih tokena čine imenovani entiteti, kao što su imena lica, organizacija, geografskih lokacija i drugih kategorija, čije značenje nije relevantno za određene klasifikacione zadatke, zbog čega se često maskiraju, odnosno zamenjuju univerzalnim oznakama [35].

- **Tokenizacija** – predstavlja podelu teksta na tokene koji čine osnovne gradivne jedinice za dalje korake u obradi teksta. Token može biti reč ili deo reči, odnosno pojedinačni karakter (eng. *Character, Chr*), a koji određujemo prema potrebama algoritma mašinskog učenja koji će se primeniti za rešavanje postavljenog zadatka. Tokeni, prema potrebama zadatka, mogu uključivati znake interpunkcije zajedno sa ostalim alfa-numeričkim karakterima ili ih posmatrati kao posebne tokene. U Pajton biblioteci *nltk* postoje implementacije čitave grupe različitih tokenizatora u okviru *tokenize* paketa[10] koji su korišćeni u okviru ovog istraživanja za tokenizaciju teksta korišćenjem različitih pristupa, kao što su:

[10] https://www.nltk.org/api/nltk.tokenize.html

– *word_tokenize* – metoda za tokenizaciju teksta koja se zasniva na Penn Treebank modelu[11];

– *wordpunct_tokenize* – metoda koja koristi jednostavan regularan izraz oblika `\w+|[^\w\s]+` za pronalaženje granica među tokenima;

– *WhitespaceTokenizer* – tokenizator koji granice tokena pronalazi u prazninama i oznakama za nove linije;

– *TweetTokenizer* – specijalizovan tokenizator koji je dizajniran da emotikone i heš oznake (eng. *hashtags*), čije je pojavljivanje karakteristično za tekstualne sadržaje sa društvenih mreža, prepozna kao samostalne tokene;

– *RegexpTokenizer* – tokenizaciju vrši na osnovu prilagođenog regularnog izraza, čime se mogućnosti za definisanje granica tokena značajno povećavaju;

– *PunktSentenceTokenizer* – omogućava pronalaženje granica između rečenica u tekstualnom sadržaju.

Pored toga, osnovna gradivna jedinica teksta može biti i sekvenca od N uzastopnih tokena (eng. *Sequence of N adjacent tokens, Ngram*), koja u specijalnom slučaju kada je $N = 1$ zapravo predstavlja pojedinačan token, odnosno sekvencu od jednog tokena (eng. *Sequence of a single token, Unigram*). Ngram tokeni se mogu posmatrati i kao pokretni prozor dužine N nad uređenom listom pojedinačnih tokena prepoznatih u sekvenci. Pojava transformer arhitekture modela, uvela je i inovativne metode tokenizacije teksta za korišćenje u ovim modelima (pogledati odeljak 4.3.3).

Izbor tehnike za normalizaciju teksta zavisi od ML algoritma i pratećeg pristupa za vektorizaciju teksta koji se koriste, zadatka koji se rešava, kao i morfoloških karakteristika jezika na kojem je napisan tekst. Tako na primer BoW tehnika vektorizacije teksta zahteva prethodno standardizaciju sadržaja i svođenje reči na osnovni oblik, dok kod Embd primena svođenja reči na osnovni oblik nije uvek nužna, već se primenjuje ukoliko se ekperimentalnim testovima nad konkretnim podacima potvrde poboljšanja u krajnjim rezultatima [33]. Razlog za uočeno ponašanje leži u činjenici da Embd, kao i arhitekture modela koje se na njih oslanjaju, prepoznaju kontekstualnu zavisnost između reči, koja se primenom ove tehnike potencijalno može narušiti.

Tabela 5.1: *Stepen zavisnosti primene tehnika normalizacije u odnosu na korišćeni algoritam, zadatak i jezik tekstualnog sadržaja*

Tehnika	Algoritam	Zadatak	Jezik
Standardizacija sadržaja	DZ	DZ	Z
Ispravljanje pravopisnih grešaka	NZ	NZ	Z
Morfološka normalizacija	Z	NZ	Z
Uklanjanje stop i specijalnih reči	NZ	Z	Z
Tokenizacija	Z	Z	Z

Tabela 5.1 prikazuje stepen uočenih zavisnosti, prema kojoj se razlikuju nezavisni (NZ), delimično zavisni (DZ) i potpuno zavisni (Z) stepeni uticaja. Uklanjanje stop i specijalnih reči direktno je zavisno od zadatka koji se rešava, u kojem se definišu skupovi reči koje bi trebalo ukloniti na način da se ne naruše performanse klasifikacije. U pojedinim slučajevima se određeni skupovi reči ne uklanjaju, već se maskiraju univerzalnim tokenima. Ovakva transformacije se primenjuje kada je potrebno očuvati informaciju o prisustvu

[11] https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html

tokena u tekstu, ali njihovo značenje nije važno za rešavanje datog zadatka. Maskiranje se primenjuje i u slučajevima kada je neophodno obezbediti zaštitu privatnosti podataka fizičkih i pravnih lica. Sa druge strane, ispravljanje pravopisnih grešaka predstavlja neophodnu tehniku koja se primenjuje nezavisno od algoritma za klasifikaciju teksta ili zadatka koji se rešava, i naročito se koristi u obradi tekstova iz neformalnog govora, kao što su tekstovi preuzeti sa društvenih mreža. Standardizacija sadržaja predstavlja postupak koji je usko povezan sa konkretnim jezikom, i podrazumeva prilagođavanje napisanog teksta u konzistentnu formu kako bi se obezbedila tačnija obrada. Morfološka normalizacija zavisi od karakteristika jezika tekstualnog sadržaja i samog algoritma, a primenjuje se sa ciljem da se različiti morfološki oblici reči svedu na njihov osnovni oblik u cilju povećanja tačnosti modela. Tokenizacija, kao fundamentalna tehnika u pripremi podataka, zavisi od jezika i zadatka, jer način segmentacije teksta na manje jedinice (tokene) može značajno uticati na performanse daljih koraka obrade teksta i klasifikacije.

6. Moralnost i emocije u klasifikaciji teksta

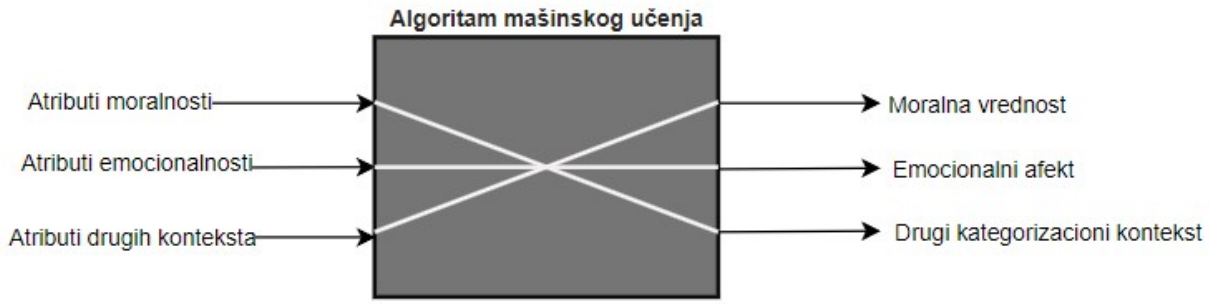
6.1. Prepoznavanje moralnosti i emocionalnosti u tekstu

Jedan pristup u uočavanju koncepata moralnosti i emocionalnosti u NLP je pravljenje semantičke baze znanja moralnih i emocionalnih reči, koji se zatim mogu koristiti za analizu teksta u pogledu prisutnosti moralnog i emocionalnog afekta. Istraživači su do sada za engleski jezik napravili brojne leksikone reči na osnovu psiholoških teorija koje ih podržavaju. Jedna od takvih je MFT teorija [66] koja daje objašnjenje kako pojedinci koriste moralne procene u različitim životnim situacijama (pogledati odeljak 2.1). Sa druge strane, složene i promenljive emocionalne reakcije mogu biti predstavljene leksikonima zasnovanim na različitim teorijama o emocijama i načinima njihovog poimanja [154]. Najrasprostranjeniji je pristup korišćenja teorija o osnovnim emocijama koje podrazumevaju postojanje malog broja osnovnih ljudskih emocija (Ekmanov model - šest emocija, Plutčikov model - osam emocija), koje su različita društva i kulture univerzalno prihvatili (pogledati odeljak 2.2). Leksikoni koji su izgrađeni nad ovim teorijama fokusiraju se na unapred definisane kategorije i reči koje su im pridružene preko indikatora pripadnosti (0 ili 1), relativnog intenziteta u odnosu na preostale reči u leksikonu ili verovatnoće pripadnosti iz numeričkog opsega [0, 1] [137, 82]. Drugi način obeležavanja reči predstavlja emocije kao vektorske prostore sa dve, tri ili četiri dimenzije [170]. Ove dimenzije se definišu na sledeći način:

- valentnost – prijatnost stimulusa;
- uzbuđenje – intenzitet emocije izazvane stimulusom;
- dominacija – stepen kontrole koji stimulus ostvaruje.

Sentiment se može smatrati posebnim slučajem ovog pristupa u merenju emocionalnog afekta u kome se valentnost, koja meri koliko je neka reč prijatna ili neprijatna, često koristi kao mera sentimenta. U leksički zasnovanim modelima, afektivno značenje reči je uglavnom fiksno, jer ne uzima u obzir kontekst upotrebe, dijalekt ili demografske karakteristike učesnika.

Drugi pristup modelovanja složenih afektivnih procesa jeste izgradnja računarskih modela korišćenjem tehnika mašinskog učenja koji su sposobni da simuliraju moralno i emocionalno rasuđivanje i donošenje odluka [4]. Ovi modeli mogu biti projektovani tako da oponašaju procese koje ljudska bića koriste kada donose odluke, kao što je odmeravanje posledica različitih postupaka, razmatranje različitih moralnih principa i emocionalnih reakcija i pravljenje kompromisa između konfliktnih vrednosti. Ovi modeli se mogu obučiti na velikim skupovima podataka o moralnim dilemama i emocionalnim reakcijama u stvarnom svetu, a zatim se koristiti za predviđanje kako će ljudska bića reagovati u novim situacijama [74, 10]. Atributi koji mere moralni ili emocionalni afekt se mogu koristiti kao zavisni ili nezavisni atributi. U prvom slučaju, klasifikacioni model se pravi da predvidi određenu moralnu ili emocionalnu kategoriju (ili više njih) iz skupa unapred definisanih kategorija. U drugom slučaju, atributi koji nose emocionalni ili moralni afekt se koriste kao nezavisni atributi klasifikacije za određivanje nekog novog kategorizacionog konteksta [185]. Treća moguća situacija upotrebe moralnih i emocionalnih atributa u procesu klasifikacije je da se oni koriste unakrsno, odnosno da se nezavisni atributi koji mere svojstvo moralnosti koriste za predviđanje emocionalne kategorije i obrnuto (pogledati sliku 6.1).



Slika 6.1: *Atributi emocionalnosti i moralnosti kao nezavisni i zavisni atributi u algoritmu mašinkog učenja*

Treći pristup je hibridni i u njemu se kombinuju tehnike prethodna dva pristupa. Hibridni pristup obično daje modele sa boljim performansama, ali je zahtevniji za razvoj i evaluaciju [178, 164].

6.2. Uspostavljeni načini obeležavanja

Raznolikost u šemama i korišćenim tehnikama za obeležavanje emocionalnog i moralnog aspekta jezika proizilazi iz složenosti prepoznavanja ovih jezičkih koncepata. Uspostavljene anotacione šeme se razlikuju u pogledu emocionalnih kategorija, načina kvantifikovanja (kontinuirano u odnosu na kategoričko) i definisanog kriterijuma kategorizacije. U naučnim istraživanjima su primenjivane različite tehnike za kreiranje leksikona i tekstualnih korpusa obeleženih kategorijama emocionalnog afekta ili moralne vrednosti. Neke od najčešće korišćenih načina obeležavanja novih jezičkih resursa su:

- **Ručno obeležavanje** – podrazumeva korišćenje ljudskih resursa za ručno dodeljivanje obeležja rečima, frazama ili tekstualnim segmentima prema unapred definisanim kriterijumima i uputstvima za obeležavanje. Anotatori su često eksperti iz datog domena (na primer lingvisti ili psiholozi u slučaju emocionalnosti ili moralnosti) ili su angažovani preko platformi za iznajmljivanje ljudskih resursa (eng. *crowdsourcing*). Anotacioni koraci zahtevaju da je rad anotatora potpuno nezavistan i višestruk, odnosno da jedan primerak obeležavanja pregleda i obeležava više anotatora (> 1). Na kraju anotacionog procesa, za svaki primerak koji je bio predmet obeležavanja se vrši usaglašavanje i izračunavanje konsenzus obeležja na osnovu utvrđenog algoritma (na primer maksimalni broj glasova, skaliranje metodom najbolji-najgori u grupi i drugih).
- **Obeležavanje zasnovano na leksičkim resursima** – podrazumeva korišćenje već postojećih, proverenih i potvrđenih leksičkih resursa (semantičkih baza znanja, leksikona i drugih) i uspostavljenih skupova pravila za dodeljivanje obeležja novim skupovima tekstualnih podataka. Ova vrsta tehnike obeležavanja se obično naziva i poluautomatsko obeležavanje, jer su obeležja dobijena na ovaj način često neprecizna i zahtevaju ručnu proveru, ali se njome ubrzava proces obeležavanja u odnosu na potpuno ručnu metodu.
- **Obeležavanje korišćenjem matematičkih metoda i tehnika mašinskog učenja** – podrazumeva korišćenje računarskih algoritama kojima se automatskim putem određuju neophodna obeležja. Ove tehnike obuhvataju algoritme za automatsko prevođenje, algoritme za izračunavanje kontinuiranih numeričkih obeležja i algorit-

me za pronalaženje harmonizovanih obeležja iz skupova pridruženih obeležja. U ovu grupu tehnika, takođe, spadaju i najnovije metode mašinskog učenja, kao što su aktivno učenje ili učenje putem transfera znanja, kojima se postiže efektivno obeležavanje podataka nad novim zadacima ili jezicima. U najnovije vreme, određene tehnike korišćenja velikih jezičkih modela, kao što su tehnike bez obučavanja (eng. *zero-shot*) ili kratkog obučavanja (eng. *few-shot*) su pokazale izuzetne performanse u obeležavanju podataka automatskim putem korišćenjem već izgrađenih modela i eventualno malog broja već obeleženih primera. Korišćenjem tehnika mašinskog učenja, obeležavanje podataka se značajno unapređuje, ali i dalje ostaje potreba za proverom obeležja dobijenih na ovaj način.

6.3. Istaknuti primeri obeleženih podataka

U ovom odeljku predstavljene su neke od semantičkih baza znanja (leksikona) i obeleženih tekstualnih korpusa koji su svojom pojavom napravili značajan iskorak i doprineli razvoju istraživanja emocionalnih i moralnih aspekata jezika u oblasti računarske lingvistike. Ovi resursi su razvijeni za engleski jezik i predstavljaju globalni standard zbog široke zastupljenosti engleskog jezika u globalnoj komunikaciji. Značaj koji ovi resursi imaju odnosi se na inicijalni razvoj, poboljšanje pokrivenosti (kod leksikona), proširenje tipova podataka ili anotacionih kategorija, kao i načina na koji se dolazi do obeleženih podataka korišćenjem najsavremenijih ML metoda i NLP tehnika. Njihova važnost za druge jezike, uključujući srpski, ogleda se u prilici da se slične metodologije primene u različitim jezičkim i kulturnim kontekstima, što omogućava istraživanje specifičnih i univerzalnih karakteristika emocionalnih i moralnih aspekata jezika.

6.3.1 Emocionalni leksikoni i korpusi

Leksikoni emocionalnosti

- **ANEW**[12] (eng. *Affective Norms for English Words*) – sadrži afektivne vrednosti za približno 1,000 engleskih reči, omogućavajući analizu emocionalnog izražavanja u tekstualnim podacima. Afektivne vrednosti su raspoređene u tri emocionalne dimenzije: valentnost (prijatnost), uzbuđenje (aktivacija) i dominacija (kontrola) [29]. **XANEW**[13] je proširena verzija ANEW leksikona, koja sadrži afektivne vrednosti za približno 14,000 engleskih reči, čime je značajno premašena pokrivenost originalnog ANEW leksikona [211].
- **SenticNet**[14] – je semantički resurs koji uključuje koncepte, entitete i sa njima povezane afektivne informacije. Ovaj resurs pruža sveobuhvatan prikaz afektivnog stanja tako što prepoznate reči i koncepte povezuje sa različitim afektivnim dimenzijama, koje uključuju sentiment, valentnost i uzročnike promene sentimenta [34, 155].
- **LIWC** (eng. *Linguistic Inquiry and Word Count*) – je softverski alat za analizu teksta koji analizira pisani ili govorni jezik na osnovu psiholoških i lingvističkih kategorija. Ovaj alat pruža informacije o lingvističkim i psihološkim dimenzijama datog teksta kategorizacijom reči u različite unapred definisane kategorije, kao što su pozitivne-/negativne emocije, kognitivni procesi, društveni odnosi i lični problemi [150].

[12] https://osf.io/y6g5b/wiki/anew/
[13] https://github.com/JULIELab/XANEW
[14] https://sentic.net/

- **WordNet leksikon afekata (eng. *WordNet-Affect Lexicon, WNA*)**[15] – je leksikon koji povezuje afektivne informacije, od kojih značajan deo zauzimaju upravo emocije, sa kontekstima Princeton verzije WordNet leksikona (eng. *Princeton WordNet Lexicon, PWN*) leksikona na engleskom jeziku [56]. Leksikon WNA proširuje PWN, dodeljivanjem afektivnih obeležja sinsetima, koji predstavljaju skupove sinonima iz jednog konteksta. Svaki sinset u WNA je obeležen afektivnim obeležjima, koji uključuju polaritete sentimenta i specifične kategorije emocija. Leksikon WNA pruža hijerarhijsku strukturu afektivnih kategorija koje su organizovane u taksonomiju[16] sa širokom i veoma specifičnom emocionalnom kategorizacijom. Kreiranje WNA uključivalo je kombinaciju ručnog obeležavanja i automatizovanih metoda. Anotatori su ručno obeležavali sinsete na osnovu njihovog ličnog doživljaja emocija, vođeni psihološkim teorijama i sopstvenim iskustvom. Automatizovane metode propagirale su afektivna obeležja sa obeleženih sinseta ka neobeleženim, koristeći karakteristike hijerarhijske strukture PWN leksikona [194].
- **NRC Grupa Leksikona**[17] – obuhvata kolekciju široko korišćenih resursa u analizi osećanja i otkrivanju emocija u okviru računarske lingvistike koje je razvio Nacionalni istraživački cantar Kanada (eng. *National Research Council Canada, NRC*).
  - **NRC leksikon asocijacija između reči i emocija (eng. *The Word-Emotion Association Lexicon, EmoLex*)** – je jedan od najviše korišćenih leksikona u okviru NRC grupe koji 14,154 engleskih reči kategoriše u 8 diskretnih emocionalnih kategorija prema Plutčikovom modelu [153] i 2 kategorije sentimenta (pozitivan i negativan) [136].
  - **NRC leksikon valentnosti, uzbuđenja i dominacije (eng. *The NRC Valence, Arousal and Dominance*, NRC-VAD)** – se fokusira na pronalaženje pozicije reči u trodimenzionalnom vektorskom prostoru, pri čemu se rečima dodeljuju numeričke vrednosti prema ovim afektivnim dimenzijama [135].
  - **NRC leksikon sentimenta heš oznaka (eng. *The NRC Hashtag Sentiment Lexicon*)** – je specijalizovan za analizu sentimenta sadržaja sa društvenih mreža, koji meri vrednost sentimenata u heš oznakama koje se na društvenim mrežama koriste za grupisanje tematskih sadržaja [138].
  - **NRC leksikon emocionalnih intenziteta (eng. *The Emotion Intensity Lexicon, EmoInt*)** – je razvijen koristeći pristup skaliranja reči poređenjem [101]. Ovaj leksikon sadrži 5,814 reči, od kojih je svaka povezana sa realnim vrednostima za intenzitet četiri primarne emocije: *anger*, *fear*, *joy* i *sadness* [137]. Pristup skaliranja reči poređenjem podrazumeva da se anotatorima prikaže skup od obično četiri reči, nakon čega im se postavlja zadatak da identifikuju reč sa najvišim i najnižim intenzitetom za određenu emociju (pogledati tabelu 6.1). Ova metoda pruža precizniju i pouzdaniju anotaciju intenziteta emocija u poređenju sa tradicionalnim metodama ručnog rangiranja, pri čemu se konačni skor računa kao odnos razike pojavljivanja reči na najboljoj, odnosno najgoroj poziciji i ukupnog broja pojavljivanja reči u skupu za obeležavanje (pogledati jednačinu 6.1).

$$score_{r_i} = \frac{|najbolja_{r_i}| - |najgora_{r_i}|}{|r_i|}, i = 1, 2, ...N \tag{6.1}$$

[15] https://wndomains.fbk.eu/wnaffect.html
[16] https://www.gsi.upm.es/ontologies/wnaffect/img/wnaffect.svg
[17] https://saifmohammad.com/WebPages/lexicons.html

Tabela 6.1: *Obeležavanje korišćenjem metode skaliranja reči poređenjem u cilju pronalaženja najbolje-najgore reči u grupi za postavljeni uslov*

Grupa Reči (n=4)				Obeležja Poređenja	
$reč_1$	$reč_2$	$reč_3$	$reč_4$	najbolja reč	najgora reč

Korpusi obeleženi emocionalnim kategorijama

- **EmotionLines**[18] – je korpus konverzacionih podataka, preciznije dijaloga iz televizijskih serija i filmova, u kojem je svaki iskaz (eng. *utterance*) obeležen kategorijom emocionalnog afekta [86]. Anotaciona šema prepoznaje sledeća obeležja: *joy, sadness, anger, surprise, fear, disgust, i neutral*, pri čemu je korišćen višeznačni pristup obeležavanja. Korpus je naknadno unapređen dodavanjem audio i video zapisa postojećim tekstualnim sadržajima. Novonastali, višemodalni korpus podataka, pod nazivom **MELD** (eng. *Multimodal EmotionLines Dataset*) - sadrži emocionalna obeležja i za pridodate audio i video sadržaje, čime se omogućava istraživanje emocionalnog izražavanja preko više modaliteta njihovog izražavanja (tekst, audio, video), a time se potencijalno poboljšava i tačnost njihovog prepoznavanja [156].
- **GoEmotions**[19] – je korpus od 58,000 poruka preuzetih sa socijalne mreže Redit koje su ručno obeležene sa 27 emocionalnih kategorija [49]. Prema svojoj veličini, ali i prema vrsti obeležavanja, kreiranje *GoEmotions* korpusa je bilo značajno za dalji razvoj istraživanja emocionalnosti sa aspekata psihologije i računarske lingvistike. Unapređenjem u *GoEmotions* anotacionoj šemi u odnosu na prethodne pristupe, koja definiše 27 različitih emocionalnih kategorija, otvaraju se mogućnosti za finije razlikovanje emocionalnog izražavanja u tekstualnim podacima. Za razliku od osnovnog skupa sa 6, odnosno 8 emocija, na kojima se zasniva većina postojećih anotacionih šema koje prepoznaju radost (eng. *joy*) kao jedinu pozitivnu emociju, u *GoEmotions* anotacionoj šemi je definisano 12 pozitivnih, 11 negativnih, 4 neodređenih i 1 neutralna emocionalna kategorija koje su grupisane na sledeći način:

  > ***pozitivne (eng. positive)**: [zabavljanje (eng. amusement), uzbuđenje (eng. excitement), radost (eng. joy), ljubav (eng. love), želja (eng. desire), optimizam (eng. optimism), briga (eng. caring), ponos (eng. pride), divljenje (eng. admiration), zahvalnost (eng. gratitude), olakšanje (eng. relief), odobravanje (eng. approval)]*
  > ***negativne (eng. negative)**: [strah (eng. fear), nervoza (eng. nervousness), kajanje (eng. remorse), neprijatnost (eng. embarrassment), razočaranje (eng. disappointment), tuga (eng. sadness), žalost (eng. grief), gađenje (eng. disgust), ljutnja (eng. anger), iritacija (eng. annoyance), neodobravanje (eng. disapproval)]*
  > ***neodređene (eng. ambiguous)**: [spoznaja (eng. realization), iznenađenje (eng. surprise), radoznalost (eng. curiosity), konfuzija (eng. confusion)]*

  Pored grupisanja u osnovne sentimentalne kategorije, *GoEmotions* kategorije se takođe mogu grupisati u 6 osnovnih Ekmanovih kategorija emocionalnosti, čime se omogućava analiza emocionalnosti na višem nivou granularnosti i poređenje sa postojećim modelima napravljenim korišćenjem ove anotacione šeme:

  > ***ljutnja (eng. anger)**: [ljutnja (eng. anger), nerviranje (eng. annoyance), neodobravanje (eng. disapproval)],*
  > ***gađenje (eng. disgust)**: [gađenje (eng. disgust)],*

[18] https://live.european-language-grid.eu/catalogue/corpus/5149
[19] https://huggingface.co/datasets/google-research-datasets/go_emotions

***strah (eng. fear)**: [strah (eng. fear), nervoza (eng. nervousness)],*
***radost (eng. joy)**: [radost (eng. joy), zabava (eng. amusement), odobravanje (eng. approval), uzbuđenje (eng. excitement), zahvalnost (eng. gratitude), ljubav (eng. love), optimizam (eng. optimism), olakšanje (eng. relief), ponos (eng. pride), divljenje (eng. admiration), želja (eng. desire), briga (eng. caring)],*
***tuga (eng. sadness)**: [tuga (eng. sadness), razočaranje (eng. disappointment), neprijatnost (eng. embarrassment), žalost (eng. grief), kajanje (eng. remorse)],*
***iznenađenje (eng. surprise)**: [iznenađenje (eng. surprise), spoznaja (eng. realization), zbunjenost (eng. confusion), radoznalost (eng. curiosity)]*

Drugačija pregrupisavanja u okviru emocionalnih kategorija bi potencijalno omogućila poklapanja sa Plutčikovim skupom osnovnih emocionalnih kategorija koji dodatno uključuje emocionalne kategorije poverenje (eng. *trust*) i iščekivanje (eng. *anticipation*). Tako bi na primer kategorije *excitement* i *curiosity* mogle odgovarati kategoriji iščekivanje, dok bi kategorije *approval* i *realization* odgovarale kategoriji poverenje, čime bi se omogućilo preslikavanje i povezivanje sa skupom emocionalnih kategorija koje prepoznaje Plutčikov model [153].

- **XED**[20] – predstavlja višejezični korpus konvezacionih podataka koji je obeležen u emocionalne kategorije prema Plutčikovom modelu. Deo korpusa na engleskom i finskom jeziku je ručno obeležen u emocionalne afektivne kategorije, dok su tekstovi na preostalih 30 jezika obeleženi korišćenjem postojeće paralelizacije među tekstovima [118], a obeležja koje su na ovaj način dobijena proverena korišćenjem BERT tansformer modela i postupkom projekcije kategorija transferom znanja na svaki jezik [145].

6.3.2 Leksikoni i korpusi moralnih vrednosti

Leksikoni moralnosti

Razvijeni leksikoni reči sa pridruženim moralnim vrednostima se prema svom dizajnu uglavnom oslanjaju na Teoriju o moralnim osnovama [66], među kojima se izdvajaju sledeći:

- **MFD**[21] (eng. *Moral Foundations Dictionary*) – je leksikon moralnih reči koji su kreirali autori MFT, sa okvirno 336 korena engleskih reči, obeleženih da pripadaju jednoj ili više dualnih moralnih kategorija definisanih ovom teorijom [67]. Leksikon je u verziji **MFD2.0** proširen na preko 2,100 engleskih reči, od kojih je svaka dodeljena jednoj ili više moralnih kategorija definisanih u okviru MFT: *care*, *fairness*, *loyalty*, *authority*, *sanctity*. Svaka od ovih kategorija dodatno je razvrstana prema pripadajućem sentimentu na vrline (eng. *virtue*) i mane (eng. *vice*) [62].
- **MoralStrength**[22] – je leksikon koji proširuje inicijalni MFD leksikon na okvirno 1,000 engleskih lema korišćenjem PWN sinseta. Kao unapređenje pristupa, autori uvode dodeljivanje intenziteta moralne vrednosti svakoj lemi u leksikonu, čime je zamenjen raniji model zasnovan na kategoričkim obeležjima. Detaljnom proverom, koje je uključilo merenje sličnosti između ugnježdenih vektorskih reprezentacija reči, potvrđeno je unapređenje performasi prilikom korišćenja leksikona koji je napravljen na ovaj način (F_1 vrednost je povećana za +25.2%) u odnosu na prethodne verzije leksikona moralnih vrednosti [12].

[20] https://github.com/Helsinki-NLP/XED
[21] https://osf.io/ezn37/
[22] https://github.com/oaraque/moral-foundations

- **eMFD**[23] (eng. *Extended Moral Foundations Dictionary*) – je leksikon reči sa pridruženim moralnim vrednostima, koji je, nasuprot prethodnim leksikonima, razvijen primenom tehnika obrade prirodnih jezika iz tekstualnih korpusa. U pažljivo projektovanom anotacionom procesu, koji je uključio veći broj anotatora za obeležavanje manjih tekstualnih korpusa, kao i predloženim načinom prepoznavanja moralne vrednosti, razvijena je nova metodologija za automatsko kreiranje rečnika moralnosti. Autori su u svom radu pokazali validnost primenjenog pristupa uspešnim prepoznavanjem moralnih vrednosti u tekstovima pomoću napravljenog rečnika. Nad obeleženim tekstovima se najpre primenjuju tehnike normalizacije teksta (uklanjaju se specijalne reči koje uključuju interpunkcijske znakove i imenovane entitete), određivanja sentimenta tekstualnih sadržaja, kreiranja vokabulara iz tekstualnog korpusa i pronalaženja karakterističnih reči za svaku od dihotomnih moralnih kategorija [82]. Rečima u leksikonu dodeljene su izračunate verovatnoće pripadnosti svakoj od suprotstavljenih moralnih vrednosti, kao i procenjeni intenzitet sentimenta zasnovan na njihovoj učestalosti u tekstovima obeleženim odgovarajućom moralnom kategorijom.

Korpusi obeleženi moralnim kategorijama

- **MFTC**[24] (eng. *Moral Foundations Twitter Corpus*) – predstavlja jedan od najznačajnih doprinosa u domenu obeleženih korpusa kategorijama moralnih vrednosti. Korpus MFTC sadrži 35,108 poruka preuzetih sa društvene mreže Tviter koje su obeležene kategorijama moralnih vrednosti [81]. Poruke su ciljano preuzimane iz konverzacija u kojima bi autori mogli pokazati svoja moralna gledišta, kao što su prirodne katastrofe, politička ili aktuelna društvena pitanja. Korpus je ručno obeležen u deset kategorija moralnog sentimenta koje prepoznaje Teorija o moralnim vrednostima. Korpus MFTC javno je objavljen i koristi se za multidisciplinarna metodološka istraživanja načina izražavanja moralnih stavova u konverzacionim tekstovima.
- **MFRC**[25] (eng. *Moral Foundations Reddit Corpus*) – predstavlja kolekciju od 16,123 poruka preuzetih sa društvene mreže Redit, dajući značajan doprinos u kreiranju korpusa obeleženih kategorijama moralnih vrednosti. Ovaj korpus tekstualnih podataka iz delimično izmenjenog domena prepoznaje sledećih osam moralnih kategorija: *care, proportionality, equality, purity, authority, loyalty, thin morality, implicit/explicit morality*. Autori kategoriju *fairness* iz Teorije o moralnim osnovama, dele u dve nove kategorije *proportionality* i *equality* i koriste u svom radu u cilju boljeg prepoznavanja i razlikovanja pravednosti u primenjenoj proceduri i jednakosti u dobijenim rezultatima. U anotacionoj šemi ovog korpusa nalazi se i kategorija *implicit/explicit morality* kojom se obeležava eksplicitno ili implicitno iskazivanje moralnog gledišta, kao i kategorija *thin morality* za obeležavanje moralnih stavova koji nisu striktno povezani sa šest osnovnih moralnih vrednosti koje ova šema obeležavanja definiše [202].
- **MIC**[26] (eng. *Moral Integrity Corpus*) – je korpus dijaloga nastalih u alatima za automatsko generisanje odgovora u kome su iskazi alata obeleženi kategorijama moralnih vrednosti uz pridružena pravila identifikovanog moralnog zaključivanja. Primena ovog korpusa se pronalazi u razrešavanju i predupređivanju moralnih nesporazuma, kao i razumevanje moralnog rezonovanja tekstova koji nastaju u automatskim alatima. U korpusu MIC nalazi se oko 38,000 parova pitanja i odgovora, koji su obeleženi

[23] https://osf.io/vw85e/
[24] https://osf.io/k5n7y/
[25] https://huggingface.co/datasets/USC-MOLA-Lab/MFRC
[26] https://github.com/SALT-NLP/mic

korišćenjem približno 99,000 različitih pravila zaključivanja. Autori u svom radu pokazuju da se većina predloženih pravila mogu automatski generisati pomoću neuronskih jezičkih modela sa prihvatljivom tačnošću. Ovaj korpus je značajan za razumevanja implicitinih moralnih pretpostavki i merenje integriteta napravljenih sadržaja u alatima za automatsko generisanje odgovora [226].

7. Predložena metodologija klasifikacije

7.1. Opšte o metodologiji

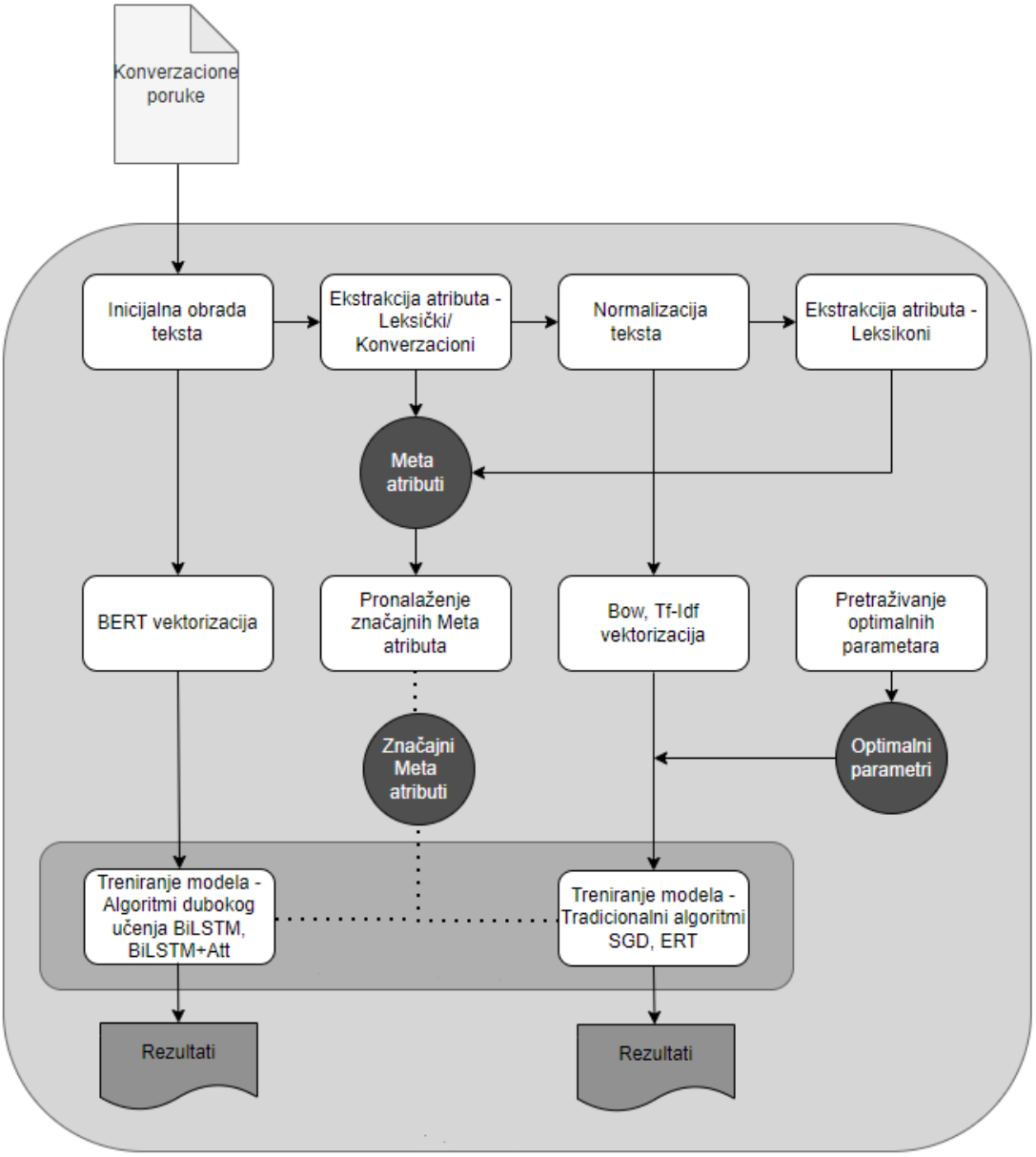
Klasifikacija konverzacionih tekstova pokazuje specifičnosti koje je razlikuju od klasifikacije standardnih tekstualnih sadržaja, kao što su pojedinačne rečenice, članci ili dokumenti, a koje proizilaze iz karakteristične strukture konverzacionih tekstova i njihovog dinamičkog, interaktivnog i često neformalnog jezičkog karaktera. Konverzaciona poruka, pored tekstualne sekvence, sadrži i potpis koji nosi prateće podatke o poruci, kao i vezu ka prethodnoj poruci u nizu (pogledati odeljak 3). Postojanje veze između poruka utiče i na postojanje kontekstualne zavisnosti sadržaja poruke od prethodnih poruka u konverzacionom nizu. Iz tog razloga, u procesu klasifikacije nad određenim zadacima, neophodno je uključiti i kontekstualne informacije u kojima se poruka nalazi.

Razvijena metodologija za klasifikaciju poruka konverzacionog niza predstavlja sistematski pristup koji uključuje niz eksperimentalnih koraka usmerenih na izbor i optimizaciju različitih tehnika obrade prirodnih jezika i algoritama mašinskog učenja. Na slici 7.1 je predstavljen dijagram toka koraka predložene metode koji obuhvata standardne korake u klasifikaciji teksta kao što su priprema i obrada podataka, kao i izbor algoritma za obučavanje modela mašinskog učenja. U okviru standardnih koraka u klasifikaciji teksta, razvijena metoda za dati zadatak i skup podataka nad kojim se izgrađuje model predlaže i čitav skup eksperimentalnih koraka u cilju pronalaženja optimalnog:

- Načina predstavljanja tekstualnog sadržaja u vektorskom obliku;
- Načina normalizacije i tokenizacije teksta;
- Skupa pridruženih atributa klasifikacije;
- Klasifikacionog algoritma;
- Skupa hiperparametara algoritma.

Ključni aspekti razvijene metodologije za analizu konverzacionog niza obuhvataju detaljan izbor tehnika za vektorizaciju teksta i metoda tokenizacije, što direktno utiče na izbor i efikasnost algoritama mašinskog učenja. Metoda započinje eksperimentalnim odabirom jedne od tehnika vektorizacije kao što su BoW, Tf-Idf ili naprednije tehnike poput BERT Embd, koja omogućava transformaciju teksta u numerički vektorski prostor uz efikasno pronalaženje semantičkog značenja reči i fraza unutar tekstualne sekvence. Izbor jedinice za tokenizaciju teksta, kao što su delovi reči, celokupne reči ili Ngram takođe igra važnu ulogu u procesu obrade teksta, jer određuje granularnost informacija koje se analiziraju i uslovljen je karakteristikama jezika i podacima koji se analiziraju. Odabrana tehnika za vektorizaciju direktno utiče na izbor tehnika za normalizaciju teksta, što je važno za smanjenje varijabilnosti ulaznih podataka. Normalizacija podrazumeva izbor odgovarajućih tehnika kao što su restauracija dijakritika, lematizacija reči, uklanjanje stop reči i specijalnih karaktera. U daljem tekstu, za BoW vektorske reprezentacije koristiće se sledeće oznake:

- BoW za reprezentacije nad pojedinačnim tokenima (Unigram),
- BoW-Ngram za N-gram reprezentacije nad tokenima ($N \in [1, k], k > 1$),
- BoW-Ngram-Chr za N-gram reprezentacije nad karakterima ($N \in [1, k], k > 1$).



Slika 7.1: *Dijagaram toka predložene metode za klasifikaciju konverzacionih poruka*

Analogno, Tf-Idf reprezentacije biće obeležene sa Tf-Idf, Tf-Idf-Ngram i Tf-Idf-Ngram-Chr. Za vektorsku reprezentaciju zasnovanu na ugrađenim vektorima (Embd) koristi se tokenizacija i težine određene BERT modelom, u oznaci BERT-Embd.

U okviru predložene metodologije uključeno je i izdvajanje dodatnih atributa iz bogate strukture konverzacionih poruka, pored atributa u okviru vektorskog predstavljanja tekstualnog sadržaja. Dodatni atributi se mogu ekstrahovati iz sadržaja poruke, ali i iz njenog potpisa i veza sa drugim porukama u okviru konverzacionog niza, čime se obogaćuje analiza i omogućava bolje razumevanje sadržaja i dinamike konverzacije. Izbor pridruženih atributa se vrši na osnovu njihove relevantnosti za zadatak koji se rešava i vrste konverzacionih podataka. Korišćenje dodatnih atributa naročito dobija na značaju u analizi konverzacionih tekstova kod kojih je uočljiva upotreba neformalnog jezika, žargona, skraćenica i emotikona, što može otežati standardne pristupe u obradi jezika koji se koriste za formalne tekstove. Dodatno, konverzacije često uključuju izražavanje ličnih mišljenja, osećanja i stavova. Analiza sentimenta, prepoznavanje emocija ili iskazanih moralnih vrednosti su stoga

često važni ulazni atributi ili zadaci za rešavanje u klasifikaciji konverzacionih tekstova, što nije uvek slučaj sa drugim vrstama tekstualnih sadržaja. Nakon konstruisanja atributa (pogledati naredni odeljak 7.2), vrši se raspoređivanje značaja pojedinih atributa teksta u klasifikaciji konverzacionih tekstova u kategorije definisane zadatkom i predlaže se njihovo dalje korišćenje ili odbacivanje u procesu klasifikacije. U okviru ovog istraživanja posebna pažnja biće posvećena analizi uticaja atributa moralnog rasuđivanja i emocionalnih reakcija na tačnost izgrađenih klasifikacionih modela.

U središtu metodologije nalaze se algoritamske arhitekture za klasifikaciju konverzacionih poruka koje, u zavisnosti od strukture ulaznih podataka, obuhvataju pristupe sa sledećim ulazima:

E1 Pojedinačna poruka (Msg) – osnova;

E2 Pojedinačna poruka sa pridruženim dodatnim informacijama (Msg-Ext);

E3 Pojedinačna poruka sa odgovorima iz iste konverzacione grane (Brch);

E4 Pojedinačna poruka sa odgovorima iz iste konverzacione grane sa pridruženim informacijama svakoj poruci (Brch-Ext).

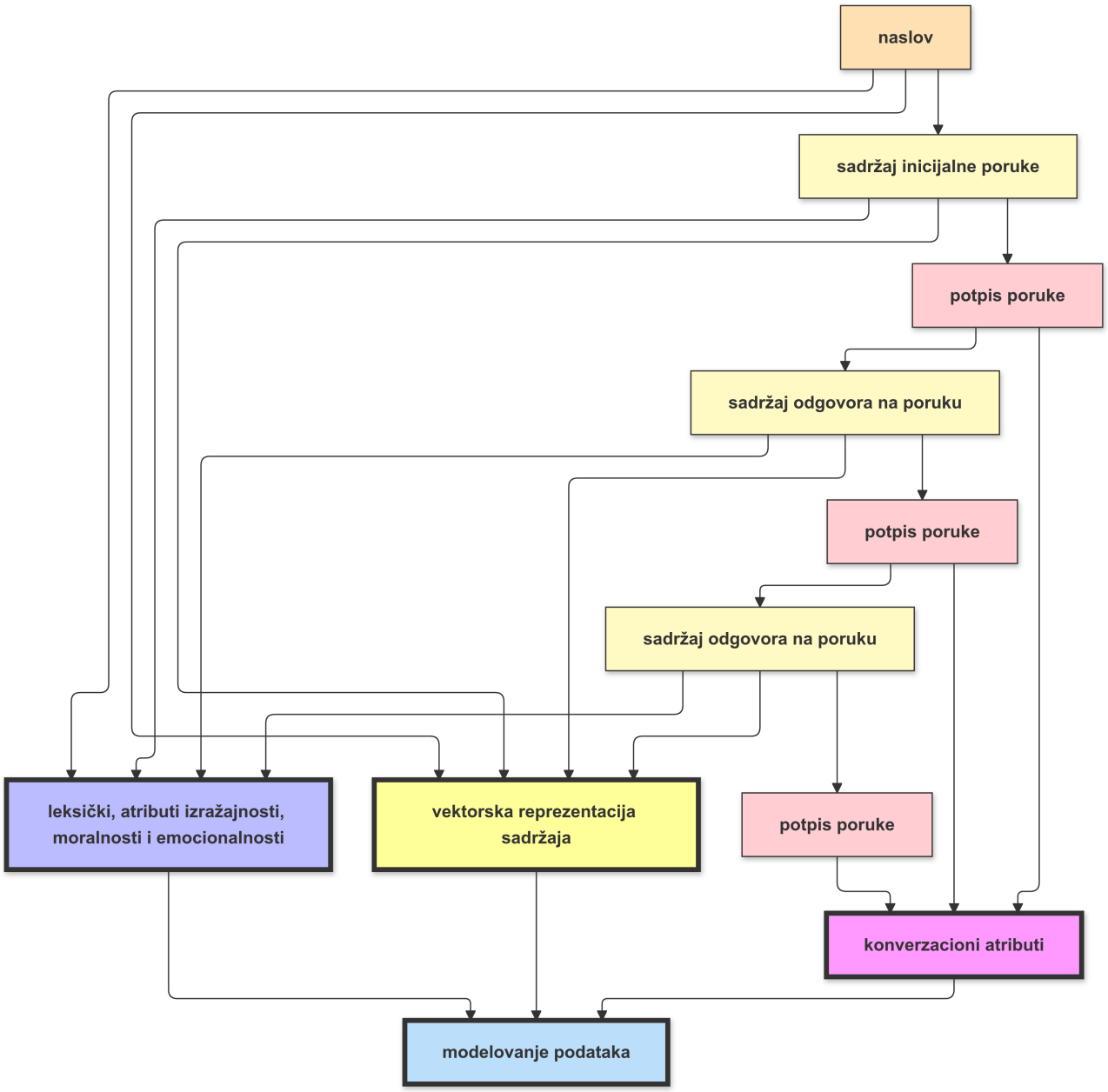
U okviru predložene metode za analizu konverzacionih poruka koristiće se nekoliko različitih algoritama i pristupa mašinskog učenja. Prvu grupu sačinjavaju algoritmi koji koriste optimizacionu tehniku SGD, a izbor samog algoritma zavisiće od konkretnog zadatka, odnosno podataka na kojima se primenjuje metoda (pogledati odeljak 4.2.1). Drugu grupu sačinjavaju algoritmi zasnovani na ansambl metodama, a cilj njihove upotrebe je efikasno merenje uspešnosti razvijene metode u odnosu na druge predložene pristupe i rešenja (ERT, pogledati odeljak 4.2.2). Algoritmi mašinskog učenja se biraju iz predefinisanog skupa algoritama i dodatno parametrizuju za specifične zadatke klasifikacije konverzacionih podataka.

Pored izbora algoritma, prilagođavanje hiperparametara za algoritme mašinskog učenja je esencijalno za izgradnju optimalnog modela. Ovaj korak uključuje eksperimentalno određivanje vrednosti kao što su broj paralelnih stabala, slojeva u mreži, broj neurona u svakom od slojeva, kao i regularizacione parametre koji su važni za obezbeđivanje sposobnosti generalizacije i sprečavanje preprilagođavanja izgrađenog modela nad podacima koji su korišćeni za obuku. Kroz ovako osmišljen pristup, teži se razvoju opšte metode koja može efikasno da se primeni na različite zadatke, od analize sentimenta pojedinačne poruke do automatskog prepoznavanja teme razgovora na čitavoj konverzaciji, čime se omogućava izgradnja robustnih modela za klasifikaciju širokog spektra konverzacionih podataka i njihovih strukturnih organizacija.

7.2. Pridruženi atributi klasifikacije

Pridruženi atributi (eng. *Associated Meta Attributes, Meta*) klasifikacije mogu imati značajnu ulogu u modelovanju teksta pomoću algoritama mašinskog učenja. Meta atributi mogu pomoći u analizi leksičke i semantičke strukture poruke i pridodaju se izgrađenim vektorskim reprezentacijama nad tekstualnim sadržajem (BoW, Tf-Idf, Embd). Na primer, autori u radu [197] koriste dodatne atribute kao što su oznake aspekata u kombinaciji sa BoW reprezentacijom za poboljšanje analize sentimenta, dok se u radu [112] uključuju informacije iz leksikona i specijalizovane oznake kako bi poboljšali klasifikaciju imenovanih entiteta. Međutim, uključivanje dodatnih atributa može povećati složenost modela, uneti

šum ako su atributi nepouzdani ili dovesti do preprilagođavanja modela. Studija [85] pokazuje da usložavanje modela dodatnim atributima može smanjiti performanse klasifikacije, dok studija [47] ilustruje negativan uticaj lošeg kvaliteta vrednosti atributa na performanse klasifikacije. Najbolje prakse iz naučnih studija ukazuju na pažljiv odabir relevantnih atributa, primenu tehnika normalizacije i standardizacije atributa, eksperimentisanje sa različitim kombinacijama atributa, kao i korišćenje tehnika regularizacije modela u toku obuke da bi se prevazišao uočeni problem. U okviru ovog istraživanja, kreirana je široka lista leksičkih, konverzacionih, atributa izražajnosti, emocionalnosti i moralnosti u cilju poboljšanja performansi klasifikacije konverzacionih poruka. Konačan izbor atributa značajnih za klasifikaciju odabran je matematičkim metodama za pronalaženje značajnih atributa (pogledati odeljak 4.1.3) za svaki zadatak i skup podataka pojedinačno. U okviru ovog istraživanja se predlaže opšta lista Meta atributa za korišćenje u klasifikacionim zadacima konverzacionih poruka:



Slika 7.2: *Ekstrakcija atributa iz različitih segmenata poruke na primeru konverzacione grane poruka sa naslovom*

- **Leksički atributi (eng. *Lexical Attributes, LexAtr*)** prikupljaju prebrojavanja i međusobne odnose između reči koje se mogu pronaći u naslovu i sadržaju poruke. Kompletna lista leksičkih atributa se može pronaći u prilogu A, tabeli A.1. Klasifikacija teksta se u velikoj meri oslanja na takve atribute i stoga pretpostavljamo da će leksičke karakteristike teksta doprineti u zadacima klasifikacije konverzacionih tekstova. Ovoj grupi atributa pripadaju:
  - **Sintaktički atributi (eng. *Syntactic Attributes, SntAtr*)** Sintaksičke karakteristike uključuju karakteristike zasnovane na strukturi reči i rečenica u sadržaju. Neke od ovih karakteristika zasnivaju se na izračunavanju:
    * broja karaktera (Chr), slogova (eng. *syllables*), reči (eng. *words*) i rečenica (eng. *sentences*),
    * složenih reči (eng. *difficult words*) – reči koje sadrže više od dva sloga,
    * praznina (eng. *spaces*) i redova (eng. *lines*),
    * broja imeničkih fraza (eng. *noun phrases*) i različitih vrsta reči,
    * prosečnog broja slogova u reči (eng. *average syllable per word, ASPW*):
    $$ASPW = \frac{\#syllables}{\#words} \tag{7.1}$$
    * prosečnog broja slogova u rečenici (eng. *average syllable per sentence, ASPS*):
    $$ASPS = \frac{\#syllables}{\#sentences} \tag{7.2}$$
    * gustine reči (eng. *words density*):
    $$words\_density = \frac{\#words}{\#spaces} \tag{7.3}$$
    * gustine rečenica (eng. *sentences density*):
    $$sentence\_density = \frac{\#sentences}{\#lines} \tag{7.4}$$

    u tekstualnom sadržaju.
  - **Atributi indikacije (eng. *Indication Attributes, IndAtr*)**: omogućavaju razumevanje kontekstualnih različitosti u kategorisanim porukama i smanjuju verovatnoću pogrešnih klasifikacija što vodi ka efikasnijem razvrstavanju poruka prema njihovom sadržaju i svrsi. Na primer, u klasifikaciji poruka elektronske pošte, poslovni indikator i indikator skraćenica, omogućavaju modelu da bolje prepozna i razume specifične termine i skraćenice koje su uobičajene u poslovnoj komunikaciji. Poslovni indikator je numerički atribut koji predstavlja broj pojavljivanja poslovnih reči i fraza („*ugovor*“, „*budžet*“ i druge) u odnosu na celokupan sadržaj. Poslovni termini se pronalaze pomoću specijalizovanih leksikona poslovnih termina[27] koji sadrži izraze i terminologiju koja se koristi u poslovnim konverzacijama. Skraćenice su u sadržaju identifikovane korišćenjem leksikona skraćenica i akronima[28], kao i pomoću regularnih izraza koji su pridodati u cilju poboljsanja preciznosti njihovog pronalaženja u sadržaju poruke.

[27] https://www.businessballs.com/glossaries-and-terminology/business-thesaurus-290/
[28] https://abbreviations.yourdictionary.com/

– **Atributi interpunkcije (eng. *Punctuation Attributes, PncAtr*)** mere prisustvo tačaka, upitnika, uzvičnika, heš („#“) i referentnih znakova („@“) u odnosu na ukupan broj interpunkcijskih znakova koji se nalaze u sadržaju poruke.

– **Atributi imenovanih entiteta (eng. *NER-based Attributes, NERAtr*)** jesu numerički pokazatelji prisustva imenovanih entiteta (eng. *named entities*, *NE*) u sadržaju. Predstavljaju broj pojavljivanja ličnih imena, imena organizacija, reči koje sadrže brojeve, reči u engleskom jeziku označene kao konektori („*in*“, „*the*“, „*all*“, „*for*“, „*and*“, „*on*“, „*but*“, „*at*“, „*of*“, „*to*“, „*a*“), imena meseci i dana, kao i ispravnih naziva elektronske i URL adrese u odnosu na ukupan broj reči u sadržaju.

- **Konverzacioni atributi (eng. *Conversational Attributes, ConAtr*)** se izdvajaju iz potpisa poruke koji sadrži informacije o primaocima ili tipovima njihovih elektronskih adresa. U tom cilju, koristićemo rečnik besplatnih domena elektronskih adresa[29]. Razmera besplatnih domena je odnos broja pojavljivanja besplatnih domena u odnosu na sve domene koji se mogu pronaći u potpisu poruke. Koherentnost domena primaoca je funkcija koja izračunava koherentnost među primaocima u odnosu na domen njihove elektronske adrese. Na primer, ovaj atribut ukazuje da li svi domeni primaoca pripadaju podrazumevanom domenu kompanije, spoljašnjim domenima ili su domeni primaoca mešovite strukture.

- **Atributi izražajnosti (eng. *Expressional Attributes, ExpAtr*)** obuhvataju informacije kao što su čitljivost teksta, subjektivnost i polaritet. Subjektivnost i polaritet su zasnovani na implementaciji Pajton biblioteke *TextBlob*[30]. Polaritet je predstavljen realnim brojem koji se nalazi u opsegu od [-1, 1], pri čemu 1 označava pozitivan, a -1 negativan sentiment. Subjektivne rečenice se odnose na lično mišljenje, emociju ili sud, dok se objektivne odnose na činjenične informacije. Subjektivnost je predstavljena realnim brojem koji se nalazi u opsegu od [0, 1], pri čemu vrednosti bliže broju 1 označavaju subjektivniji kontekst.

  Atributi čitljivosti izračunati su korišćenjem uspostavljenih mera za izračunavanje čitljivosti teksta na engleskom jeziku [99], kao što su:

  – Automatski indeks čitljivosti (eng. *Automatic Readability Index, ARI*) koristi broj karaktera po reči i broj reči po rečenici kako bi procenio uzrast (školski nivo) potreban za razumevanje teksta [182]. Veća vrednost znači da je tekst složeniji, a data je sledećom jednačinom:

$$ARI = 4.71(\frac{\#characters}{\#words}) + 0.5(\frac{\#words}{\#sentences}) - 21.43 \tag{7.5}$$

  – Flešova ocena lakoće čitanja (eng. *Flesch Reading Ease Score, FRES*) ocenjuje tekst na skali od 0 do 100, pri čemu više vrednosti označavaju lakše tekstove [60]. Zasniva se na prosečnom broju slogova po reči i reči po rečenici i data je sledećom jednačinom:

$$FRES = 206.835 - 1.015(\frac{\#words}{\#sentences}) - 84.6(\frac{\#syllables}{\#words}) \tag{7.6}$$

  – Linsearova mera čitljivosti (eng. *Linsear Write Metric, LWM*) je razvijena za procenu čitljivosti tehničkih i vojnih dokumenata. Koristi broj „lakih“ i „teških“ reči u

[29] https://github.com/villvhite/freemail
[30] https://tektblob.readthedocs.io/en/dev/

uzorku od 100 reči kako bi izračunala obrazovni nivo potreban za razumevanje teksta (pogledati algoritam 7.1).

Algoritam 7.1: Linsearova mera čitljivosti (*LWM*)

```
LWM
    inputs: text (tekst na engleskom jeziku od 100 reči)
    output: r (obrazovni nivo), LWM
    foreach word ∈ text do
        if #syllabeles(word) < 3 then r += 1 ;
        else r += 3 ;
    r = r / sentences(text) if r > 20 then LWM = r / 2 ;
    else LWM = r / 2 - 1 ;
return LWM;
```

- **Atributi emocionalnosti (eng. *Emotional Attributes, EmoAtr*)** se izračinavaju korišćenjem EmoLex leksikona koji sadrži 10,170 reči kategorizovanih u 8 emocionalnih kategorija određenih prema Plutčikovom modelu osnovnih emocija [137]. Za izračunavanje emocionalnih atributa koristi se Pajton *NRCLex*[31] biblioteka koja proširuje izvorni leksikon na 27,000 reči, oslanjajući se na bazu PWN sinonima. Atributi se na nivou tekstualne sekvence izračunavaju pronalaženjem emocionalno afektivnih reči i agregacijom njihovih obeležja. Na osnovu ovako projektovanog pristupa, svakoj tekstualnoj sekvenci se dodeljuje 8 atributa koji predstavljaju relativnu zastupljenost reči iz pojedinačnih emocionalnih kategorija.
- **Atributi moralnosti (eng. *Morality Attributes, MorAtr*)** se izračunavaju iz teksta obeležavanjem reči koje imaju potvrđenu povezanost sa nekom od pet moralnih osnova definisanih u okviru MFT teorije: *authority/subversion*, *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *sanctity/degradation* (pogledati odeljak 2.1). Za identifikovanje moralno relevantnih reči u sadržaju poruke korišćen je leksikon *eMFD*[32], kao i prateća Pajton biblioteka za izračunavanje moralnog sentimenta. Svakom tekstualnom sadržaju je dodeljeno pet atributa koji označavaju prosečnu verovatnoću pripadnosti jednoj od pet moralnih osnova i pet atributa koji predstavljaju prosečan intenzitet sentimenta za svaku moralnu osnovu. Dodatno, izračunat je i atribut koji predstavlja odnos između broja moralno obeleženih i ukupnog broja reči u tekstu, čime je omogućen uvid u relativnu moralnu izraženost napisanog sadržaja.

Zadatak koji se rešava, vrsta konverzacione poruke i jezik na kom je poruka napisana mogu uticati na relevantnost i uspešnost kreiranja atributa iz određenih predloženih grupa. Atributi iz leksičke grupe (LexAtr) su delimično zavisni (DZ) od vrste poruke i jezika na kome su poruke napisane. Delimična zavisnost se može prikazati na primeru naslova poruke koji nije prisutan u svim vrstama konverzacionih tekstova. Drugi primer jesu imenički izrazi ili teške reči čije je identifikovanje visoko zavisno (Z) od morfologije određenog jezika. U potpisu konverzacione poruke se nalaze vremenske odrednice, vrsta medija koja se koristi za slanje poruke, lična imena učesnika u konverzaciji i druge informacije koje su specifične za određenu vrstu konverzacije. U zavisnosti od zadatka koji se rešava, ove informacije mogu biti od važnosti za unapređenje tačnosti klasifikacije. Atribute koji struktuiraju ove informacije u predloženoj metodi nazivamo konverzacioni atributi (ConAtr). Ova grupa atributa pokazuje visoku zavisnost od zadatka i vrste konverzacionih poruka

[31] https://pipi.org/project/NRCLex/
[32] https://github.com/medianeuroscience/emfd

Tabela 7.1: *Sumarni opis grupa pridruženih atributa sa stepenom zavisnosti izračunavanja u odnosu na zadatak, vrstu i jezik konverzacione poruke*

Grupa	Podgrupa	Opis	Zadatak	Vrsta	Jezik
LexAtr	SntAtr	Pomažu u analizi leksičke kompleksnosti i strukture jezika poruke	NZ	DZ	DZ
	IndAtr	Mere pojavljivanje specijalnih indikatora u sadržaju koji su karakteristični za zadatak koji se rešava	Z	Z	Z
	PncAtr	Analiziraju upotrebu interpunkcijskih znakova, čime pružaju uvid u komunikacione stilove poruka	NZ	NZ	NZ
	NERAtr	Omogućavaju identifikaciju i analizu relevantnih imenovanih entiteta unutar poruka, što pomaže u prepoznavanju ličnih i specifičnih informacija	Z	NZ	Z
ConAtr		Mere karakteristike komunikacije, što pomaže u razumevanju socijalne dinamike u interakcijama	Z	Z	DZ
ExpAtr		Fokusiraju se na ocenu čitljivosti i subjektivnosti sadržaja, što omogućava razumevanje koliko je poruka lako dostupna i sentimentalno obojena	NZ	NZ	Z
EmoAtr		Pružaju uvid u emocionalni afekt u sadržaju merenjem broja ili intenziteta afektivnih reči	NZ	NZ	Z
MorAtr		Koriste analizu reči kako bi izmerili prisustvo moralnih vrednosti u poruci preko verovatnoće pripadnosti određenoj kategoriji ili intenziteta moralnog sentimenta	NZ	NZ	Z
Meta		Kombinuje sve atribute, koji zajedno omogućavaju sveobuhvatnu analizu i razumevanje poruka	Z	Z	Z

koje se analiziraju, kao i delimičnu zavisnost od jezika (pogledati tabelu 7.1). Ove osobine atributa ukazuju na promenljivu karakteristiku skupa Meta atributa, odnosno da je njenu konačnu strukturu moguće odrediti na konkretnom zadatku koji se rešava, kao i da je u zavisnosti od vrste konverzacionih poruka i jezika na kojem su poruke napisane.

7.3. Specijalizovane arhitekture algoritama dubokog učenja

Predložene arhitekture neuronskih mreža za rešavanje zadataka klasifikacije konverzacionih tekstova uključuju:

1. BiLSTM nad Msg – koristi dvosmerni BiLSTM za obradu tekstualne sekvence u oba smera.
2. BiLSTM nad Brch – obrađuje nizove poruka tako što svaka poruka prolazi kroz BiLSTM sloj i agregaciju, a zatim se svi izlazi obrađuju kroz dodatne BiLSTM slojeve na nivou sekvence poruka.

Svi predloženi modeli dubokog učenja se zasnivaju na arhitekturi dvostruko povezane LSTM mreže (BiLSTM, odeljak 4.3.1), koja je prevazilaženjem nedostataka iz prethodnih arhitektura (RNN, odeljak 4.3.1; LSTM, odeljak 4.3.1), pokazala dobre performanse u analizi sekvencijalnih podataka u opštem slučaju. Ove arhitekture se mogu primeniti sa i bez mehanizma pažnje (BiLSTM+/-Att, odeljak 4.3.2), koji uključuje usmerenost na relevantne delove teksta tokom klasifikacije. Dodatno, arhitekture se opciono mogu obogatiti Meta atributima klasifikacije (BiLSTM+/-Att+/-Meta, odeljak 7.2), što može poboljšati performanse modela u prepoznavanju i razumevanju složenih obrazaca u tekstovima. Kombinovanjem navedenih tehnika, omogućava se potencijalno postizanje veće tačnosti u klasi-

fikaciji tekstova, kao i bolje razumevanje njihovog sadržaja i pojedinačnih aspekata jezika tih sadržaja. Sve arhitekture koriste *ReLU* aktivacionu funkciju u skrivenim slojevima i σ ili $softmax$ funkcije u izlaznim slojevima, zavisno od zadatka i vrste klasifikacije. Funkcije greške na nivou pojedinačnih instanci u slučaju binarne klasifikacije je binarna unakrsna entropija (eng. *Binary Cross-Entropy, BCE*), prikazana jednačinom 7.7, i kategorička unakrsna entropija (eng. *Categorical Cross-Entropy, CCE*), prikazana jednačinom 7.8.

$$L(y, \hat{y}) = -\left[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right] \tag{7.7}$$

gde je y stvarna oznaka (0 ili 1) i $\hat{y}$ predviđanje modela (verovatnoća od 0 do 1).

$$L(y, \hat{y}) = -\sum_{i=1}^{K} y_i \cdot \log(\hat{y}_i) \tag{7.8}$$

gde je y_i stvarna oznaka (0 ili 1) za klasu i, $\hat{y}_i$ predviđanje modela za klasu i (verovatnoća od 0 do 1) i K ukupan broj klasa.

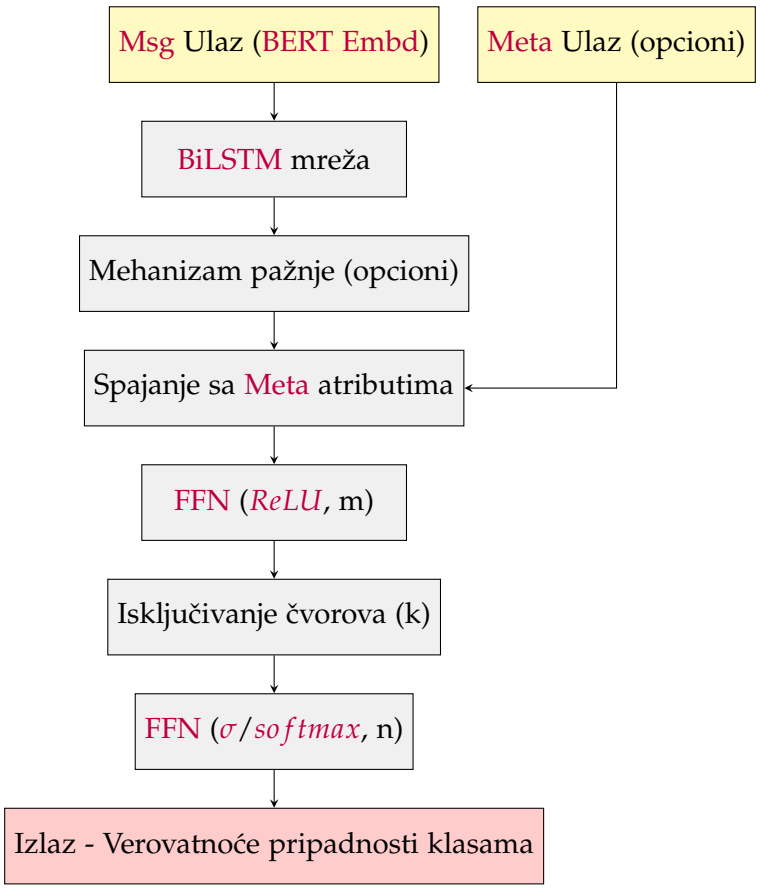
U oznakama eksperimenata, predložene arhitekture, zasnovane na BiLSTM, imaju dodeljene oznake strukture podataka nad kojima su izgrađene, odnosno Msg za arhitekturu nad pojedinačnom porukom i Brch za arhitekturu nad konverzacionom granom poruka.

7.3.1 Pojedinačna poruka

BiLSTM-Msg arhitektura podrazumeva klasifikaciju pojedinačne poruke bez uzimanja u obzir šireg konteksta u kojem se poruka nalazi. Arhitektura u osnovi koristi BiLSTM mrežu i podrazumeva sekvencijalno povezivanje sledećih slojeva neuronske mreže:

1. **Ulazni sloj** – predstavlja ulazni tekstualni niz u kome je svaka reč u tekstualnoj sekvenci predstavljena kao BERT Embd.
2. **BiLSTM sloj** – predstavlja jednu ili više dvosmernih LSTM slojeva koji obezbeđuju uočavanje kontekstualnih zavisnosti iz oba smera tekstualne sekvence. Izlazi iz dvosmernih slojeva se spajaju u cilju stvaranja kombinovane reprezentacije prethodećeg i sledujućeg konteksta za svaku reč.
3. **Opcioni Att sloj** – predstavlja izračunavanje težinskih koeficijenata za svaku reč u nizu na osnovu njenog relevantnosti za zadatak koji se rešava. Izračunati koeficijenti se koriste za izračunavanje težinskog zbirnog izlaza LSTM slojeva.
4. **Opcioni sloj spajanja sa Meta atributima** – podrazumeva opciono uključivanje pridruženih atributa tekstualne sekvence koje mogu pružiti dopunske informacije o tekstualnom sadržaju.
5. **FFN sloj** – podrazumeva jedan ili više potpuno povezanih slojeva neurona, kojima se kao parametri dodeljuju broj neurona i *ReLU* aktivaciona funkcija. U okviru ovog sloja postoji opcioni sloj isključivanja (eng. *dropout*) koji obezbeđuje regularizaciju i smanjenje preprilagođavanja modela sa parametrom za stepen isključivanja.
6. **Izlazni sloj** – predstavlja sloj koji daje verovatnoću pripadnosti određenoj klasi na osnovu obrađene reprezentacije ulazne tekstualne sekvence kroz sve prethodne slojeve.

Svaka tektualna poruka se predstavlja preko vektorskih reprezentacija (BERT Embd) reči. Reprezentacije poruka se prosleđuju nastupajućim BiLSTM slojevima. Izlazi iz BiLSTM slojeva se prosleđuju FFN slojevima i koriste se za konačni klasifikacioni sloj sa σ ili $softmax$ aktivacionom funkcijom, koji predviđa oznaku klase za poruku.

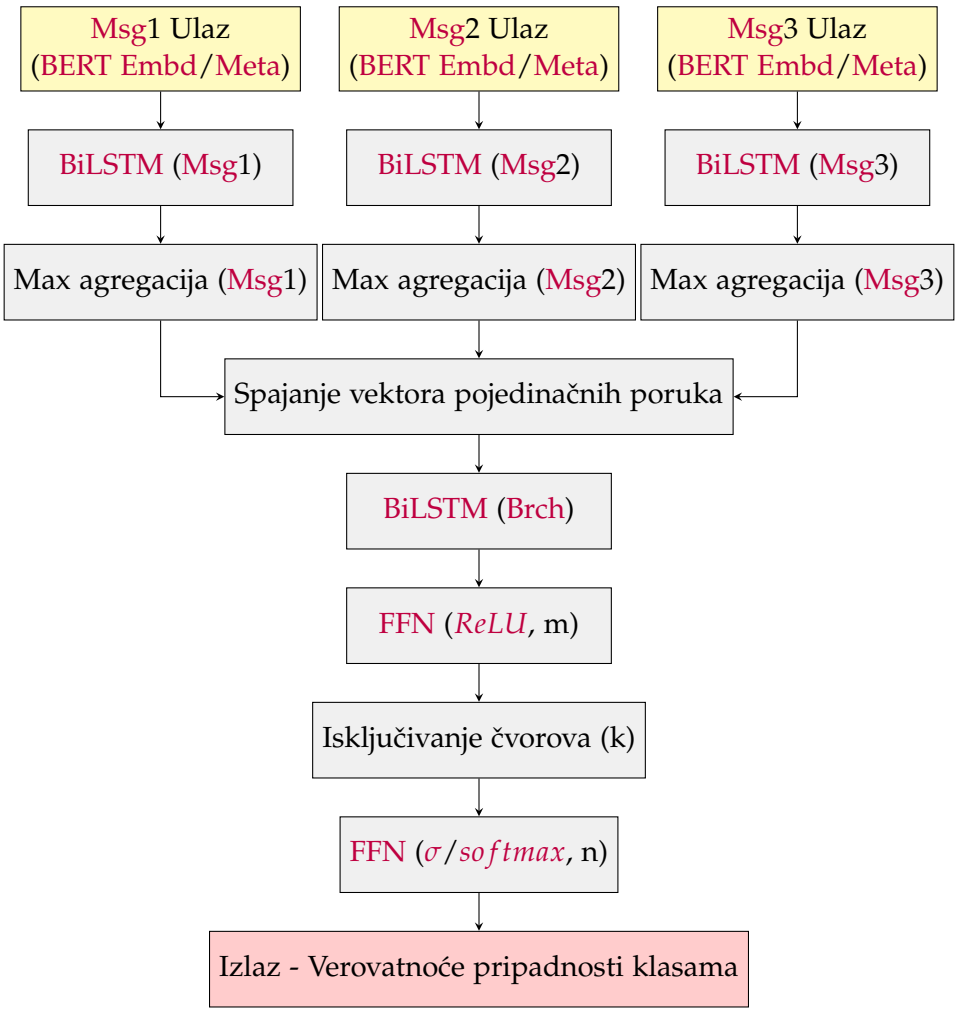


Slika 7.3: *Tok algoritma dubokog učenja nad ulaznom tekstualnom sekvencom pojedinačne poruke*

7.3.2 Konverzaciona grana neposrednih poruka

BiLSTM–Brch arhitektura podrazumeva klasifikaciju celokupne konverzacione grane sa pripadajućim porukama. Arhitektura u osnovi koristi dva nivoa BiLSTM mreža u cilju da pronađe kontekstualne zavisnosti između reči u tekstualnoj sekvenci (nivo reči) i kontekstualne zavisnosti između neposrednih poruka (nivo poruke). Predložena arhitektura podrazumeva sekvencijalno povezivanje sledećih slojeva neuronske mreže:

1. **BiLSTM sloj (reči)** – predstavlja obradu tekstualnih sekvenci pojedinačnih poruka kroz jedan ili više BiLSTM slojeva, uz opciono uključivanje mehanizma pažnje i spajanje sa Meta atributima (pogledati BiLSTM-Msg arhitekturu, korake **1**.-**4**.).
2. **Sloj agregacije (Max)** – predstavlja agregiranje izlaza BiLSTM mreže nad tekstualnim sekvencama pojedinačnih poruka (nivo reči) u cilju smanjenja njihovih dimenzija i mogućnosti obrade u narednom sloju.
3. **BiLSTM sloj (poruke)** – predstavljaju jedan ili više BiLSTM slojeva koji obezbeđuju prepoznavanje kontekstualnih zavisnosti iz oba smera sekvence poruka u konverzacionoj grani.
4. **FFN sloj** – podrazumeva jedan ili više potpuno povezanih slojeva neurona, kojima se kao parametri dodeljuju broj neurona i *ReLU* aktivaciona funkcija. U okviru ovog sloja postoji opcioni sloj isključivanja (eng. *dropout*) koji obezbeđuje regularizaciju i smanjenje preprilagođavanja modela sa dodeljenim parametrom za stepen isključivanja.
5. **Izlazni sloj** – predstavlja sloj koji daje verovatnoću pripadnosti određenoj klasi na osnovu obrađene reprezentacije ulazne tekstualne sekvence kroz sve prethodne sloje-



Slika 7.4: *Tok algoritma dubokog učenja nad ulaznom sekvencom neposrednih poruka na konverzacionoj grani*

ve.

Svaka tektualna poruka se projektuje u odgovarajuću vektorsku reprezentaciju (BERT Embd) koja se zatim obrađuje pomoću nastupajućih BiLSTM slojeva. Vektorska reprezentacija poruke se uči iz vektorskih reprezentacija reči: prva neuronska mreža obrađuje svaku reč ulazne poruke kroz jedan ili više BiLSTM slojeva. Izlazna sekvenca se zatim agregira kako bi se eliminisala zavisnost od dužine poruke, čime se dobijaju vektorske reprezentacije poruka fiksne dužine. Nakon toga, dobijene reprezentacije poruka se prosleđuju drugoj neuronskoj mreži. Jedan ili više BiLSTM slojeva obrađuju ulazne vektore, a završna stanja FFN slojeva se koriste za konačni klasifikacioni sloj sa σ ili $softmax$ aktivacionom funkcijom, koji predviđa oznaku klase na svakom vremenskom koraku, odnosno poruci.

Struktura podataka za obučavanje ovakve arhitekture ima određene specifičnosti, jer zahteva dvostruki proces popunjavanja vektorskih reprezentacija do određene fiksirane dužine. Svaka grupa ulaznih podataka sadrži b grana, pri čemu svaka od njih sadrži promenljiv broj poruka *n_messages*, a svaka poruka sadrži promenljiv broj reči *n_words*. Dakle, grupa ulaznih podataka je predstavljena matricom oblika (b, *max_n_messages*, *max_n_words*), gde *max_n_messages* označava najveći broj poruka u jednoj grani (između b grana), a *max_n_words* označava najdužu poruku (u broju tokena) u svim porukama i granama. Tokeni koji se koriste za popunjavanje su specijalni tokeni za tu namenu (obično praznine) i zanemaruju se u toku obučavanja modela.

Predložena metoda se može primeniti u slučajevima kada je potrebno klasifikovati pojedinačne poruke ili celokupan niz poruka koje pripadaju jednoj grani. Kod klasifikacije celokupnog niza neposrednih poruka, arhitektura se u završnim slojevima delimično menja tako da vrši klasifikaciju svih vremenskih koraka u jednu klasu, umesto svakog koraka pojedinačno. U pojedinim zadacima klasifikacije teksta, postoji potreba za klasifikovanjem celokupnih konverzacija, kao što je na primer provera istinitosti objavljenih glasina na društvenim mrežama [65]. U slučaju klasifikacije celokupne konverzacije, predložena arhitektura se može iskoristiti za klasifikovanje svih njenih grana nad čijim rezultatima se primenjuje metoda većinskog glasanja za dobijanje klasifikacione oznake konverzacije.

7.4. Klasifikacija poruka elektronske pošte u poslovnu i ličnu

Klasifikacija poruka elektronske pošte u kategorije **Poslovna** i **Lična**, u oznaci zadatka **PL**, izvršena je na osnovu predložene metodologije u okviru ovog istraživanja (pogledati odeljak 7.1). Bogata struktura poruka elektronske pošte u kojoj se mogu identifikovati tri glavna segmenta kao što su naslov, sadržaj i prateće informacije (ne primer informacije o primaocima ili vremenskim okvirima), prati generalnu strukturu svih konverzacionih poruka (pogledati poglavlje 3). Automatsko kategorizovanje poruka elektronske pošte omogućava korisnicima da na jednstavan način razdvoje važne poslovne poruke od lične komunikacije, što doprinosi boljoj produktivnosti i smanjenju vremena potrebnog za pronalaženje relevantnih informacija. Ova vrsta klasifikacije se najčešće oslanja na ML algoritme koji analiziraju sadržaj, ton i strukturu poruka kako bi ih pravilno svrstali u odgovarajuće kategorije. Na primer, poruke elektronske pošte u poslovnom kontekstu obično koriste formalni jezik, poslovne termine, fraze i skraćenice, kao što su skraćenice na engleskom jeziku „*ROI*“ ili „*FYI*“, dok su lične poruke uglavnom neformalne i sadrže sentimentalne izraze i oznake. Značaj ove klasifikacije leži u njenoj sposobnosti da olakša upravljanje velikim brojem poruka elektronske pošte, optimizuje vreme, i omogući efikasnu organizaciju poslovne komunikacije, smanjujući pri tome rizik da važne poslovne informacije budu zagubljene među ličnim porukama.

Tabela 7.2: *Karakteristične reči u kategorijama* **Poslovna** *i* **Lična**, *kao i celom korpusu*

Korpus	**Poslovna**	**Lična**	Ceo korpus
$Enron_C$	„*energy*“, „*agreement*“, „*information*“, „*power*“, „*market*“, „*attached*“, „*gas*“, „*price*“, „*FYI*“, „*trading*“, „*issues*“, „*credit*“, „*review*“, „*questions*“, „*contract*“	„*love*“, „*hotmail*“, „*night*“, „*weekend*“, „*hey*“, „*msn*“, „*man*“, „*mom*“, „*yahoo*“, „*fun*“, „*god*“, „*really*“, „*game*“, „*house*“	„*enron*“, „*ferc*“, „*hotmail*“, „*http*“, „*attached*“, „*dynegy*“, „*aol*“, „*fyi*“, „*carrfut*“, „*counterparty*“, „*ect*“, „*cpuc*“, „*com*“, „*org*“, „*trading*“, „*nymex*“, „*eol*“, „*thanks*“, „*www*“, „*enrononline*“, „*gas*“, „*tomorrow*“, „*calpine*“, „*pge*“

Za potrebe **PL** zadatka korišćen je javno dostupan korpus *Enron*, koji sadrži približno 600,000 poruka elektronske pošte na engleskom jeziku. *Enron* korpus sadrži poslovne i lične elektronske poruke koje su nastale od strane više od stotinu zaposlenih u *Enron* korporaciji u periodu od 3.5 godina (od 1998. do 2002. godine). Korpus je javno objavljen od američke nacionalne regulatorne komisije za energetiku (eng. *Federal Energy Regulatory Commission*, *FERC*) tokom pravnog istraživanja stečaja kompanije. Sa tačke naučnog istraživanja, korpus je najpre obrađen i objavljen na Univerzitetu *Carnegie Mellon* [102], a kasnije su pojedini delovi ovog korpusa korišćeni u istraživanjima. U dosadašnjim istraživanima posvećenih rešavanju **PL** zadatka, primenjivane su metode ručnog obeležavanja i evaluacije obeleženih podataka korišćenjem tradicionalnih ML metoda [89], poređenja različitih načina vektorizacije sadržaja poruka i tradicionalnih ML algoritama [8], kao i metode koje analiziraju

Tabela 7.3: *Oznake i njihova značenja u $Enron_C$ skupu podataka*

Oznaka	Značenje
Poslovni	Čisto poslovni sadržaj
Uglavnom poslovni	Poslovnim sadržaj koji sadrži i lične delove
Mešoviti	Kombinovani poslovni i lični sadržaji
Uglavnom lični	Ličnim sadržaj koji sadrži poslovne delove
Lični	Čisto lični sadržaj
Neodređen	Kategorija se ne može odrediti iz sadržaja

grafovske strukture nizova poruka [9] u cilju pronalaženja što uspešnijeg načina za klasifikaciju poruka. Na zadacima klasifikacije konverzacionih tekstova jedan od najvećih izazova predstavlja nepostojanje označenih podataka koji su vezani za predmet istraživanja. Iz razloga zaštite privatnosti, dodatan problem predstavlja i nedostupnost konverzacionih tekstova koji uključuju i privatne sadržaje. Na primer, korpus poruka elektronske pošte *Avocado*[33] je licenciran i zahteva izvesna materijalna ulaganja da bi mu se moglo pristupiti. Zahvaljujući podršci istraživača sa Univerziteta *Columbia*, koji su poruke iz korpusa *Enron* samostalno obeležavali i klasifikovali u pogledu privatnosti sadržaja [8], u ovom istraživanju je korišćena ova verzija obeleženih *Enron* poruka, pod nazivom *Enron Columbia* i dodeljenom oznakom $Enron_C$. U $Enron_C$ skupu podataka postoji šest različitih kategorija određenih prema stepenu poslovnog, odnosno ličnog sadržaja u poruci koje su korišćene prilikom označavanja (pogledati tabelu 7.3). U našim eksperimentima, ove kategorije su grupisane u dve osnovne kategorije **`Poslovna`** i **`Lična`** na sledeći način:

- **`Lična`**: Lične (eng. *Personal*), Uglavnom lične (eng. *Somehow personal*) i Mešovite (eng. *Mixed*);
- **`Poslovna`**: Poslovne (eng. *Business*) i Uglavnom poslovne (eng. *Somehow business*).

U svim eksperimentima izvršena je inicijalna normalizacija sadržaja poruka koja je uključila formatiranje teksta u mala slova, uklanjanje adresa elektronske pošte, internet adresa i višestrukih razmaka. Nakon toga iz skupa podataka su uklonjene prazne i ponovljene poruke, čime je nastao skup $Enron_{Cp}$. Broj poruka elektronske pošte u svakoj klasi u $Enron_C$ i $Enron_{Cp}$ skupovima podataka, pre i nakon inicijalne obrade, predstavljen je u tabeli 7.4.

Tabela 7.4: *Statistika $Enron_C$ skupa podataka pre i nakon obrade praznih i ponovljenih poruka elektronske pošte*

Korpus	**`Poslovna`**	**`Lična`**	Total
$Enron_C$	9,738 (86.5%)	1,523 (13.5%)	11,261
$Enron_{Cp}$	8,651 (86.6%)	1,340 (13.4%)	9,991

Na **`PL`** zadatku, internet domeni adresa primalaca mogu imati važnu ulogu u pronalaženju poruka lične prirode. Lična komunikacija često se odvija između osoba van organizacije, a internet domeni elektronske pošte mogu biti pokazatelj takvih aktivnosti. U okviru eksperimenata otkriveno je da su reči kao što su „*hotmail*“ i „*yahoo*“ učestalo pojavljuju u klasi **`Lična`** (pogledati tabelu 7.2). Iz tog razloga su internet domeni primalaca prepoznati kao značajna informacija koja je korišćena u okviru eksperimenata sa proširenim sadržajem (Ext). Dodatno, naslov poruke elektronske pošte, koji je karakterističan za ovu vrstu konverzacionih poruka, je u svim eksperimentima pridodat na sadržaj poruke pre vektorizacije, ali

[33] https://catalog.ldc.upenn.edu/LDC2015T03

je nezavisno posmatran kod izračunavanja dodatnih semantičkih atributa poruke. Koristeći predloženu metodologiju za klasifikaciju konverzacionih tekstova (pogledati odeljak 7), u okviru ovog zadatka izvršeni su sledeći eksperimenti u kojima se kao ulaz u algoritam koristi:

E1 Pojedinačna poruka (Msg) – osnova;

E2 Pojedinačna poruka sa internet domenima adresa primalaca iz potpisa poruke (Msg-Ext);

E3 Pojedinačna poruka sa odgovorima iz iste konverzacione grane (Brch);

E4 Pojedinačna poruka sa odgovorima iz iste konverzacione grane i internet domenima adresa primalaca iz potpisa poruke (Brch-Ext);

E5 Cela konverzaciona grana sa pratećim informacijama (Brch*) – osnova.

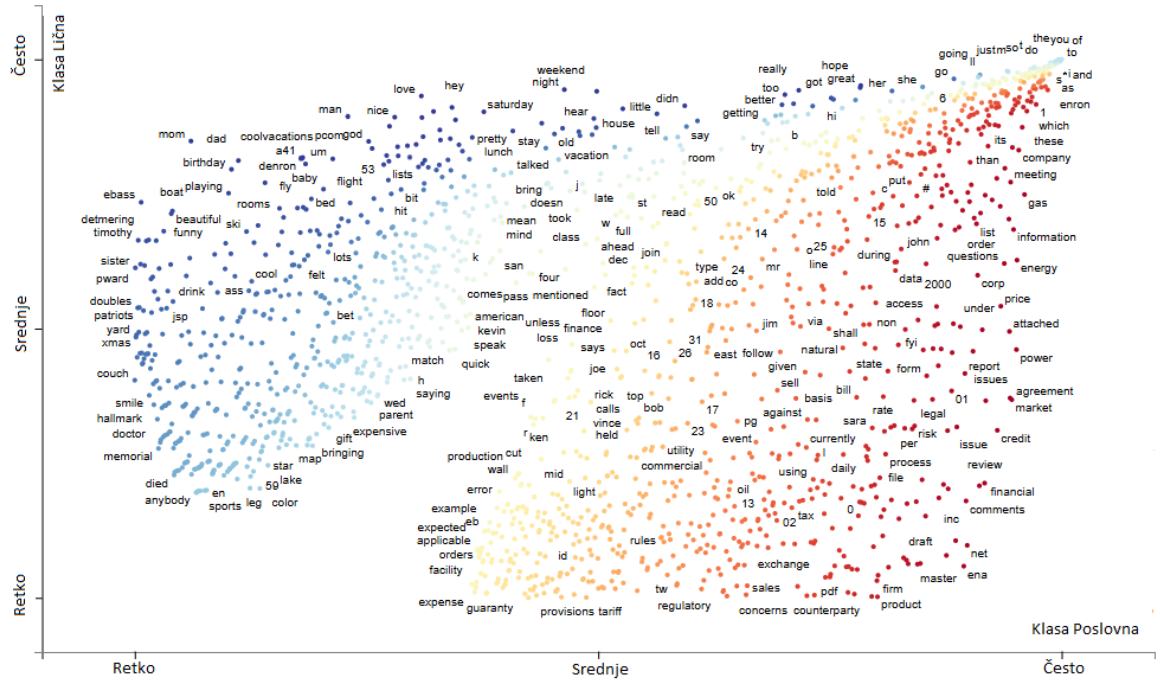
Specifično, u skupu označenih podataka elektronske pošte $Enron_C$ dostupan je ceo konverzacioni niz sa pratećim informacijama iz zaglavlja svih poruka, koji će biti iskorišćen kao osnova za poređenje u okviru dodatnog **E5** eksperimenta (Brch*). Nesegmentiran tekstualni sadržaj može poslužiti za analizu uticaja procesa pravilnog razdvajanja pojedinih segmenata konverzacione poruke i njihovog uticaja na tačnost klasifikacije. Ovaj uticaj se prvenstveno odnosi na karakteristične tokene koji se mogu pojaviti u zaglavljima poruka, koji mogu uticati na preprilagođavanje ML modela podacima za obučvanje ukoliko se podaci koriste na takav način.

Sadržaji poruka u skupu podataka $Enron_C$ mogu se predstaviti pomoću alata *ScatterText*[34] i ugrađenog dijagrama rasejanja učestalosti pojavljivanja reči u klasama. **Poslovna** i **Lična**. Učestalost pojavljivanja reči je mera koju alat *ScatterText* koristi kao koordinate za svaku tačku na dijagramu (slika 7.5). X-osa na dijagramu označava učestalost pojavljivanja u klasi **Poslovna**: ukoliko se reč učestalo pojavljuje u poslovnim porukama elektronske pošte, ona je pozicionirana udesno. Slično, Y-osa na dijagramu označava učestalost pojavljivanja u klasi **Lična**: reč koja se učestalo pojavljuje u ličnim porukama elektronske pošte je pozicionirana ka vrhu. Iz ovih razloga, najznačajnije oblasti na dijagramu koje daju dobar pregled o raspodeli reči po klasama su:

- Gore-levo: reči sa učestalim pojavljivanjem u klasi **Lična**;
- Dole-desno: reči sa učestalim pojavljivanjem u klasi **Poslovna**;
- Gore-desno: reči sa učestalim pojavljivanjem u obe klase.

Sa dijagrama prikazanog na slici 7.5 možemo uočiti da uobičajene poslovne reči kao što su „*sporazum*“ (eng. „*agreement*“), „*energija*“ (eng. „*energy*“) i „*prilog*“ (eng. „*attachment*“) naglašavaju zvanični ton koji se može naći u sadržajima poslovnih poruka elektronske pošte. Nasuprot tome, poruke koje se nalaze u kategoriji **Lična** imaju opuštajući ton, često koriste reči kao što su „*ljubav*“ (eng. „*love*“), „*vikend*“ (eng. „*weekend*“) i „*zabava*“ (eng. „*fun*“). Boje na dijagramu prikazuju vrednost skora koji se naziva skalirana F_1 mera, koja je uvedena od strane autora ove vizuelizacije u [97]. Reči sa skaliranom F_1 merom blizu nule, prikazane žutom i narandžastom bojom na dijagramu, imaju slične frekvencije za obe klase i nisu od velike važnosti. Kada je učestalost reči izraženija u jednoj klasi, vrednosti se pomeraju ka -1 (**Poslovna**) ili 1 (**Lična**), koje su na dijagramu obeležene crvenom ili plavom bojom, u tom redosledu. Tamnija nijansa crvene ili plave ukazuje na jaču pripadnost reči odgovarajućoj klasi.

[34] https://github.com/JasonKessler/scattertext



Slika 7.5: *Vizuelni prikaz karakterističnih reči za kategorije* **Poslovna** *i* **Lična** *napravljenog pomoću ScatterText alata*

Eksperimenti normalizacije teksta nad porukama elektronske pošte

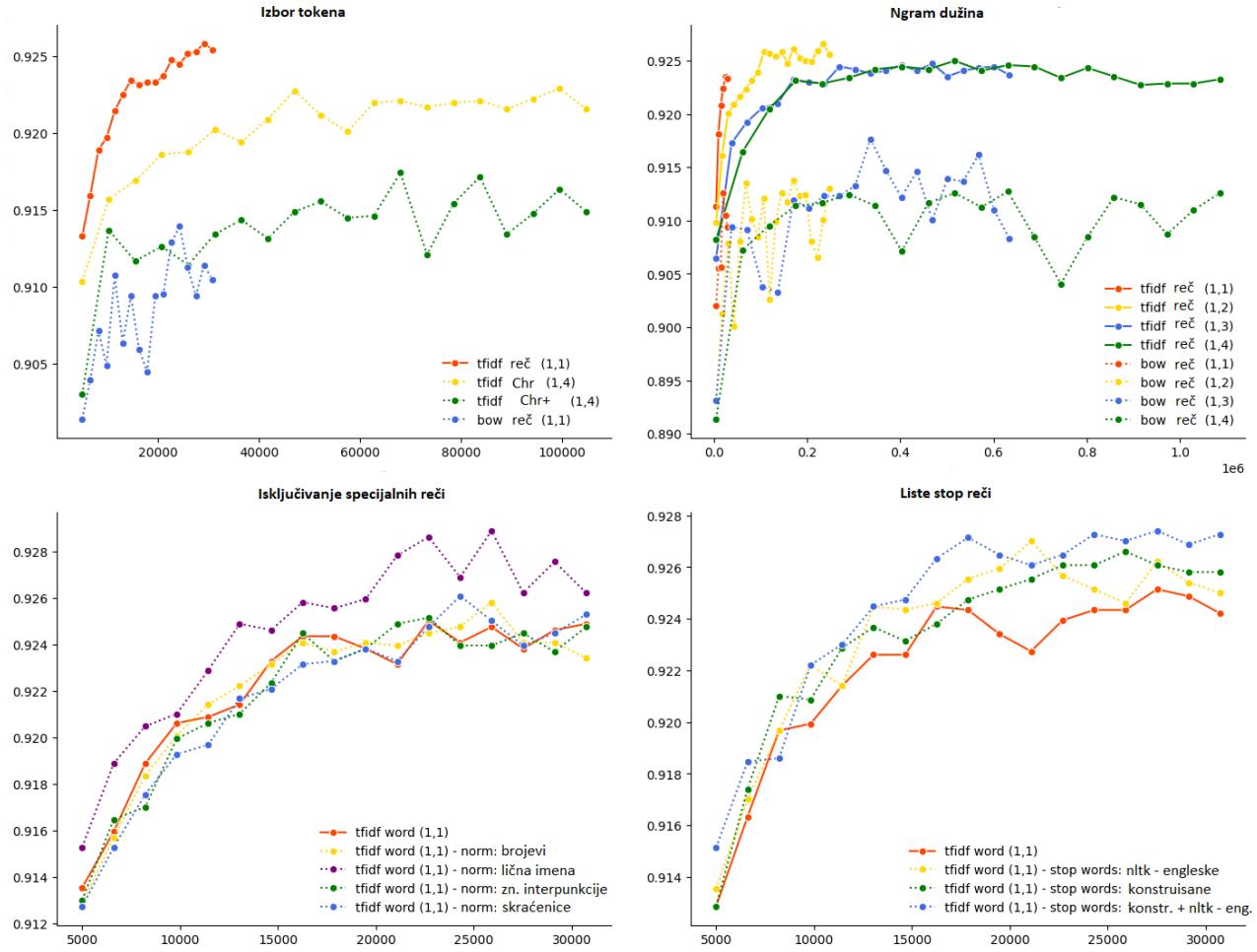
Da bi se utvrdile najbolje tehnike obrade i predstavljanja teksta, poređene su različite tehnike normalizacije teksta sprovođenjem velikog broja eksperimenata koristeći SGD-SVM klasifikacioni algoritam. U okviru ovog eksperimenta, lematizacija je primenjena na vrste reči kao što su glagoli, imenice, pridevi i prilozi. Za lematizaciju je korišćen *WordNetLemmatizer* iz *nltk*[35] Pajton biblioteke. Specijalne reči kao što su brojevi, lična imena, interpunkcijski znakovi i skraćenice su u okviru tekstualnih sadržaja pronalažene korišćenjem regularnih izraza ili leksikona ličnih imena. Takođe, konstruisana je lista stop reči specifičnih za korpus korišćenjem osobina alata *ScatterText*[36] i pratećih matematičkih formulacija. Autor alata u svom radu [97] uvodi mere preciznosti i odziva reči u korpusu i upućuje na njihov recipročan odnos. U vizuelnoj interpretaciji, reči sa visokim vrednostima odziva teže ka gornjem desnom uglu grafikona, dok su reči sa visokom preciznošću pozicionirane bliže osama (pogledati sliku 7.5). Činjenica da su reči sa izuzetno visokim stepenom odziva obično stop reči je korišćena za konstruisanje liste stop reči specifičnih za dati korpus. Kao što je prikazano na slici 7.6, rezultati eksperimenata su pokazali da:

- Ngram za $n = 2$ pokazuje bolje performanse od Ngram drugih dužina ($n \neq 2$) nad rečima kao tokenima;
- Tf-Idf težine nadmašuju BoW težine koje su bile najzastupljeni prisup za vektorizaciju teksta u prethodnim radovima na istom zadatku;
- Ngram nad rečima pokazuju bolje performanse od Ngram nad karakterima;
- Uključivanje lematizacije i isključivanje specijalno konstruisane liste stop reči igra važnu ulogu u poboljšanju performansi modela, zajedno sa uklanjanjem ličnih imena i znakova interpunkcije;

[35] https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html
[36] https://github.com/JasonKessler/scattertext

- Ograničavanje minimalnog ili maksimalnog broja (ili procenta) dozvoljenih tokena (reči ili karaktera) u vokabularu ne poboljšava performanse modela;
- Ograničavanje veličine rečnika u BoW i Tf-Idf vektorskim reprezentacijama značajno smanjuje performanse modela.

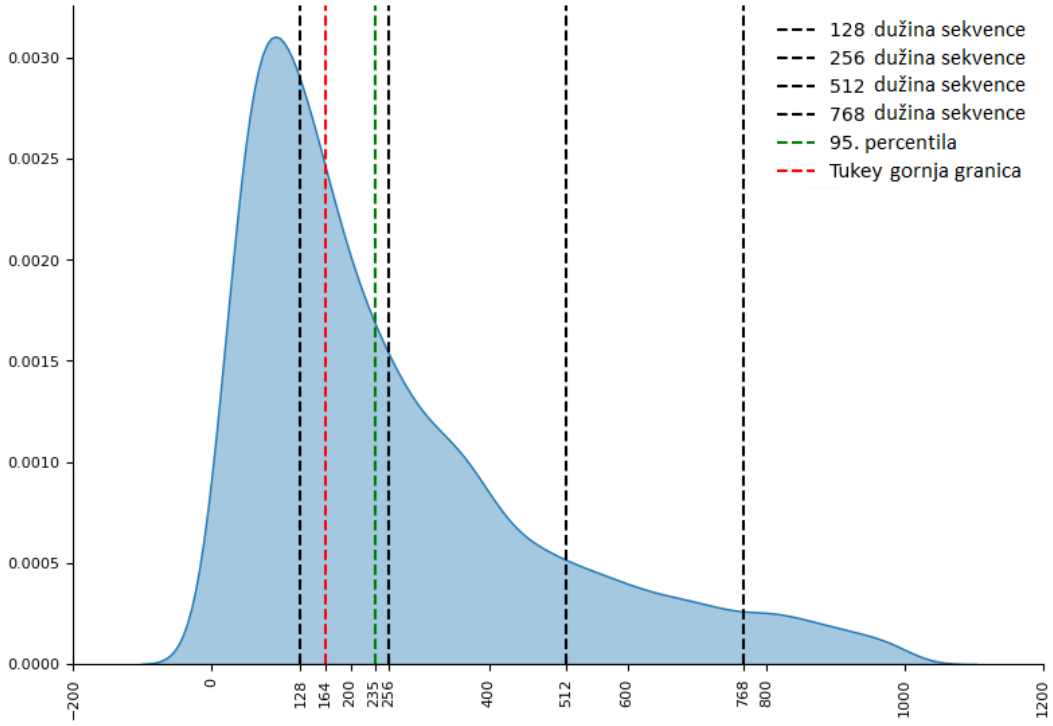
Na osnovu ovih zaključaka, za obradu ulaznih tekstualnih sadržaja su izabrani koraci normalizacije koji su pokazali najveći uticaj na poboljšanje performansi modela klasifikacije.



Slika 7.6: *Poređenje tehnika za obradu ulaznih tekstualnih sadržaja za različite težine i vrste tokena, dužine Ngram, stop i specijalnih reči. Za eksperimente je korišćen SGD-SVM algoritam za klasifikaciju sa podrazumevanim skupom parametara nad Msg-Ext sadržajem*

Još jedna tehnika za predstavljanje teksta koja je u eksperimentima korišćena jesu BERT-Embd. Tehnika Embd ima za cilj korišćenje vektorskih reprezentacija tekstualne sekvence koje čuvaju karakteristike reči preuzetih iz konteksta [130]. Kompaktni unapred obučeni BERT model, prethodno obučen na Vikipediji i korpusima knjiga, izabran je iz TensorFlow skladišta modela[37] (eng. *Google TensorFlow Hub*). Izabrani kompaktni model, sa L=2 skrivena sloja, dimenzije od H=256 i A=4 višestrukih ponavljanja mehanizma pažnje, korišćen je za inicijalizaciju ulaza svih modela dubokog učenja. Arhitektura standardnog BERT modela zadržana je i u manjim modelima, bez obzira na broj parametara, što ove BERT modele pogodnim za okruženja sa ograničenim računarskim resursima. Ovi modeli se mogu doobučavati na isti način kao originalni BERT modeli [204]. Deskriptivna statistika dužine sadržaja poruka elektronske pošte i *Tukey* statistički test značajnosti za

[37] https://tfhub.dev/s?module-tipe=text-embedding



Slika 7.7: *Analiza dužine ulazne tekstualne sekvence u Msg-Ext eksperimentu*

izračunavanje granica ekstremnih vrednosti su pomogli u izboru odgovarajuće granice za maksimalnu dužinu sekvence (eng. *Maximum Sequence Length, MSL*) u modelima (pogledati sliku 7.7). Vrednost MSL na **PL** zadatku postavljena je na 512 tokena kako bi se iskoristio što veći broj dostupnih informacija iz ulaznih podataka (broj tokena), a u skladu sa prethodnom analizom, dostupnim distribucijama BERT modela i računarskih resursa.

Konstrukcija pridruženih atributa

Za zadatak klasifikacije poruka elektronske pošte u klase **Poslovna** i **Lična** (**PL**) su kreirani Meta atributi, koji odgovaraju strukturi ove vrste konverzacionih poruka i jeziku na kojem su napisane (pogledati odeljak 7.2). Pored standardne liste atributa koja je preporučena i primenljiva u većini zadataka, za **PL** zadatak kreirani su dodatni atributi koji su specifični za **PL** zadatak (pogledati prilog A, tabelu A.1), kao što su:

- *acronyms_indicator* – indikator prisustva akronima, što može ukazivati na formalnost ili tehnički karakter poruke,
- *business_indicator* – indikator koji prepoznaje poslovne termine kako bi se razlikovale poslovne od ličnih poruka,
- *subject_lex_count* i *subject_lex_length* – broj reči i karaktera u naslovu poruke, što može dati naznake o važnosti ili formalnosti poruke, i koji je, dodatno, dostupan samo u određenim vrstama konverzacionih poruka,
- *free_domains_ratio* – procenat adresa elektronske pošte sa besplatnih domena (na primer, *yahoo.com* ili *hotmail.com*), što može biti relevantno za identifikaciju lične ili poslovne komunikacije,
- *recipients_domains_coherency* – koherentnost domena primalaca, koji ukazuje na ciljanu grupu primalaca (na primer, svi primaoci pripadaju istoj organizaciji).

Sve grupe konstruisanih atributa su objedinjene u jednu listu, označenu sa Meta, koja sadrži ukupno 57 dodatnih atributa za **PL** zadatak (pogledati prilog A, tabelu A.1). Na sve konstruisane atribute primenjena je L2 normalizacija, čime je vektorska reprezentacija atributa svake poruke prilagođena tako da njena euklidska norma (eng. *Euclidean norm, L2*) bude jednaka 1.

Izabrani algoritmi mašinskog učenja

U toku eksperimenata, izvršeno je poređenje performansi dva tradicionalna (SGD i ERT) i dva DL algoritma (BiLSTM, BiLSTM+Att) nad različitim vektorskim predstavljanjima sadržaja konverzacione poruke. Najbolji parametri za oba klasična algoritma odabrani su pomoću algoritama za opsežno pretraživanje[38] (eng. *grid search*) parametara. U tabeli 7.5 predstavljeno je poređenje performansi tradicionalnih SGD i ERT algoritama klasifikacije za različite vrednosti parametara ovih algoritama. Balansirana vrednost za težine klasa u parametrima za pretraživanje predstavlja težine klasa koje su inverzno proporcionalne njihovim frekvencijama u skupu za obučavanje. Za pokretanje eksperimenata su korišćene implementacije ovih algoritama u *scikit-learn*[39] Pajton biblioteci [149].

Tabela 7.5: *Izbor optimalnih parametara za klasifikaciju poruka elektronske pošte u klase* **Poslovna** *i* **Lična** *korišćenjem algoritma za opsežno pretraživanje*

Algoritam	Parametar	Prostor vrednosti	Izabrana Vrednost
SGD	Funkcija greške	granična (hinge), logaritamska, modifikovani huber, kvadratna granična, perceptron	modifikovani huber
	Metoda normalizacije	L1, L2, ElasticNet	L2
	Parametar regularizacije	0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000	0.0001
	Stopa učenja	konstantna, optimalna, inverzna skalirajuća, adaptivna	optimalna
	Težine klasa	{L: 0.5, P: 0.5}, {L: 0.6, P: 0.4}, {L: 0.7, P: 0.3}, balansirana	{L: 0.7, P: 0.3}
	Brzina učenja	1, 10, 100	10
ERT	Broj stabala	10, 20, 30, 50, 100, 200	10
	Kriterijum podele	gini koeficijent, entropija	entropija
	Minimalni broj primera po listu	1, 3, 10	3
	Težine klasa	{L:0.5, P:0.5}, {L:0.6, P:0.4}, {L:0.7, P:0.3}, balansirana	{L:0.7, P:0.3}

Za SVM algoritam, optimalna funkcija greške modifikovani huber zapravo predstavlja kvadratnu SVM grešku za parametar *gamma*=2 [223]. U daljem delu teksta, SGD-SVM će označavati SVM algoritam sa funkcijom greške modifikovani huber. Za DL modele, parametri su odabrani ručno korišćenjem opsežnog skupa eksperimenata. Obučavanje je optimizovano korišćenjem metode za stohastičku optimizaciju *Adam* [100]. Parametar za brzinu učenja je postavljen na 3×10^{-5}, u skladu sa preporukama za početne vrednosti ovog parametra za doobučavanje BERT modela [94]. Strategija ranog zaustavljanja je korišćena da bi se sprečilo preterano prilagođavanje modela [159]. Pored toga, sve DL arhitekture koriste:

- Gradijentno spuštanje sa 64 instanci u svakom skupu za proveru;
- Dimenziju ugnježdenih vektora jednaku 256 (zavisnu od modela);
- MSL postavljenoj na 512 tokena;

[38] https://scikit-learn.org/0.15/modules/grid_search.html
[39] https://scikit-learn.org

- Dva BiLSTM sloja sa brojem jedinica jednakom MSL;
- Odnos isključivanja neurona od 0.1 za BiLSTM i 0.4 za FFN slojeve;
- Aktivacionu funkciju *ReLU* u FFN slojevima i σ u završnom sloju.

Izabrana funkcija greške je BCE, koja se koristi kao standardna funkcija greške za binarnu klasifikaciju (pogledati jednačinu 7.7). Podaci su podeljeni na skupove za obuku, proveru i testiranje u odnosu 50:25:25. Modeli su obučavani na skupu za obuku, izbor hiperparametara je izvršen korišćenjem skupa za proveru (u toku obučavanja), dok je konačna prediktivna tačnost modela proverena na nezavisnom test skupu (nakon obučavanja). Za obučavanje tradicionalnih modela (SGD-SVM, ERT) korišćena je CV sa pet ponavljanja. Da bi se ilustrovale performanse predloženih pristupa, izvršeno je poređenje dobijenih rezultata sa osnovnim modelima izgrađenim na pojedinačnoj poruci (Msg) i celoj grani poruka sa pridruženim informacijama iz zaglavlja poruka (Brch*).

7.5. Klasifikacija istinitosti glasine i tipa delovanja na objavljenu glasinu

Predložena metoda za klasifikaciju konverzacionih tekstova biće proverena nad konverzacionim podacima iz drugih domena i struktura konverzacije. U tom cilju biće iskorišćen skup podataka objavljen na *SemVal2019*[40] takmičenju na zadatku za utvrđivanje istinitosti glasina i tipu delovanja na ovakve vrste tekstualnih objava[41]. Ovaj zadatak je unapređenje zadatka sa istog takmičenja iz 2017. godine [50], jer je sa porastom širenja dezinformacija i lažnih vesti na društvenim mrežama postalo neophodno razviti efikasne i napredne metode za automatsku proveru verodostojnosti objavljenih informacija. Skup podataka sadrži konverzacione nizove, njihove delove, objave i komentare na objave sa Tviter i Redit društvenih mreža napisanih na engleskom jeziku. Objave se odnose na glasine poznate u svetskoj javnosti za koje u trenutku objavljivanja nije bila poznata njihova tačnost [65]. Tipovi delovanja u komentarima na objave o glasinama su klasifikovani na sledeći način (tabela 7.6):

- podrška (eng. *support*, *S*) – podržava istinitost objave,
- opovrgavanje (eng. *deny*, *D*) – odbija da poveruje u istinitost objave,
- upit (eng. *query*, *Q*) – traži dodatne potvrde za istinitost objave,
- komentar (eng. *comment*, *C*) – iznosi svoje mišljenje bez jasnih naznaka o stavu prema istinitosti iznete objave.

Početne objave su kategorizovane kao istinite (eng. *True*, *T*), neistinite (eng. *False*, *F*) ili neverifikovane (eng. *Unverified*, *UVF*), koje su označene od strane anotatora naknadnom proverom njihove tačnosti. Raspodela poruka prema kategorijama tipova delovanja i istinitosti glasina je prikazana u tabeli 7.6. Raspodela poruka prema tipu delovanja je nebalansirana, odnosno primetan je značajno veći broj komentara u odnosu na upite, poruke podrške ili opovrgavanja. Zadaci postavljeni za rešavanje su:

A. Klasifikacija tipa delovanja u komentaru na objavljenu glasinu (**TD**);

B. Klasifikacija istinitosti glasine (**IG**).

[40] https://alt.qcri.org/semeval2019
[41] https://competitions.codalab.org/competitions/19938

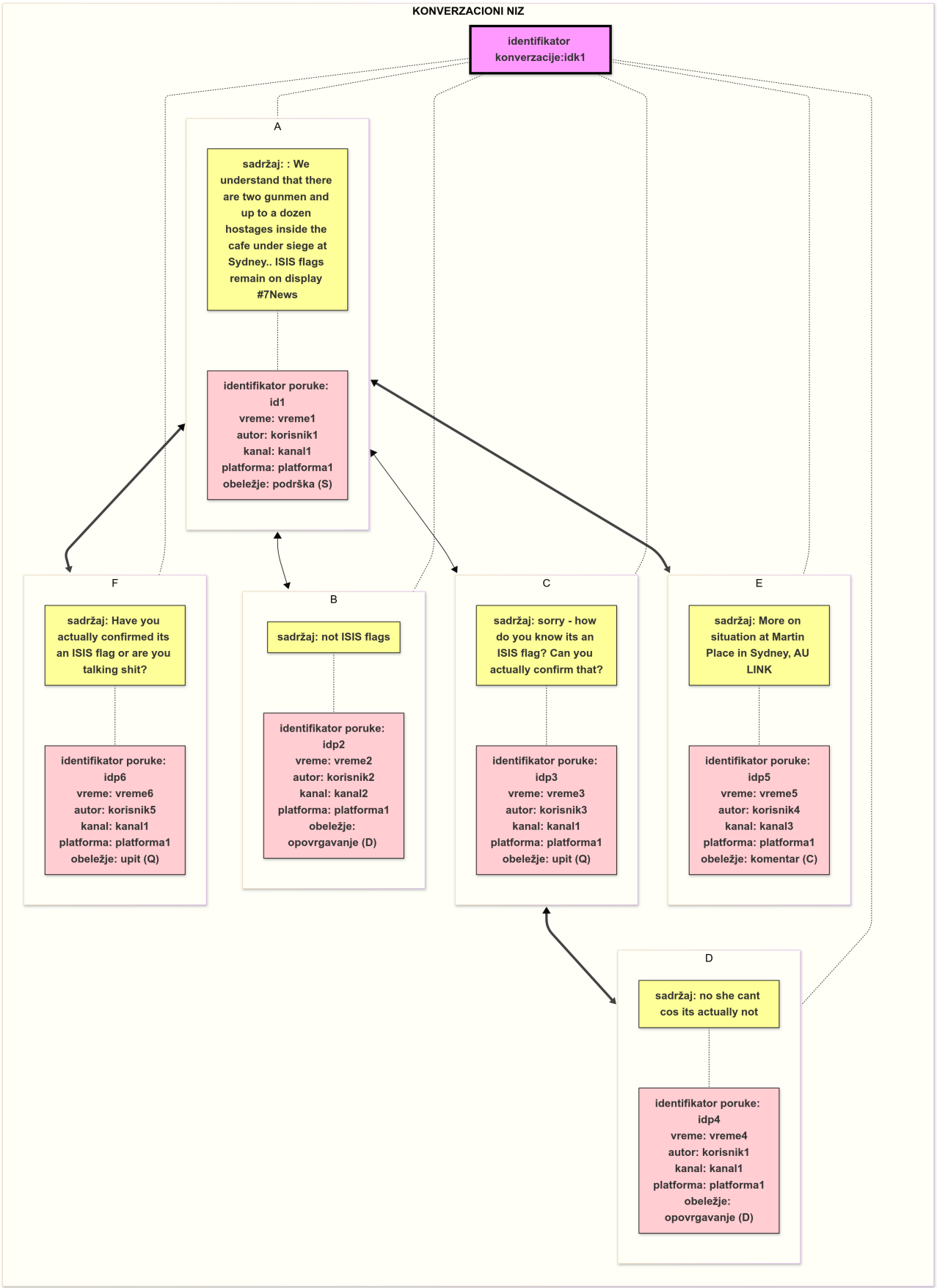
Tabela 7.6: *Raspodela poruka u skupovima za obuku i testiranje po kategorijama delovanja i istinitosti glasina [65]*

		S	D	Q	C	Total	T	F	UVF	Total
Train	Tviter	1004	415	464	3685	5568	145	74	106	325
	Redit	23	45	51	1015	1134	9	24	7	40
	Total	1027	460	515	4700	6702	154	98	113	365
Test	Tviter	141	92	62	771	1066	22	30	4	56
	Redit	16	54	31	705	806	9	10	6	25
	Total	157	146	93	1476	1872	31	40	10	81
Total		1184	606	608	6176	8574	185	138	123	446

Autori takmičenja su učesnicima ponudili nekoliko implementacija koje su poslužile kao osnovna rešenja za poređenje performansi različitih pristupa. Za zadatak klasifikacije tipa delovanja, učesnicima je bio dostupan DL model BrchLSTM, koji je predstavljao pobedničko rešenje na istom takmičenju iz 2017. godine [104]. Ovaj model koristi strukturu konverzacije tako što je deli na Brch. Radi se o neuronskoj mreži sa LSTM slojevima za obradu sekvenci poruka, u kojoj se u svakom koraku predviđa oznaka tipa delovanja. Za klasifikaciju istinitosti glasine (zadatak **B**.), ponuđena su dva osnovna pristupa. Prvi je proširenje BrchLSTM modela [105], koji koristi iste atribute kao i za klasifikaciju stava, ali daje jedan izlaz po Brch, dok se konačno predviđanje za konverzaciju odlučuje većinskim glasanjem nad izlazima iz grana te konverzacije. Drugi model je NileTMRG [53], linearni SVM sa BoW reprezentacijom poruke, kombinovan sa atributima sadržaja kao što su prisustvo URL adresa, heš oznaka i proporcija podržavajućih, opovrgavajućih i upitnih tvitova u Brch.

Struktura skupa podataka, koja obuhvata konverzacione nizove i pojedinačne poruke sa društvenih mreža Tviter i Redit, je pogodna za primenu metode za klasifikaciju konverzacionih tekstova predstavljene u ovom radu, kao što je prikazano na slici 7.8. Na ovaj način bi se proverila uspešnost predložeme metode nad konverzacionim podacima sa društvenih mreža poređenjem sa rezultatima ostvarenim u okviru takmičenja. Dodatno, ispitala bi se uspešnost tehnika koje su u ovom istraživanju predložene, a koja prijavljena rešenja nisu pokrila, kao što su korišćenje Att mehanizma, kao i značaj emocionalnih i moralnih atributa za rešavanje ovih zadataka. Na zadatku lažnih vesti, odnosno istinitosti glasina, istraživački radovi su se bavili pitanjem na koji način emocije utiču na prepoznavanje dezinformacija, i ukazali na uspešnost ovog prepoznavanja kada se kombinuju s drugim leksičkim atributima. Jedan od pristupa uporedo analizira emocije u tekstualnom sadržaju vesti (na primer, strah ili bes) i emocije u komentarima korisnika (na primer, iznenađenje ili radost), i na taj način otkrivaju da nesklad između ove dve vrste emocija (dualne emocije) koje mogu biti indikator lažnih vesti [224]. Neke druge metode kombinuju emocionalne oznake s psiholingvističkim obeležjima iz leksikona LIWC (na primer, analitičko mišljenje ili emocionalni ton) kako bi detektovali glasine, čime je povećana tačnost modela [76]. Najnoviji pristupi koriste multimodalnu fuziju (tekst + slike/video) i upozoravaju na zanemarivanje važnih kulturoloških razlika u izražavanju emocija, promenu emocionalnih izražavanja tokom vremena (na primer, eskalacija besa u diskusijama), i fokusiranje na pojedinačne platforme umesto na univerzalne modele. Složenost postavljenog zadatka zahteva integraciju DL metoda sa interpretabilnim metodama, analizu u realnom vremenu, te etičko osmišljavanje sistema kako bi se izbegla potencijalna pristrasnost u podacima [121].

Prema predloženim DL arhitekturama u ovom istraživanju, sa i bez uključivanja Att mehanizma i Meta atributa, eksperimenti koji su korišćeni za rešavanje zadatka **A**. su:



Slika 7.8: *Primer konverzacionog niza Tviter poruka iz skupa označenih podataka na istinitost glasina i tipa delovanja na objavljenu glasinu. Preuređena slika na osnovu primera iz rada [65]*

E1 Pojedinačna poruka (Msg) – osnova;

E3 Konverzaciona grana poruka (Brch).

dok su za zadatak **B.** korišćeni:

E3 Konverzaciona grana poruka (Brch) – osnova;

E4 Konverzaciona grana poruka sa tipom delovanja pridruženim svakoj poruci (Brch-Ext).

Inicijalna semantička analiza sadržaja poruka je pokazala da se mogu uočiti određene pravilnosti u rečima koje se učestalo pojavljuju u pojedinačnim kategorijama. U sadržaju poruka je uočena značajna zastupljenost imenovanih entiteta koji su povezani sa temama za koje su analizirane glasine, koje mogu imati značaja na tačnost klasifikacije. Iz sadržaja poruka, nakon uklanjanja stop reči i imenovanih entiteta, izdvojene su karakteristične reči za svaku klasu u zadacima klasifikacije tipa delovanja na glasinu (`TD`) i istinitosti glasina (`IG`). Najzastupljenije reči povezane sa svakom kategorijom sugerišu obrasce i teme koje su obično povezane sa tipom delovanja ili istinitosti glasine. Na zadatku klasifikacije tipa delovanja (pogledati tabelu 7.7), identifikovane su katekteristične reči za svaku klasu kojima su pridružene emocionalne i moralne kategorije korišćenjem *NRC.EmoLex* i *eMFD* leksikona, u tom redosledu:

- Klasa `S`: *„vesti“* (eng. *„news“* → [*anticipation*, *trust*] [*authority*, *care*]), *„izveštaj“* (eng. *„report“* → [*trust*] [*authority*]), *„napad“* (eng. *„attack“* → [*anger*, *fear*] [*loyalty*, *authority*]) i *„policija“* (eng. *„police“* → [*trust*] [*authority*]), ukazuju na izjave koje pružaju činjenične ili podržavajuće informacije, često vezane za izveštavanje ili potvrđivanje.
- Klasa `D`: *„osumnjičeni“* (eng. *„suspect“* → [*fear*] [*fairness*, *authority*]), *„pogrešno“* (eng. *„false“* → [*disgust*] [*fairness*]), *„laž“* (eng. *„lie“* → [*disgust*, *anger*] [*fairness*]) i *„falsifikat“* (eng. *„fake“* → [*disgust*] [*fairness*]), koje impliciraju skepticizam ili osporavanje.
- Klasa `Q`: *„potvrditi“* (eng. *„confirm“* → [*trust*] [*authority*]), *„značenje“* (eng. *„mean“* → [*neutral*] [*non-moral*]), *„reći“* (eng. *„tell“* → [*trust*] [*fairness*]), i *„video“* (eng. *„video“* → [*surprise*] [*non-moral*]), koje odražavaju jezik upita ili traženja pojašnjenja.
- Klasa `C`: *„pogledati“* (eng. *„look“* → [*neutral*] [*care*]), *„godina“* (eng. *„year“* → [*neutral*] [*non-moral*]), *„potreba“* (eng. *„need“* → [*neutral*] [*care*]), i *„misliti“* (eng. *„think“* → [*neutral*] [*fairness*]), koje ukazuju na posmatranja ili neodlučne komentare.

Tabela 7.7: *Karakteristične reči za kategorije* `S`, `C`, `D` *i* `Q` *na zadatku klasifikacije tipa delovanja na glasinu*

`S`	`C`	`D`	`Q`
„least“, *„video“*, *„people“*, *„crash“*, *„supermarket“*, *„siege“*, *„news“*, *„shot“*, *„soldier“*, *„hold“*, *„amp“*, *„report“*, *„attack“*, *„parliament“*, *„dead“*, *„french“*, *„suspect“*, *„gunman“*, *„say“*, *„rt“*, *„breaking“*, *„kill“*, *„cafe“*, *„shoot“*, *„police“*, *„hostage“*	*„call“*, *„come“*, *„much“*, *„look“*, *„kill“*, *„still“*, *„year“*, *„way“*, *„amp“*, *„thing“*, *„police“*, *„need“*, *„right“*, *„well“*, *„want“*, *„also“*, *„use“*, *„even“*, *„take“*, *„see“*, *„time“*, *„good“*, *„gt“*, *„know“*, *„think“*, *„go“*, *„make“*, *„say“*, *„get“*, *„people“*	*„suspect“*, *„year“*, *„number“*, *„photo“*, *„cop“*, *„kill“*, *„happen“*, *„police“*, *„still“*, *„evidence“*, *„need“*, *„make“*, *„right“*, *„believe“*, *„true“*, *„stop“*, *„go“*, *„get“*, *„people“*, *„fact“*, *„report“*, *„not“*, *„know“*, *„lie“*, *„news“*, *„flag“*, *„say“*, *„fake“*	*„kill“*, *„mean“*, *„need“*, *„give“*, *„confirm“*, *„video“*, *„tell“*, *„know“*, *„did“*, *„hostage“*, *„right“*, *„news“*, *„look“*, *„show“*, *„vote“*, *„please“*, *„people“*, *„make“*, *„still“*, *„many“*, *„report“*, *„debunk“*, *„police“*, *„source“*, *„get“*, *„say“*, *„true“*

U klasifikaciji istinitosti glasine (pogledati tabelu 7.8), identifikovane su katekteristične reči za svaku klasu kojima su pridružene emocionalne i moralne kategorije korišćenjem *NRC.EmoLex* i *eMFD* leksikona, u tom redosledu:

- Klasa **T**: „*ispravno*“ (eng. „*true*“ → [*trust*][*fairness*]), „*potvrditi*“ (eng. „*confirm*“ → [*trust*][*authority*]), „*vojnik*“ (eng. „*soldier*“ → [*fear*][*authority*]), i „*policija*“ (eng. „*police*“ → [*fear*][*authority*]), ukazuje na sadržaj koji se smatra tačnim ili verifikovanim.
- Klasa **F**: „*nesreća*“ (eng. „*crash*“ → [*fear*, *sadness*][*care*]), „*pokazati*“ (eng. „*show*“ → [*anticipation*, *trust*][*loyalty*]), „*verovati*“ (eng. „*believe*“ → [*trust*][*loyalty*]), i „*demantovati*“ (eng. „*debunk*“ → [*anger*, *anticipation*][*fairness*]), što sugeriše potencijalne dezinformacije ili netačne tvrdnje.
- Klasa **UVF**: „*radnja*“ (eng. „*act*“ → [*neutral*]), „*nenaoružan*“ (eng. „*unarmed*“ → [*fear*][*care*]), „*prosek*“ (eng. „*average*“ → [*neutral*][*fairness*]), i „*starost*“ (eng. „*age*“ → [*sadness*][*care*]), što može ukazivati na neproverene ili nesigurne informacije koje nisu jasno tačne ili netačne.

Tabela 7.8: *Karakteristične reči za kategorije* **T**, **F**, **UVF** *na zadatku klasifikacije istinitorsti glasine*

T	**F**	**UVF**
„*attack*“, „*flag*“, „*war*“, „*memorial*“, „*true*“, „*place*“, „*situation*“, „*say*“, „*hill*“, „*live*“, „*suspect*“, „*kill*“, „*shot*“, „*breaking*“, „*parliament*“, „*confirm*“, „*gunman*“, „*shoot*“, „*soldier*“, „*people*“, „*dead*“, „*cafe*“, „*police*“, „*hostage*“	„*crash*“, „*attack*“, „*war*“, „*state*“, „*breaking*“, „*tonight*“, „*show*“, „*secret*“, „*day*“, „*believe*“, „*hold*“, „*hostage*“, „*call*“, „*emergency*“, „*back*“, „*muslim*“, „*police*“, „*debunk*“, „*get*“, „*report*“, „*say*“, „*prince*“	„*store*“, „*still*“, „*unarmed*“, „*make*“, „*give*“, „*get*“, „*use*“, „*cop*“, „*look*“, „*turnout*“, „*tell*“, „*kill*“, „*report*“, „*medium*“, „*brown*“, „*hostage*“, „*age*“, „*show*“, „*suspect*“, „*release*“, „*shoot*“, „*robbery*“, „*vote*“, „*say*“

Na osnovu karakterističnih reči unutar svake klase, moguće je izvršiti preliminarnu analizu njima pridruženih emocionalnih i moralnih kategorija. Klasa koja izražava opovrgavanje (**D**) karakterišu reči sa pridruženim emocionalnim afektom [*anger*, *disgust*, *fear*] i moralnim vrednostima [*fairness*], koje odgovaraju neslaganju ili percepciji manipulacije informacijama. S druge strane, klasa upita (**Q**) najčešće se povezuju sa emocijama [*surpsise*, *trust*], koje odražavju potrebu za dodatnim informacijama ili nesigurnost u pogledu istinitosti. Klasa za iskazivanje podrške (**S**) dominantno pripadaju kategorijama [*trust*] i [*authority*], dok komentari (**C**) često sadrže neutralne kategorije [*neutral*], [*non-moral*], što ukazuje na lične refleksije bez jasne potvrde ili osporavanja informacija. Sa druge strane, u klasi potvrđenih glasina (**T**) dominantne su kategorije [*trust*, *authority*], dok su neistinite glasine (**F**) povezane sa raznovrsnim skupom emocionalnih kategorija i najdominantnijom moralnom kategorijom [*loyalty*]. Ovi obrasci ukazuju na potencijalne mogućnosti za unapređenje algoritama klasifikacije na zadacima **TD** i **IG** pomoću semantičkih analiza sadržaja poruka, koje uključuju emocionalne i moralne atribute teksta.

Konstrukcija pridruženih atributa

Slično kao u zadatku **PL** (pogledati odeljak 7.4), za **TD** i **IG** zadatke su, pored standardne liste Meta atributa, kreirani atributi koji su specifični za konverzacione poruke sa društvenih mreža, kao što su:

- *is_retweet* – pokazuje da li je poruka podeljena, što može ukazivati na popularnost ili podršku određenoj izjavi,
- *is_mention* – označava prisustvo spominjanja drugih korisnika, što može biti indikator interakcije ili ciljane komunikacije,

- *is_plain* – identifikuje poruke bez specijalnih formata ili oznaka, što ukazuje na originalnosti i jednostavnost sadržaja.

Sve grupe konstruisanih atributa su objedinjene u jednu listu, označenu sa Meta, koja sadrži ukupno 55 dodatnih atributa za **TD** i **IG** zadatke (pogledati prilog A, tabelu A.1). Na sve konstruisane atribute primenjena je tehnika L2 normalizacije.

Izabrani algoritmi mašinskog učenja

Treniranje modela uključuje nekoliko ključnih parametara za optimizaciju i poboljšanje efikasnosti modela, kao što su broj koraka po epohi, ukupan broj koraka treniranja, broj prolazaka kroz skup za obučavanje, broj koraka zagrevanja, kao i optimizator *Adam* sa promenljivim rasporedom brzine učenja (10^{-3}, 10^{-4}, 10^{-5}). Parametri DL modela su podešeni ručno kroz niz pokrenutih eksperimenata kako bi se pronašle optimalne vrednosti za svaki zadatak, pristup i skup podataka. Skup podataka je podeljen u skupove za obuku, proveru i testiranje u odnosu 80:10:10. Pored toga, sve DL arhitekture koriste:

- Gradijentno spuštanje sa 32 instance u svakom skupu za proveru;
- Dimenziju ugnježdenih vektora jednaku 256 (zavisnu od modela za ugnježdavanje);
- MSL postavljenoj na 128 tokena;
- Dva BiLSTM sloja sa brojem jedinica jednakom MSL;
- Odnos isključivanja neurona od 0.1 za BiLSTM i 0.3 za FFN slojeve;
- Aktivacionu funkciju $ReLU$ u FFN slojevima i $softmax$ u završnom sloju.

Izabrana funkcija greške je CCE (pogledati jednačinu 7.8), koja se koristi kao standardna funkcija greške za višeklasnu klasifikaciju. Performanse predloženih pristupa su proverene korišćenjem F_1^{Ma} mere poređenjem sa performansama osnovnih i najboljih modela koji su predstavljeni za rešavanje **TD** i **IG** zadataka na *SemVal2019* takmičenju [65]. U pogledu izabranih tehnika za obradu ulaznih sekvenci, na oba zadatka je primenjeno maskiranje heš oznaka (#*hash*), korisničkih imena (@*user*) i adresa internet strana (*http*), koji se uobičajeno pojavljuju u sadržajima poruka preuzetih sa društvenih mreža. Ovom tehnikom postiže se uklanjanje nepotrebnih informacija iz teksta, uz očuvanje informacija o njihovom prisustvu.

8. Modelovanje emocionalnih i moralih aspekata u srpskom jeziku

8.1. Jezičke zavisnosti i ograničenja

Primena NLP tehnika na jezike sa manje razvijenim resursima, koji uključuju i srpski, suočava se sa brojnim izazovima, uključujući nedostatak jezičkih resursa (obeleženih korpusa i leksikona, kao i predobučenih modela), lingvističku složenost (morfološku, sintaksnu i fonološku) i ograničenu digitalnu pristupačnost, koji otežavaju razvoj i evaluaciju NLP alata. Rešenja se često oslanjaju na transfer znanja iz drugih jezika, polu-nadgledano učenje, kolaborativne inicijative za efikasno prikupljanje, obeležavanje i proveru jezičkih resursa, kao i uključivanje višemodalnih komponenti u korpusima. Ova ograničenja u obradi srpskog jezika karakterišu:

- **Morfološka složenost** – srpski jezik ima bogatu morfologiju sa sedam padeža, tri roda i složenim sistemom fleksija. Ova karakteristika jezika otežava normalizaciju teksta, posebno kada su dostupni alati ograničeni ili nisu dovoljno precizni. Za srpski jezik su razvijeni resursi za određivanje vrste reči i lematizaciju reči, u obliku leksikona [192] ili naprednih ML modela [191]. Dodatno, složene reči i fraze (izrazi), zahtevaju posebnu pažnju prilikom analize tekstualnih sekvenci. Za pronalaženje složenih reči u tekstovima na srpskom jeziku razvijeni su algoritmi za efektivnu podelu srpskih reči na slogove [92, 169], koji se mogu unaprediti primenom naprednih ML tehnika.
- **Izgovorne i pravopisne varijacije** – srpski jezik koristi dva pisma (ćirilicu i latinicu) i ima dva izgovora (ekavski, ijekavski), što može izazvati probleme u unifikaciji prilikom obrade teksta. Dodatno, način na koji se jezik koristi u svakodnevnoj komunikaciji može značajno varirati između različitih kultura i odstupati od formalnog načina izražavanja. Za srpski jezik je, prilikom korišćenja latiničnog pisma u neformalnom izražavanju (lične poruke, poruke na društvenim mrežama i druge), karakteristično često odstupanje od ispravnog korišćenja slova sa dijakriticima. Da bi se tekst napisan na ovakav način pravilno pripremio za dalju obradu u ML algoritmima, neophodno je ispravljanje svih uočenih nepravilnosti. Restauracija dijakritika je u ovom radu rešena korišćenjem REDI[42] ML modela za srpski jezik [123]. Drugi vidovi unapređenja bi podrazumevali razvoj matematičkih algoritama i leksičkih resursa za efektivno prepoznavanje izgovornih i leksičkih varijacija, kao i pronalaženje pravopisnih grešaka u tekstualnim sadržajima.
- **Nedovoljno razvijeni obeleženi resursi** – objavljeni obeleženi korpusi za srpski jezik su ograničeni u pokrivenosti, broju i veličini u poređenju sa korpusima za engleski jezik. Neki od do sada obeleženih i objavljenih korpusa u srpskom jeziku jesu:
  - *SrpKor4Tagging*[43] korpus književnih i administrativnih tekstova (∼343k tokena) obeležen na nivou lema i vrsta reči [191];
  - *INTERA*[44] korpus paralelnih uparenih rečenica na srpskom i engleskom jeziku sa obeležjima vrsta reči [208];

[42] https://github.com/clarinsi/redi
[43] https://live.european-language-grid.eu/catalogue/corpus/9295
[44] https://live.european-language-grid.eu/catalogue/corpus/685

– *SETimes.SR*[45] korpus novinskih tekstova (∼87k tokena) obeležen na nivou lema i vrsta reči, morfosintakse, sintaktičkih zavisnosti i imenovanih entiteta [20];

– *ELTeC*[46] (eng. *The European Literary Text Collection*) je kolekcija evropskih romana koju sačinjava i 100 romana na srpskom jeziku (∼4.93M reči), obeleženih na nivou lema, vrsta reči i 7 tipova imenovanih entiteta [190];

– *SentiComments.SR*[47] korpus filmskih recenzija obeleženih prema sentimentu [19];

– *SrpMD4Tagging*[48] morfološki rečnik srpskog jezika za obeležavanje leme i vrste reči [192].

Još uvek nedovoljan broj obeleženih korpusa na srpskom jeziku ograničava mogućnosti za obučavanje i dostizanje optimalne preciznosti modela za zadatke kao što su prepoznavanje imenovanih entiteta i njihovih relacija, sintaksne analiza teksta ili rešavanje specijalizovanih zadataka kao što su predviđanje emocionalnog afekta ili lažnih objava na internetu. Drugi problem predstavlja harmonizacija različitih načina obeležavanja (skupova obeležja i pristupa) u prethodno pomenutim obeleženim skupovima, uz napomenu da nijedan od pomenutih skupova nema ručno obeležene ili verifikovane sve slojeve, što je potrebno za obučavanje složenih modela.

- **Nedostatak obučenih modela** – višejezični modeli kao što su mBERT, XLM-R ili Čet-GPT, koji uključuju srpski, nisu optimalno podešeni za složenu morfologiju i fleksiju srpskog jezika. Razvoj lokalizovanih i jednojezičkih modela rešava ovaj potencijalni izazov. Do sada su za srpski jezik razvijeni lokalizovani transformer modeli kao što su Jerteh-81/Jerteh-355 [181] ili BERTić [124], koji zahtevaju dalje doobučavanje u cilju postizanja optimalnih performansi na širokom spektru zadataka nad tekstovima na srpskom jeziku.

- **Nedostatak specijalizovanih resursa** – za specifične domene (medicina, pravo, literatura) nedostaju specijalizovani korpusi ili modeli, što ograničava primenu NLP u realnim slučajevima korišćenja. Neke od prvih inicijativa u tom pravcu jesu razvoj književnog korpusa *SrpELTeC* [189], paralelnog italijansko-srpskog književnog korpusa *SerbItaKor* [134], leksikon dvojezičnih termina iz elektroenergetike [88] ili modela specijalizovanog za razumevanje pravnih tekstova [25].

Dostupnost neophodnih leksičkih resursa, obeleženih korpusa i ML modela za srpski jezik je često ograničena, što predstavlja jedan od najvećih izazova za analizu specifičnih jezičkih karakteristika. Rad na izgradnji i poboljšanju postojećih resursa za srpski jezik je obezbeđen kroz rad članova akademske zajednice, Društva za razvoj jezičkih resursa za srpski jezik – *JeRTeh*[49], nacionalnog naučnog projekta TESLA[50], međunarodnih naučnih projekata *ELE*[51] (eng. *European Language Equality*) i *ELG*[52] (eng. *European Language Grid*) uspostavljenih za razvoj i razmenu jezičkih resursa, kao i međunarodnih naučnih organizacija *ReLDI*[53] i *CLARIN*[54] sa usmerenjem ka istraživanju južnoslovenskih jezika. Za srpski jezik

[45] https://vukbatanovic.github.io/SETimes.SR/
[46] https://live.european-language-grid.eu/catalogue/corpus/11214
[47] https://vukbatanovic.github.io/SentiComments.SR/
[48] https://live.european-language-grid.eu/catalogue/lcr/9294
[49] https://jerteh.rs
[50] https://tesla.rgf.bg.ac.rs
[51] https://european-language-equality.eu
[52] https://live.european-language-grid.eu
[53] https://reldi.rs/
[54] https://www.clarin.si

do sada nisu razvijeni resursi za prepoznavanje emocionalnog afekta i moralne vrednosti, te će naredni deo istraživanja biti posvećen razvoju inicijalnih verzija ovih resursa.

8.2. Izgradnja sentimentalnih i emocionalnih semantičkih leksikona

U okviru ovog odeljka prikazan je razvoj semantičkih leksikona reči za prepoznavanje intenziteta sentimenta i emocionalnog afekta u rečima na srpskom jeziku. Razvijeni leksički resursi su značajni za izračunavanje dodatnih atributa teksta iz navedenih aspekata jezika u predloženoj metodologiji, ali i kao nezavisni resursi za razvoj naprednijih resursa za srpski jezik. Prikupljanje znanja neophodnih za izgradnju leksikona se može vršiti prevođenjem reči, odnosno leme reči na engleskom jeziku, na srpski jezik, ručnim putem ili korišćenjem automatskih alata, kao što je naširoko upotrebljavani Gugl prevodilac (eng. *Google Translate*, *GT*), a najčešće u upotrebi je poluautomatski pristup koji uključuje automatsko prevođenje i ručnu korekciju dobijenog prevoda. U tom procesu, od izuzetne je važnosti da se za datu lema_{En} ispravno odredi lema reči na srpskom jeziku koja bi joj u srpskom jeziku bila prevodni ekvivalent. Ovaj zadatak je zahtevan jer postojeći leksikoni na engleskom jeziku često ne sadrže PoS obeležja, zbog čega je postupak pronalaženja odgovarajućeg prevodnog ekvivalenta na srpskom jeziku značajno otežan.

8.2.1 SentiWords.SR

Po uzoru na leksikon engleskih reči *SentiWords*, u kojem su rečima dodeljene vrednosti sentimenta [63], razvijena je poluautomatska metoda za kreiranje srpske verzije ovog leksikona, nazvanog *SentiWords.SR*. Autori *SentiWords* leksikona uporedili su najčešće korišćene formule u istraživanjima zasnovanim na leksikonu *SentiWordNet* sa novim tehnikama i koristili ih kao ulazne atribute za ML modele, čime su postigli bolje rezultate u izračunavanju polariteta (sentimenta) reči u odnosu na druge metode zasnovane na *SentiWordNet* leksikonu. Za razliku od *SentiWordNet*, *SentiWords* pridružuje vrednosti sentimenta direktno rečima, nezavisno od konteksta u kojem se reč pojavljuje. Ove vrednosti, poznate kao prvobitni (eng. *prior*) polariteti, označavaju polaritet reči bez obzira na kontekst upotrebe, dok su posteriorni (eng. *posterior*) polariteti zavisni od konteksta. Pošto je izveden iz verzije 3.0 PWN leksikona, ovaj rečnik obuhvata približno 155,000 reči, što ga čini jednim od najobimnijih leksikona polariteta za engleski jezik. Reči u leksikonu sa intenzitetom polariteta većim od 0 obeležene su kao *Pozitivne*, manjim od 0 kao *Negativne*, a jednakim 0 kao *Neutralne* (tabela 8.1).

Tabela 8.1: *Statistika SentiWords leksikona prema vrsti reči i obeležjima sentimenta*

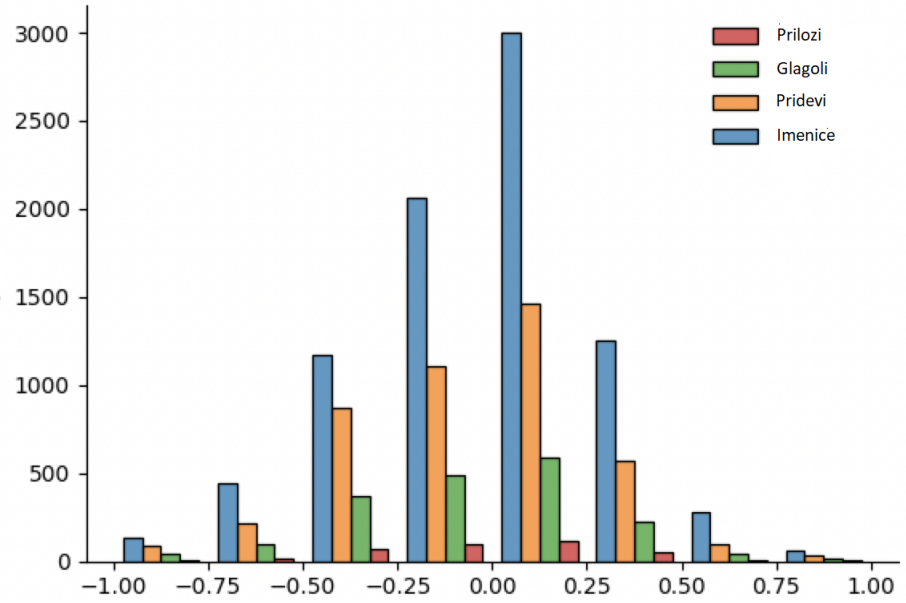
PoS	Pozitivne	Negativne	Neutralne	Total
Imenice	13,287	14,680	89,692	117,659
Pridevi	5,814	7,598	7,835	21,247
Glagoli	35,88	3,052	4,889	11,529
Prilozi	2,271	5,40	1,670	4,481
Total	24,957	25,870	104,086	15,4913

Za izgradnju srpske verzije leksikona razvijen je algoritam (pogledati algoritam 8.1), koji koristi lema_{En}-PoS parove iz engleskog *SentiWords* leksikona i korišćenjem GT alata, automatski prevodi sve leme od jedne reči (PoS $\neq$ „MWE“) u leksikonu sa dodeljenim intenzitetom sentimenta $\neq 0$. Na ovaj način je dobijeno ukupno 41,843 lema_{En}-PoS parova,

čija je ispravnost proverena u narednim koracima. Tokom provere, prevodni ekvivalenti su kategorisani u potpuno tačne, delimično tačne - prevod leme ili dodeljeno PoS obeležje je pogrešno, i potpuno pogrešne - prevod nedostaje ili je nekorektan zajedno sa PoS obeležjem. Sakupljeni su svi potpuno tačni prevodi (20,244 ili 48.3%) i ispravljene verzije delimično pogrešnih prevoda (7,528 ili 18%), dok su potpuno pogrešni prevodi (14,071 ili 33.6%) izostavljeni iz dalje obrade.

Algoritam 8.1: Kreiranje *SentiWords.SR* leksikona

FindPolarity
- **inputs:** *SentiWords*; *GT*
- **output:** *SentiWords.SR*
- **foreach** $lema_{En}$, *PoS* $\in$ *SentiWords* **do**
  - score $\leftarrow$ score(lema_{En}, PoS);
  - lema_{Sr}, PoS $\leftarrow$ *clean*(*GT*(lema_{En}, PoS));
  - lema_{Sr}, PoS $\leftarrow$ *score*(lema_{En}, PoS);
  - *SentiWords.SR* $\leftarrow$ *evaluate*(lema_{Sr}, PoS, *score*);
- **foreach** $lema_{Sr}$, *PoS* $\in$ *SentiWords.SR* **do**
  - *score* $\leftarrow$ *mean*(lema_{Sr}, PoS, *score*);
  - *std* $\leftarrow$ *std*(lema_{Sr}, PoS, *score*);
  - *count* $\leftarrow$ *count*(lema_{Sr}, PoS);
- **return** *SentiWords.SR*;



Slika 8.1: *Distribucija intenziteta polariteta reči u SentiWords.SR leksikonu prikazana po vrstama reči na segmentima dužine 0.25 u okviru intervala [-1, +1]*

Ispravnim kombinacijama lema_{Sr}-PoS na srpskom jeziku (27,772 ili 66.4%) je dodeljena ista vrednost intenziteta polariteta koju ima odgovarajući lema_{En}-PoS par. Konačno, za svaku kombinaciju lema_{Sr}-PoS je izračunata srednja vrednost intenziteta polariteta i dodatne statistike kao što su standardna devijacija intenziteta i broj pojavljivanja u leksikonu. Dodeljene vrednosti intenziteta polariteta su ocenjivali anotatori, koji su proveravali dodeljeni intenzitet i njegovu orijentaciju. Dodatna statistika je pomogla u procesu provere, jer reči sa standardnom devijacijom > 0 i brojem pojavljivanja > 1 mogu biti pokazatelj potencijalnih grešaka koje mogu nastati prilikom automatske obrade. Uspostavljena šema

za obeležavanje ograničava vrednosti polariteta na opseg [-1, +1], pri čemu su najjače vrednosti polariteta bliže granicama opsega, a slabije vrednosti polariteta bliže 0.

Tabela 8.2: *Statistika SentiWords.SR leksikona prema vrsti reči i obeležjima sentimenta*

PoS	Pozitivne	Negativne	Neutralne	Total
Imenice	4,512	3,791	105	8,408
Pridevi	2,143	2,258	44	4,445
Glagoli	866	988	18	1,872
Prilozi	164	178	0	342
Total	7,685	7,215	167	15,067

Nakon ručne provere, svi lema_{Sr}-PoS parovi sa agregiranim intenzitetom polariteta jednakim 0 su izostavljeni iz leksikona. Reči u leksikonu su upoređene sa rečima u *SentiPol.SR* leksikonu, sa kategoričkim obeležjima sentimenta (pozitivan, negativan) koji je napravljen za analizu sentimenta u korpusu srpskih novela iz perioda 1840–1920. [189]. Leme prisutne u *SentiPol.SR*, ali nedostajuće u *SentiWords.SR* leksikonu, u ukupnom broju od 1,281, su prikupljene, pregledane i obeležene istom šemom obeležavanja. U konačnoj verziji, *SentiWords.SR* leksikon sadrži 15,067 jedinstvenih lema_{Sr}-PoS parova (tabela 8.2), koje su u mogućnosti da identifikuju ukupno ∼210,000 različitih flektivnih oblika srpskih reči.

Kao što je prikazano u tabeli 8.2 i vidljivo na slici 8.1, leksikon ima blagu pristrasnost prema lemama sa pozitivnim sentimentom (7,685, što odgovara 51.7%) u poređenju sa negativno obeleženim lemama (7,215, što odgovara 48.3%). Pored toga, većina lema je centrirana oko 0, sa približno 57% lema sa intenzitetom polariteta u opsegu [-0.25, 0.25]. Nasuprot tome, primećuje se da leksikon sadrži nešto više izrazito negativnih lema ($\leq$-0.75) nego izrazito pozitivnih ($\geq$0.75), pri čemu 6.2% lema ima skor u opsegu [-1, -0.5], dok pozitivni rezultati iz opsega [0.5, 1] čine 5.4% lema. U odnosu na PoS distribuciju, leksikon sadrži približno 56% imenica, 29% prideva, 12% glagola i 2% priloga obeleženih intenzitetom polariteta. Leksikon *SentiWords.SR* je javno objavljen na *ELG* platformi, koji naučna zajednica može da koristi i dalje unapređuje[55].

Merenje intenziteta sentimenta u tekstovima

Po završetku procesa izgradnje leksikona srpskih reči *SentiWords.SR* sa pridruženim vrednostima intenziteta sentimenta, razvijen je programski alat za izračunavanje intenziteta sentimenta za srpski jezik (eng. *Serbian Polarity Framework, SRPOL*), koji implementira algoritam za izračunavanje intenziteta sentimenta u tekstualnim sekvencama na srpskom jeziku [184]. SRPOL algoritam procenjuje kontekstualno značenje reči u tekstualnoj sekvenci, primenom sledećih kontekstualnih pravila koja imaju ključnu ulogu u izračunavanju intenziteta sentimenta tekstualne sekvence:

- Prilozi kao modifikatori intenziteta sentimenta nastupajuće fraze u tekstu;
- Negacijski signali koji preokreću polaritet sentimenta u nastupajućoj frazi;
- Uzvičnici kao pojačivači sentimenta prethodećeg tekstualnog segmenta;
- Produžene reči (reči sa višestruko ponovljenim slovima) kao pojačivači sentimenta originalne reči;

[55] https://live.european-language-grid.eu/catalogue/lcr/23614

- Segmentacija na smislene morfološke delove kao što su rečenica ili deo rečenice kao jedinica mere za izračunavanje sentimenta.

SRPOL algoritam najpre ulazni tekst deli na segmente koji su u inicijalnoj konfiguraciji podešeni na rečenice ili delove rečenica do pojave znaka zareza. Primarni cilj podele teksta na segmente je da se pomogne u poboljšanju tačnosti prilikom izračunavanja sentimenta na dužim tekstualnim sekvencama na čijim delovima se mogu uočiti delovi različitog polariteta. Podela teksta na segmente može se izvesti na nekoliko različitih načina i predstavlja polje širokog spektra istraživačkih studija [146]. Na primer, autori u [87] koriste kontrastivni veznik *„ali“* za promenu polariteta sentimenta, pri čemu se sentiment teksta koji sledi posle veznika smatra dominantnim i diktira ukupnu ocenu sentimenta tekstualne sekvence. U ovom radu segment je predstavljen jednom rečenicom ili delom rečenice do pojave znaka zareza. Polaritet segmenta je zbir intenziteta pojedinačnih reči podeljen brojem reči koje doprinose ukupnom intenzitetu sentimenta, odnosno srednja vrednost intenziteta pojedinačnih sentimentalnih reči koje se pojavljuju na tom segmentu (pogledati jednačinu 8.1).

$$P_s = \frac{\sum_i^k P_w^i}{k} \tag{8.1}$$

Sentiment tekstualne sekvence izračunavamo kao težinsku sredinu intenziteta njegovih segmenata (pogledati jednačinu 8.2):

$$P_{text} = \frac{\sum_i^S w_i * P_s^i}{\sum_i^S w_i}, w_i = \sum_m^S |sign(P_s^i) = sign(P_s^m)| \tag{8.2}$$

gde je S broj segmentata u tekstu, P_i je polaritet i w_i je faktor težine i-tog segmenta. Na ovaj način smo u mogućnosti da merimo polaritet dužih tekstova sa međusobno suprotnim intenzitetima sentimenta identifikovanim na njegovim segmentima.

Segmenti, odnosno rečenice u tekstu, se pronalaze korišćenjem specijalizovane metode iz *nltk* paketa *PunktSentenceTokenizer*[56], kojoj se mogu definisati obeležja za kraj rečenice. SRPOL algoritam za kraj rečenice koristi znakove interpunkcije kao što su ., ,, ..., ?, !, kao i specijalne kombinacije znakova koji predstavljaju emotikone. Svaki identifikovani segment se zatim deli na tokene pomoću *RegexpTokenizer*[57] tokenizatora, uz istovremeno uklanjanje stop reči. Uklanjanje stop reči nije od presudnog značaja za rad SRPOL algoritma jer se on prvenstveno oslanja na imenice, prideve, priloge i glagole koji se u tekstu mogu pronaći, već predstavlja dodatnu proveru ispravnosti vrste reči koje se u okviru algoritma obrađuju i ubrzava njegovo izvršavanje.

Nakon tokenizacije i uklanjanja stop reči, tokeni su obeleženi PoS obeležjima i lematizovani uz pomoć izgrađenog modela za lematizaciju i određivanje vrste reči u srpskom jeziku [191]. U skladu sa BoW tehnikom, svaki lema_{Sr}-PoS par se zatim upoređuje sa SRPOL leksikonom sentimenta *SentiWords.SR*. Ukoliko je lema_{Sr}-PoS prisutna u leksikonu, odgovarajuća vrednost intenziteta sentimenta leme se dodaje u konačni zbir segmenta. Prilikom pojave negacijskog signala, rezultat tekućeg intenziteta sentimenta menja svoj znak. Ukoliko se pojavi modifikator sentimenta, sentiment tekuće fraze se množi intenzitetom modifikatora. U složenijim slučajevima, kod pojavljivanja negacija sa modifikatorima, primenjuje se kompozicija pravila za modifikovanje intenziteta sentimenta (pogledati tabelu 8.3).

Slično, ukoliko tekst sadrži uzvičnik, rezultat rečenice se množi sa 1.06 ($n < 2$) ili 1.18 ($n \geq 2$), gde je n broj neposrednih uzvičnika u tekstu. Produžene reči se menjaju

[56] https://www.nltk.org/api/nltk.tokenize.PunktSentenceTokenizer.html
[57] https://www.nltk.org/api/nltk.tokenize.RegexpTokenizer.html

Tabela 8.3: *Efekat negacijskih signala u kombinaciji sa prilozima i negacijama kao modifikatorima intenziteta sentimenta*

Jezik	Fraza	Sent.
Sr	„**Nije** (→ NEG) uradio (p=+0.2)...“	-0.2
En	„**Not** (→ NEG) done (p=+0.2)...“	
Sr	„**Nije** (→ NEG) **zaista** (→ MOD=1.2) uradio (p=+0.2)...“	-0.17
En	„**Not** (→ NEG) **really** (→ MOD=1.2) done (p=+0.2)...“	
Sr	„**Niko** (→ MOD=1.2) **nije** (→ NEG) uradio (p=+0.2)...“	-0.24
En	„**Nobody** did [**not** (→ NEG)] do (p=+0.2)...“	
Sr	„**Niko** (→ MOD=1.2) **nikada**(→w MOD=1.2) **nije** (→ NEG) uradio (p=+0.2)...“	-0.29
En	„**Nobody** has **never** (→ MOD=1.2) [**not** (→ NEG)] done(p=+0.2)...“	

standardnim oblikom reči iz koje je produžena reč izvedena, a rezultat sentimenta se množi sa 1.05^n, gde je n broj ponovljenih slova u produženoj reči. Konačno, polaritet pojedinačnih segmenata i tekstualne sekvence koja se sastoji iz datih segmenata se izračunava kao što je predstavljeno jednačinama 8.1 i 8.2, i ograničen je na [-1, +1] numerički opseg.

Na osnovu brojnih izvršenih eksperimenata, pokazano je da je SRPOL algoritam primenljiv za izračunavanje intenziteta sentimenta na tekstovima na srpskom jeziku iz različitih domena [184]. Alat se razlikuje od tradicionalne BoW tehnike po tome što uključuje kontekstualna pravila koja uključuju negacije, modifikatore priloga, uzvičnike, produžene reči, emotikone, kao i segmentaciju teksta. Prema svojoj konstrukciji, SRPOL alat sa ugrađenim skupovima pravila razrešava složenost kontekstualnih zavisnosti u srpskom jeziku, pošto se sentiment ne nalazi samo u pojedinačnim rečima, već zavisi i od konteksta u kojem se reči pojavljuju.

Glavna metodološka ograničenja SRPOL alata proizilaze iz pretpostavki koje prate korišćenje BoW tehnike. Analizirajući reči u tekstu pojedinačno, ova tehnika potencijalno kompromituje aspekte značenja jer ne uzima u obzir kontekst u kome se reči pojavljuju. SRPOL takođe ima nedostatak u tumačenju izraza u kojima su emocije iskazane metaforično, sarkastično ili ironično. Pored toga, sličan nedostatak se susreće i sa tumačenjem homonima i homografa. Važno je napomenuti da se SRPOL oslanja na pretpostavku da sve fleksije reči nose isti sentiment. Na primer, flektivni pridevi kao što je pridev *„loš“* (eng. *„bad“*) imaju isti intenzitet sentimenta za sve flektivne komparacije, iako komparativ *„gori“* (eng. *„worse“*) i superlativ *„najgori“* (eng. *„the worst“*) imaju relativno veći stepen negativnog sentimenta. Pored toga, tačnost SRPOL alata u velikoj meri zavisi od tačnosti modela za lematizaciju i određivanje vrsta reči koji se koriste za procesiranje teksta.

8.2.2 SWN-Affect

Proširenje PWN leksikona dodavanjem kategorija afekata prilikom kreiranja WNA leksikona [194] predstavlja važan korak koji unapređuje razumevanje emocija i osećanja u engleskom jeziku (pogledati odeljak 6.3.1 u okviru odeljka 6.3). Iako se WNA odnosi na semantiku engleskog jezika, u okviru naučnih istraživanja napravljeni su napori da se WNA resurs prilagodi za korišćenje u drugim jezicima poput rumunskog i ruskog [23] ili japanskog [201]. Poravnavanjem PWN sinseta sa sinsetima u srpska verzija WordNet leksikona (eng. *Serbian WordNet Lexicon, SWN*) leksikonu, koristeći uspostavljeni *ILI* konektor za grupu evropskih jezika u okviru EuroWordNet - EWN [209] projekta, kategorije afekta iz WNA leksikona se potencijalno mogu proširiti i na srpski jezik. Ograničenje WNA leksikona leži u njegovoj povezanosti sa verzijom PWN-1.6, što sprečava direktnu integraciju

sa leksikonima PWN-3.0 ili SWN, koji je zasnovan na PWN-3.0, za prikupljanje kategorija afekta. Identifikovani izazov je rešen primenom:

1. Preslikavanja između engleskih sinseta u verzijama PWN 1.6 i 3.0, uz očuvanje semantičkog značenja između sinseta;
2. Spajanja WNA sinseta sa PWN 3.0 sinsetima koristeći mapiranje napravljeno u prethodnom koraku;
3. Spajanja SWN sinseta sa WNA sinsetima koristeći uspostavljeni *ILI* konektor u SWN leksikonu.

Tabela 8.4: *Isečak iz WNA.SR leksikona za kategoriju joy*

lema_{Sr}	PoS	Poz.	Neg.	lema_{En}	Sinset	Offset-v16	Offset-v30
„veselje“	noun	0.500	0.250	joy	joy.n.01	5596218	7527352
„slavljenje“	noun	0.375	0.000	joy	joy.n.01	5596218	7527352
„razdragan“	adj	0.125	0.625	joyful	joyful.a.01	1100759	1363613
„ushićen“	adj	0.500	0.375	elated	elated.s.02	1313944	1367211
„raspoložen“	adj	0.750	0.125	elated	elated.s.02	1313944	1367211
„radosno“	adv	0.750	0.125	gleefully	gleefully.r.01	342819	348247
„vedro“	adv	0.750	0.125	gleefully	gleefully.r.01	342819	348247
„obradovati“	verb	0.125	0.125	gladden	gladden.v.01	1237403	1813499
„razveseliti“	verb	0.125	0.125	gladden	gladden.v.01	1237403	1813499

Preslikavanje između sinseta iz dve različite verzije PWN leksikona uspostavljeno je korišćenjem Wu-Palmerove mere sličnosti između naziva sinseta. Ova mera računa povezanost uzimajući u obzir dubine dva sinseta u taksonomijama PWN leksikona i čvor najspecifičnijeg pretka [215]. Za svaki sinset u WNA, pronađen je najsličniji PWN-3.0 sinset, kome je dodeljena ista kategorija afekta iz WNA taksonomije koju ima originalni PWN-1.6 sinset [201]. Konačno, SWN sinseti su povezani sa najsličnijim pronađenim PWN-3.0 sinonimskim skupom u prethodnom koraku pomoću *ILI* konektora. Konstruisanim pristupom, SWN sinsetima i odgovarajućim leksičkim oblicima koji im pripadaju, su dodeljene kategorije afekta iz WNA afektivnog leksikona kako je prikazano u Tabeli 8.4 sa primerima unosa za kategoriju *joy*.

Tabela 8.5: *Mapiranje između Plutčikovih i WNA emocionalnih kategorija*

NRC	WNA
fear	*ambiguous-fear, gravity, daze, shame, anxiety, negative-fear, scare, horror, shyness, timidity, diffidence, annoyance, negative-concern*
sadness	*apathy, pensiveness, compassion, sadness, melancholy, regret-sorrow, grief, lost-sorrow*
anger	*despair, ingratitude, general-dislike, anger, annoyance, bad-temper, oppression, hate, displeasure, bad-temper*
anticipation	*ambiguous-expectation, positive-expectation, positive-hope, forgiveness*
surprise	*ambiguous-agitation, surprise, stupefaction, astonishment*
disgust	*neutral-unconcern, thing, ingratitude, general-dislike, disgust, antipathy, dislike, repugnance*
trust	*fearlessness, self-pride, humility, encouragement, approval, confidence, security, belonging, closeness, favor*
joy	*levity, positive-fear, enthusiasm, calmness, joy, gratitude, affection, love, liking, happiness, joy-pride, cheerfulness, euphoria, satisfaction, general-gaiety, jollity, sympathy, positive-concern*

U završnici, konstruisani *WNA.SR* leksikon sadrži 630 sinseta i 1,003 leksička oblika iz SWN leksikona obeleženih u jednu ili više, od 226 postojećih, kategorija afekata. Ukrštanjem *WNA.SR* sa *EmoLex.SR* emocionalnim leksikonom i korišćenjem uspostavljenog mapiranja između emocionalnih kategorija u dva resursa (pogledati Tabelu 8.5), dodeljena obeležja u *EmoLex.SR* se mogu uporediti sa emocionalnim obeležjima *WNA.SR* leksikona za odgovarajuće leksičke oblike i na taj način se može izvršiti dodatna provera obeležavanja. Prilikom uspostavljanja veze između WNA i SWN leksikona izvršavanjem navedenih koraka, identifikovano je 219 sinseta u PWN za koje još uvek nisu postojali unosi u SWN leksikonu i oni su, tokom ovog istraživanja, iskorišćeni da se SWN leksikon dodatno unapredi. Šema SWN leksikona je, takođe, obogaćena obeležjima afekata uspostavljanjem novih XML etiketa <AFFECT> i <EMOCAT> u okviru etikete <SYNSET> u postojećoj SWN XML šemi. Etiketa <AFFECT> predstavlja emocionalni afekt preuzet iz WNA leksikona, dok etiketa <EMOCAT> predstavlja njoj odgovarajuću Plutčikovu kategoriju dobijenu korišćenjem uspostavljenog preslikavanja prikazanog u tabeli 8.5. Unapređena SWN šema je prikazana na primeru *ENG30-01797051-v* sinseta:

```
</SYNSET>
   <SYNSET>
      <ID>ENG30-01797051-v</ID>
      <POS>v</POS>
      <SYNONYM>
         <LITERAL>oplakati<SENSE>1</SENSE>
            <LNOTE>V123+Perf+Tr+Iref</LNOTE></LITERAL>
         <LITERAL>tugovati<SENSE>1b</SENSE>
            <LNOTE>V518+Imperf+It+Iref</LNOTE></LITERAL>
         <LITERAL>jadikovati<SENSE>1</SENSE>
            <LNOTE>V518+Imperf+It+Iref</LNOTE></LITERAL>
         <LITERAL>ožaliti<SENSE>1</SENSE>
            <LNOTE>V153+Perf+Tr+Iref</LNOTE></LITERAL>
      </SYNONYM>
      <DEF>Osećati tugu.</DEF> ...
      <SUMO>IntentionalPsychologicalProcess<TYPE>+</TYPE></SUMO>
      <SENTIMENT>
         <POSITIVE>0.00000</POSITIVE>
         <NEGATIVE>0.75000</NEGATIVE>
      </SENTIMENT>
      <DOMAIN>psychology</DOMAIN>
      <AFFECT>regret-sorrow</AFFECT>
      <EMOCAT>sadness</EMOCAT>
</SYNSET>
```

Fragment koda 8.1: *Primer sinseta iz SWN leksikona sa dodatnim oznakama afekta <AFFECT> i <EMOCAT>*

Na ovaj način su emocionalne kategorije, zajedno sa ostalim atributima sinseta, postale dostupne za detaljnu analizu semantike srpskog jezika i za druga istraživanja u kojima se koristi SWN leksikon.

8.2.3 EmoLex.SR

Koncept konstrukcije leksikona

Kako bi se prevazišao nedostatak resursa za emocionalni afekt u srpskom jeziku, razvijena je poluautomatska metoda za formiranje srpskog leksikona emocionalnog afekta, koji se u daljem tekstu naziva *EmoLex.SR*. Reči u leksikonu su obeležene korišćenjem osam osnovnih emocionalnih Plutčikovih kategorija: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*,

surprise i *trust*, uz dodatnu kategoriju *neutral* koja označava odsustvo emocionalnog afekta u reči. Kako bi se olaksalo reprodukovanje rezultata dobijenih u procesu konstrukcije leksikona *EmoLex.SR*, kompletan izvorni kod iz ovog dela istraživanja objavljen je na *GitHub* repozitorijumu[58]. Napravljeni leksikoni su javno objavljeni na *ELG* repozitorijumu[59] jezičkih resursa da bi se obezbedila njihova dostupnost za korišćenje i dalje unapređivanje u naučnim istraživanjima.

Inspirisani uspehom alata Čet-GPT na mnogim zadacima iz oblasti obrade prirodnih jezika [106], ovaj alat je iskorišćen kao pomoćni alat za razvoj leksikona na srpskom jeziku. Kroz ovo istraživanje, mogućnosti alata Čet-GPT, specifično LLM *gpt-3.5-turbo*, su korišćene za rešavanje sledećih zadataka (pogledati tabelu C.1 i C.4 u Dodatku C za posebna podešavanja u svakom zadatku):

T1 Prevođenje reči sa dodeljenom vrstom reči – u cilju smanjivanja dvosmislenosti značenja reči u toku prevođenja;

T2 Obeležavanje reči sa dodeljenom vrstom reči u kategorije emocionalnog afekta – u cilju podrške proveri kategorija emocionalnog afekta preuzetih iz engleskog jezika;

T3 Dobijanje sinonima za reč sa dodeljenom vrstom reči – u cilju proširivanja postojećeg skupa sinonima;

T4 Generisanje paralelnih rečenica obeleženih kategorijama emocionalnog afekta – u cilju provere ispravnosti leksikona u jednojezičnom i višejezičnom kontekstu.

Iako rezultati dobijeni iz LLM modela pružaju bogate i raznolike informacije, često mogu pokazivati pristrasnost i nekonzistentnost, što zahteva pažljivu proveru prilikom formiranja resursa. Iz pomenutog razloga su sve rezultate generisane uz pomoć Čet-GPT alata ručno proverili lingvistički stručnjaci ili je verifikacija sprovedena putem postojećih resursa za srpski jezik, kao što su SWN leksikon i modeli za obeležavanje vrste reči i lematizaciju. Osim SWN leksikona, koji je u ovom istraživanju korišćen za dobijanje sinonima, njegova verzija obogaćena emocionalnim afektom, predstavljena u Sekciji 8.2.2, korišćena je za proveru usvojenih i generisanih obeležja emocionalnog afekta.

Pristup korišćen za izračunavanje kategorije afekta u ovoj studiji se oslanja na ponavljanja reči sa istim značenjem u svakoj fazi procesa konstrukcije leksikona. U koraku agregacije kategorija, binarna obeležja afekata su inicijalno normalizovana između kategorija, sa pretpostavkom uspostavljenom u koraku anotacije da dodeljena obeležja nose jednake težine:

$$flag_w = (flag_w^1, flag_w^2, ..., flag_w^{|L|}), |L| = 8 \tag{8.3}$$

gde je $flag_w$ vektor sa binarnim obeležjima ($flag_w^i$ sa vrednostima 0 ili 1) za reč w i sve emocionalne kategorije L.

$$score_w^i = \frac{flag_w^i}{\sum_{i=1}^{L} flag_w^i}, i = 1, 2, ..., L \tag{8.4}$$

gde je $score_w^i$ težina za i-tu kategoriju reči w.

Analogno metodi skaliranja poređenjem sa najboljim-najgorim u grupi [101], ukoliko se reč pojavljuje više puta u različitim kontekstima, vrednost emocionalnog afekta za svaku

[58] https://github.com/milena-sosic/EmoLex.SR
[59] https://live.european-language-grid.eu/catalogue/lcr/23616

kategoriju je uprosečena (jednačina 8.5). Na primer, frekvencija pojavljivanja reči za specifičnu kategoriju (najbolja osobina) se deli sa ukupnim brojem pojavljivanja date reči u svim identifikovanim kategorijama (najbolje plus najgore osobine):

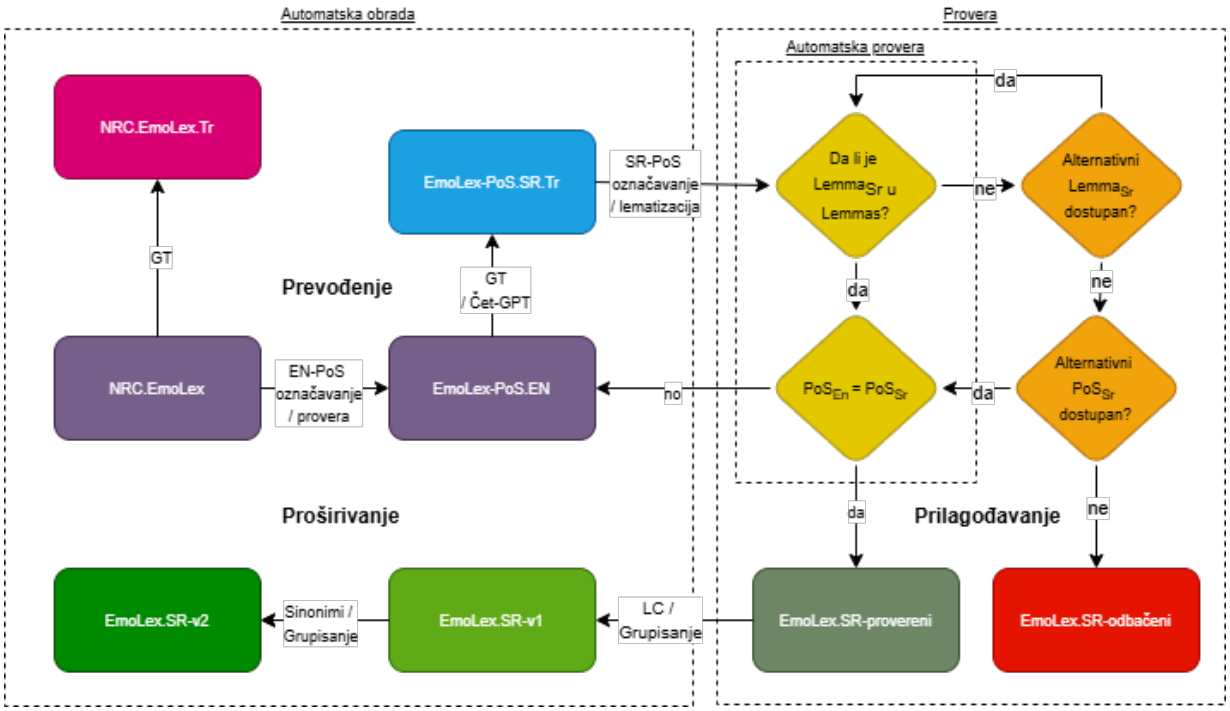
$$ascore_w^i = \frac{\sum_{j=1}^{n} score_w^{i_j}}{n}, n = count_w, i = 1, 2, ..., L \tag{8.5}$$

gde je $ascore_w^i$ prosečna težina za i-tu kategoriju reči w.

Koristeći ovaj pristup, rečima su dodeljene numeričke težinske vrednosti iz opsega [0, +1] za svaku emocionalnu kategoriju. Pored leksikona sa kontinuiranim vrednostima obeležja, razvijen je i kategorički leksikon sa diskretnim vrednostima kako bismo omogućili analizu pojavljivanja reči povezanih sa specifičnim afektima u tekstovima, često potrebnim u psihološkim i društvenim studijama:

$$aflag_w^i = \underset{i}{\operatorname{argmax}}\, ascore_w^i, i = 1, 2, ..., L \tag{8.6}$$

gde je $aflag_w^i$ novo binarno obeležje za kategoriju i i reč w.



Slika 8.2: *Opšti poluautomatski algoritam za kreiranje leksikona srpskog jezika počevši od engleske verzije*

Predložena metoda za razvoj leksikona obuhvata proces iz tri koraka koji uključuje automatsku obradu, automatsku i ručnu proveru, naizmenično primenjene kroz tri glavne faze razvoja leksikona: prevođenje, prilagođavanje i proširivanje (pogledati Sliku 8.2). Automatska obrada podrazumeva prevođenje engleskog leksikona NRC.EN korišćenjem dostupnih alata za prevođenje kao što je GT kao i daljih koraka za strukturiranje leksikona (dodavanje sinonima, grupisanje reči, agregacija vrednosti kategorija). Prema našim saznanjima, GT Programski interfejs aplikacije (eng. *Application Programming Interface, API*) nema mogućnost razlikovanja reči po vrsti, već pruža prevode zasnovane na najdominantnijim značenjima tih reči. Čet-GPT alat koji je izgrađen nad generativnim modelom ima mogućnost da se vrsta reči dostavi kao kontekstualna informacija prilikom prevođenja.

Prevedene reči su zatim automatski proverene pomoću izgrađenih resursa za lematizaciju i prepoznavanje vrste reči. Deo reči koji nije potvrđen kroz automatsku proveru valjanosti je ručno je proveren radi mogućeg uključivanja. Automatski i ručni koraci provere valjanosti prevodnih ekvivalenata su deo faze prilagođavanja leksikona. Faza proširivanja podrazumeva unošenje sinonima i dodatnih reči za koje je u istraživanjima jezika potvrđeno da nose emocionalni afekt.

Prevođenje

Reči iz NRC.EN leksikona su prevedene na srpski jezik uz očuvanje semantičkih i emocionalnih značenja, koristeći kombinaciju alata za automatsko prevođenje GT i Čet-GPT. Algoritam najpre uzima sve reči iz NRC.EN leksikona i automatski prevodi sve reči sa obeležjem $\neq$ *neutral* koristeći GT alat. Prevodi se zatim automatski proveravaju kako bi se osigurala tačnost srpske leme i pripadajuća PoS obeležja, u skladu sa sledećim pravilom:

$$PoS_{Sr} = PoS_{En} \wedge lema_{Sr} \in Leme_{Sr} \tag{8.7}$$

gde je $Leme_{Sr}$ skup ispravnih lema u srpskom jeziku [191, 192]. Tokom automatske provere ispravnosti, prevodi su kategorizovani u kategorije *ispravni* i *neispravni*, prema ispunjenosti pravila 8.7. *Neispravni* prevodi su definisani kao rezultat netačnih oblika $lema_{Sr}$ ili neslaganja u dodeljenim PoS obeležjima u poređenju između dva jezika. Prema korišćenom pravilu 8.7, GT je ispravno preveo 3,696 (62.5%) od ukupno 5,917 reči iz NRC.EN engleskog leksikona.

Sve reči leksikona su takođe prevedene pomoću Čet-GPT modela. Pomoću tehnike prompt inženjeringa, od Čet-GPT je zatraženo da prevede reč sa pridruženim PoS obeležjem na srpski jezik, kroz tri odvojene sesije za svaki $lema_{En}$-PoS par (pogledati prilog C). Najbolji $lema_{Sr}$-PoS par je izabran pomoću većinskog glasanja iz tri dobijena odgovora koje zadovoljavaju pravilo $PoS_{Sr} = PoS_{En}$. Prateći pravilo (8.7), Čet-GPT je ispravno preveo 5,030 (85.0%) od 5,917 reči iz leksikona. Identični prevodi dobijeni pomoću dva alata su kategorizovani kao *Potvrđeni*, dok su preostali prevodi obeleženi kao *Nepotvrđeni* i zahtevali su dalju ručnu proveru ispravnosti.

Prilagođavanje

U ovoj fazi smo proverili i prilagodili stavke od prethodnog koraka uzimajući u obzir lingvistička i kulturološka prilagođavanja (eng. *Linguistic and Cultural adjustments, LC*) karakteristike specifične za govornike srpskog jezika. Ovaj proces je podrazumevao ekspertsku analizu, ručnu doradu i kontekstualno prilagođavanje kako bi se obezbedilo da:

- *Nepotvrđeni* prevodi mogu biti ispravljeni;
- Dodeljene kategorije emocionalnog afekta relevantno odslikavaju srpske jezičke i kulturološke karakteristike.

Za *Nepotvrđene* prevode, anotatori su proveravali da li prevod postoji na srpskom jeziku u izvornom obliku ili kao njegova derivacija. Na primer, prevod $lema_{Sr}$-PoS para („*adventure*“, VERB) na srpski može biti („*doživeti avanturu*“, MWE). Engleska reč „*adventure*“ primarno se koristi kao imenica („*avantura*“, NOUN), dok se u slučaju glagola prevodni ekvivalent je višerečni izraz koji ima značenje učestvovanja u uzbudljivoj ili rizičnoj aktivnosti. Ukoliko se jednorečni prevod ne može pronaći tako da sačuva značenje i vrstu reči, potencijalno se zameniti prevodnim ekvivalentima kao što su „*istraživati*“, „*krenuti*“, „*usuditi se*“, „*rizikovati*“ ili „*iskusiti*“. Slično tome, polisemne reči u engleskom jeziku, kao što su „*bank*“ ili

„*interest*“, imaju različite prevodne oblike i prateća emocionalna obeležja u srpskom jeziku, koje nije moguće lako otkriti automatskim alatima bez razumevanja konteksta u kojem su upotrebljene. Nakon ovakve provere, sve moguće i odgovarajuće korekcije su sprovedene, a pripadajući lema_{Sr}-PoS parovi označeni su kao *Potvrđeni*. Suprotno tome, *Nepotvrđeni* lema_{Sr}-PoS parovi obeleženi su za isključivanje.

Pored toga, za svaku *Potvrđenu* reč u leksikonu su napravljena dodatna obeležja emocionalnih afekata korišćenjem dostupnih resursa za srpski jezik (*WNA.SR*) i naprednih LLM modela (Čet-GPT). Leksikon je najpre objedinjen sa *WNA.SR* leksikonom iz koga su preuzeta emocionalna obeležja koja ovaj leksikon ima. Obeležja u *WNA.SR* leksikonu su mapirana na emocionalna NRC obeležja koristeći pravila mapiranja navedena u Tabeli 8.5. Reči su zatim obrađene pomoću Čet-GPT alata da bi se dobilo obeležje kategorije emocionalnog afekta (pogledati prilog C, tabela C.1). Koristeći pristup većinskog glasanja nad Čet-GPT obeležjima dobijenim kroz tri nezavisne sesije, izabrano je konačno Čet-GPT obeležje za svaku srpsku reč.

Emocionalna obeležja reči preuzeta iz NRC.EN leksikona su proverila dva anotatora, srpskog maternjeg jezika. Provera je izvršena upoređivanjem sa emocionalnim obeležjima dobijenim iz *WNA.SR* leksikona i modela Čet-GPT. Za procenu međusobne usaglašenosti između anotatora korišćena je statistička mera Fleisova kapa (eng. *Fleiss's Kappa*, F_k) [59]. Analiza je pokazala visok nivo saglasnosti među anotatorima, sa F_k vrednošću od 0.86, što je i očekivano budući da su reči bile unapred obeležene za prisustvo emocionalnog afekta. Konačno obeležje za svaku reč određeno je na osnovu saglasnosti između dva nezavisna anotatora. Preostali lema_{Sr}-PoS parovi su agregirani radi uklanjanja potencijalnih ponovljenih unosa nastalih tokom opsežnih procesa prevođenja i validacije. Dobijeni leksikon, nazvan *EmoLex.SR-v1*, sadrži ukupno 4,135 lema_{Sr}-PoS parova kategorizovanih u emocionalne kategorije (pogledati tabelu 8.6).

Proširivanje

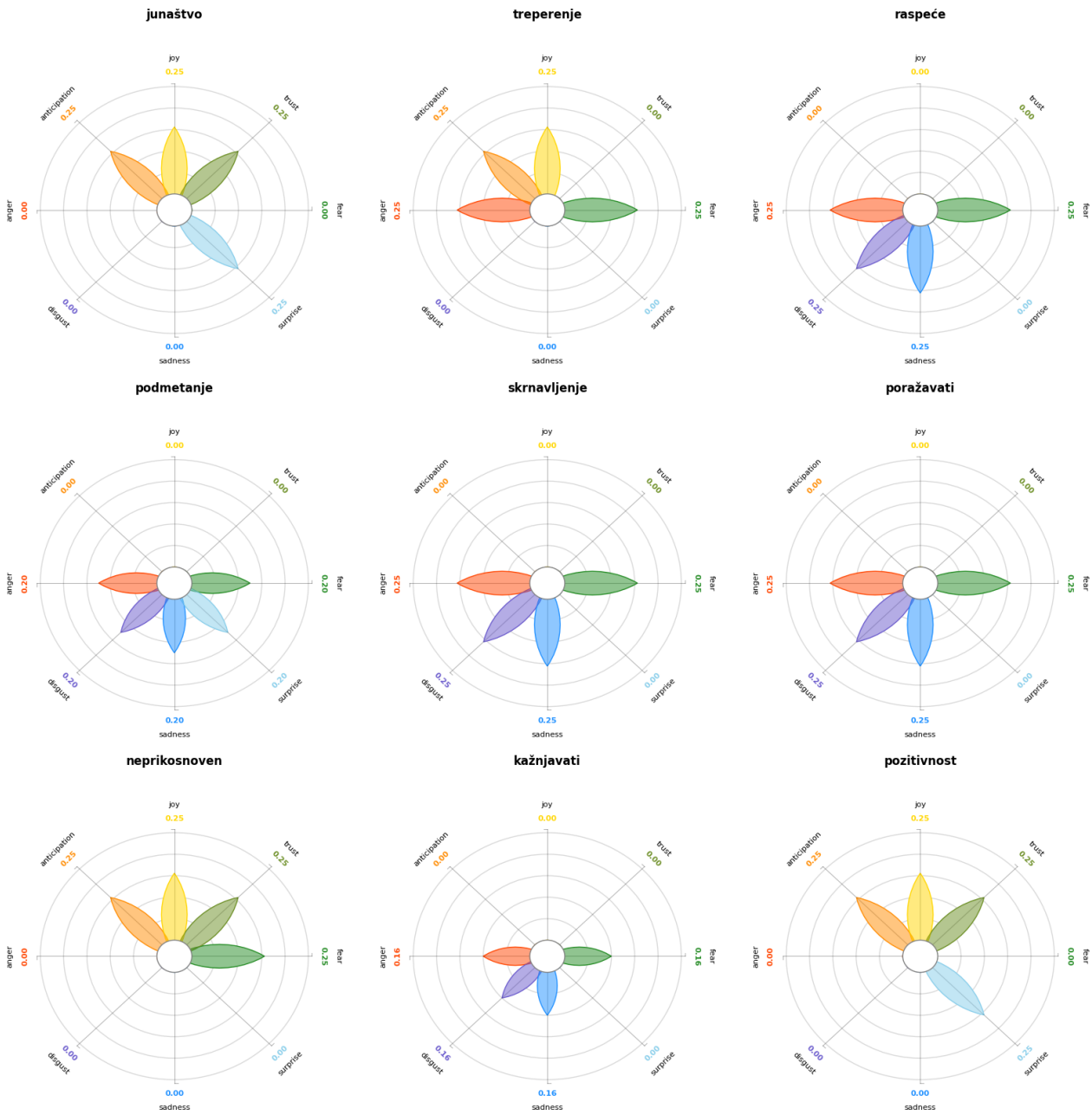
U fazi proširivanja *EmoLex.SR-v1* leksikona, korišćeni su SWN leksikon i Čet-GPT model za prikupljanje reči sa istim ili sličnim značenjem. Osnovni princip korišćen prilikom proširivanja jeste da se rečima sa sličnim značenjem pridružuju iste emocionalne kategorije. Inicijalno su napravljene dve lista sinonima (eng. *Synonyms List*, *Syn*) za svaku reč, zasnovane na Čet-GPT modelu i SWN leksikonu, koje su nazvane $Syn_{\text{Čet-GPT}}$ i Syn_{SWN}, u tom redosledu. Od Čet-GPT alata je zatraženo da generiše sinonime za datu lema_{Sr}-PoS kombinaciju, pri čemu se od modela očekivalo da napravljene reči budu iste vrste (pogledati prilog C). Ispravnost prikupljenih sinonima (Syn_{Total}, jednačina 8.8) je ručno proverena, pri čemu su svi nekorekntno (eng. *Incorrect*, *Inc*) obeleženi sinonimi dodati na listu netačnih sinonima (Syn_{Inc}) za svaku reč. Sinonim je klasifikovan kao netačan ukoliko nije iste vrste kao originalna reč, predstavlja flektivni oblik leme, frazu, ili ako sadrži imena i terminologiju specifičnu za određeni domen (npr. medicinsku).

$$Syn_{Total} = Syn_{\text{Čet-GPT}} + Syn_{SWN} \tag{8.8}$$

Konačna lista sinonima, u oznaci Syn_{Gold}, je sastavljena od Čet-GPT i SWN sinonima, sa izostavljanjem netačnih sinonima i ručno (eng. *Manual*, *Man*) dodatim sinonimima (Syn_{Man}) identifikovanim u toku procesa provere ispravnosti (jednačina 8.9).

$$Syn_{Gold} = Syn_{Total} - Syn_{Inc} + Syn_{Man} \tag{8.9}$$

Najzad, u leksikon su dodate reči kategorisane u emocionalne afektivne kategorije, koje su prikupljene kroz istraživanja srpskog jezika. Ove reči sačinjavaju 1,126 prideva i

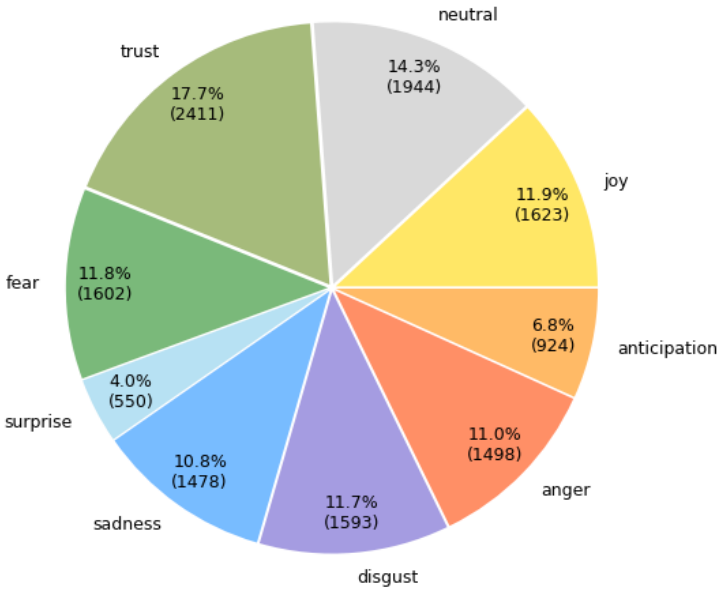


Slika 8.3: *Vizuelni prikaz kategorizovanih emocionalnih reči iz EmoLex.SR-v2 leksikona*

priloga [108], kao i 1,602 glagola i imenica [131]. Na kraju, vrednosti za sve emocionalne kategorije za sve lema_{Sr}-PoS parove su agregirane zbog ponovljenih kombinacija koje su se pojavile u procesu proširivanja.

Statistika leksikona

Završni leksikon, nazvan *EmoLex.SR-v2*, sadrži 9,782 jedinstvenih lema_{Sr}-PoS parova, sa ukupnom statistikom po svakoj emocionalnoj kategoriji predstavljenoj na slici 8.4. Tokom procesa izgradnje leksikona, ručna korekcija doprinela je proširenju leksikona od 4.7%. Ovaj proces je podrazumevao izbor ispravnih lema, sinonima i ispravljanje emocionalnih obeležja automatski prevedenih reči. Faza proširivanja je uticala na povećanje od 146% broja lema_{Sr}-PoS parova *EmoLex.SR-v1* leksikona, što je rezultovalo sa 9,782 lema_{Sr}-PoS kombinacija prisutnih u završnom leksikonu *EmoLex.SR-v2*. Vizuelni prikaz dodeljenih emocionalnih kategorija i vrednosti intenziteta na primerima iz *EmoLex.SR-v2* leksikona



Slika 8.4: *Raspodela emocionalnih kategorija u EmoLex.SR-v2 leksikonu*

prikazan je na slici 8.3 pomoću biblioteke *pyplutchik*[60].

Tabela 8.6: *Statistika broja emocionalnih reči u NRC.EN, EmoLex.SR-(v1, v2) leksikonima*[61]

Leksikon	Jezik	Konstrukcija	#Lema	#Emo-Lema
NRC.EmoLex	EN	Man	14,154	4,454 (31.5%)
NRC.EN	EN	Man + SWN	5,917	3,686 (62.3%)
NRC.EmoInt.tr	SR	GT	5,891	3,925 (66.6%)
NRC.EN.tr	SR	GT	3,778	2,315 (61.3%)
NRC.EN.val	SR	GT*	3,940	2,418 (61.4%)
EmoLex.SR-v1	SR	GT* + LC	4,134	3,353 (81.1%)
EmoLex.SR-v2	SR	GT* + LC + Syn	9,782	8,383 (85.7%)

Postepeni pristup izgradnji leksikona ilustruje evoluciju leksikona *EmoLex.SR*, prelaskom sa osnovnog automatskog prevođenja reči na sofisticiranu integraciju osnovnih oblika reči specifičnih za srpski jezik i kulturu. Kao što je prikazano na Slici 8.4, leksikon ima gotovo ravnomernu raspodelu između kategorija emocionalnog afekta, pri čemu većina reči pripada kategoriji *trust* (17.7%), dok su kategorije *surprise* (4.0%) i *anticipation* (6.8%) znatno manje zastupljene. Statistika prema vrsti reči pokazuje da leksikon sadrži ∼50% imenica, ∼23% prideva, ∼25% glagola i ∼2% priloga koji su obeleženi na prisustvo afekta.

Pažljiva konstrukcija leksikona, kroz faze prevođenja, prilagođavanja i proširivanja, proizvela je verzije *EmoLex.SR-(v1, v2)* leksikona. Iz Tabele 8.6 je primetno da prevod sa očuvanim semantičkim značenjem (*EmoLex.EN-val*) obuhvata 66.5% engleskog skupa reči (NRC.EN). Međutim, broj potpuno ispravnih srpskih lema dobijenih na takav način je relativno mali, sa samo 2,418 lema u NRC.EN.val i 3,925 lema u NRC.EmoInt.tr leksikonima.

[60] https://github.com/alfonsosemeraro/pyplutchik

[61] Oznaka "tr" u nazivu leksikona označava leksikon preveden na srpski jezik nekom od metoda automatskog prevođenja.

8.3. Obeležavanje korpusa konverzacionih podataka prema emocionalnom afektu i moralnoj vrednosti

Korpus konverzacionih podataka na srpskom jeziku, pod nazivom *Social.SR*, nastao je prikupljanjem tekstualnih poruka sa društvenih mreža Tviter i Redit sa ciljem njihove kategorizacije u odgovarajuće emocionalne ili moralne kategorije. Kategorizacija za emocionalni afekt je izvršena prema Plutčikovom modelu, koji daje optimalan balans između složenosti i preciznosti klasifikacije emocionalnih stanja u tekstovima [214]. Za kategorizaciju moralnih vrednosti korišćene su kategorije osnovnih moralnih vrednosti koje prepoznaje MFT. Poruke su razdvojene prema platformi kako bi se omogućila analiza specifičnih karakteristika komunikacije na konverzacijama sa svake platforme. Pored emocionalnog, na jednom delu emocionalnog korpusa je sprovedeno analogno obeležavanje moralnih vrednosti. Korpus obeležen na moralnost je prema svojoj konstrukciji nešto manje veličine, ali uvodi potpuno novu kategorizaciju moralnih vrednosti, čime se omogućava unakrsna analiza emocionalnih i moralnih aspekata komunikacije na srpskom jeziku. Postupak izgradnje obeleženih konverzacionih korpusa na ova dva aspekta jezika se sastojao iz sledećih neposrednih koraka izvršavanja:

- **Prikupljanje podataka** – prikupljanje poruka na srpskom jeziku sa društvenih mreža Tviter i Redit.
- **Predobeležavanje poruka** – obeležavanje poruka korišćenjem dostupnih jezičkih resursa za prepoznavanje aspekata u tekstovima na srpskom jeziku.
- **Odabir poruka** – odabir poruka sa emocionalnim afektivnim stanjima ili moralnim vrednostima prema unapred definisanim pravilima nad automatski obeleženim podacima.
- **Provera obeležja** – ručna provera automatski dodeljenih obeležja porukama na osnovu uspostavljenih šema za anotaciju.
- **Harmonizovanje korpusa** – računanje usaglašenosti obeležja, efikasnosti ručne provere i kreiranje konačnih obeležja.

Harmonizovani emocionalni korpus je nazvan **`Social-Emo.SR`**, dok njegovi podkorpusi, imaju odgovarajuće nazive **`Twitter-Emo.SR`** za poruke prikupljene sa Tviter i **`Reddit-Emo.SR`** za poruke prikupljene sa Redit platforme. Analogno, moralni korpus je nazvan **`Social-Mor.SR`**, odnosno **`Twitter-Mor.SR`** i **`Reddit-Mor.SR`**, prema platformama na kojima su poruke nastale.

8.3.1 Prikupljanje konverzacionih poruka

Konverzacione poruke su prikupljene sa platformi Tviter i Redit sa primarnim ciljem da se razume dinamika razgovora i interakcija unutar digitalne zajednice koja govori srpski jezik. Popularnost ovih platformi omogućila je pristup velikoj količini raznovrsnih tekstualnih poruka na srpskom jeziku, što ih čini pogodnim za rešavanje raznovrsnih istraživačkih zadataka.

Poruke su preuzete programski korišćenjem otvorenog API pristupa za obe platforme. Za pristup porukama je korišćen Tviter API[62] kojem se pristupilo putem Pajton biblioteke *Tweepy*[63]. Biblioteka *Tweepy* sa sveobuhvatnim skupom funkcija omogućava pretragu

[62] https://developer.twitter.com/en/products/twitter-api
[63] https://www.tweepy.org/

poruka kroz jednostavne parametarske upite. Uvođenjem identifikatora konverzacije kao parametra pretrage dostupnog u API.v2, omogućeno je preuzimanje kompletnih konverzacija, umesto pojedinačnih poruka. Međutim, srpski jezik nije konzistentno prepoznat u Tviter bazi podataka. Iz tog razloga je ručno odabrano 94 ličnih ili poslovnih Tviter naloga, za koje je potvrđeno da proizvode poruke na srpskom jeziku. Lični nalozi pripadaju poznatim ličnostima, sportistima, političarima ili pojedincima uticajnim u Tviter zajednici, dok su poslovni nalozi povezani sa organizacijama, institucijama ili kompanijama. Deljene poruke su uklonjene iz početne selekcije poruka kako bismo osigurali prikupljanje samo poruka koje je originalno kreirao autor i kako bismo izbegli moguću pojavu poruka napisanih na drugim jezicima. Poruke su prikupljene kroz nekoliko iteracija u periodu od aprila do jula 2020. godine. Ovim pristupom je prikupljeno ukupno 116,272 poruka, koje čine 6,931 inicijalnih objava i 109,341 komentar na objave.

Poruke sa Redit platforme su prikupljene korišćenjem Redit API[64] kojem se pristupilo putem Pajton biblioteke Redit API omotač - *PRAW*[65]. Na platformi Redit odabrano je osam grupa na kojima su aktivni korisnici koji govore i pišu na srpskom jeziku. Odabrane grupe pokrivaju teme kao što su ekologija, finansije, pravni saveti, programiranje, istorija Srbije ili studentska pitanja, a vremenski okvir je postavljen na period od dve godine, od početka 2020. do kraja 2022. Poruke su prikupljene u okviru konverzacija, što znači da su prikupljene inicijalne objave, kao inicijatori razgovora, kao i pridruženi komentari na objave. Ukupno je sa platforme Redit prikupljeno 9,145 objava i 131,431 komentara, koje u ukupnom zbiru čine 140,576 Redit poruka napisanih na srpskom jeziku.

8.3.2 Predobeležavanje i odabir poruka

Prilikom kreiranja emocionalnog korpusa od izuzetne je važnosti da tekstualni sadržaji iskazuju neki emocionalni afekt ili moralni stav. Kako bismo ispunili ovaj uslov i olakšali iscrpan proces ručnog obeležavanja, primenili smo NLP tehnike na prikupljenim skupovima podataka. Konkretno, proces odabira poruka u našem pristupu sledi dva uzastopna koraka:

- Automatsko predobeležavanje poruka korišćenjem naprednih ML modela i semantičkih leksikona za detektovanje emocionalnih i moralnih signala u sadržajima.
- Odabir poruka na osnovu uspostavljenih kriterijuma nad rezultatima iz prethodnog koraka.

Predobeležavanje u kategorije emocionalnog afekta

Podaci su automatski obeleženi korišćenjem modela *Twiter-XLM-RoBERTa-Es*[66] koji je fino obučen za emocionalnu klasifikaciju Tviter poruka na španskom jeziku [207] u diskretne emocionalne kategorije [*anger*, *fear*, *sadness*, *joy*, *surprise*, *disgust*, *others*]. Osnova za ovaj model je *Twiter-XLM-RoBERTa*[67] model, koji je obučen na 198 miliona Tviter poruka na različitim jezicima [18]. Izabrani emocionalni model je postigao najbolji F_1^{Mi} rezultat od $\sim$71% na zadatku prepoznavanja emocija u porukama na španskom jeziku [152] i ispunjava nekoliko važnih preduslova:

- Postigao je prihvatljivu ukupnu tačnost ($F_1^{Mi} > 70\%$);

[64] https://www.reddit.com/dev/api/
[65] https://praw.readthedocs.io/en/stable/
[66] https://huggingface.co/daveni/twitter-xlm-roberta-emotion-es
[67] https://huggingface.co/cardiffnlp/Twitter-xlm-roberta-base

- Osnovna arhitektura modela XLM-R prepoznaje srpski jezik [44];
- Koristi šemu za kategorizaciju emocija koja je približna Plutčikovom skupu osnovnih emocija, pri čemu su kategorije *anticipation* i *trust* identifikovane kao nedostajuće;
- Obučen je na korpusu konverzacionih poruka sa društvene mreže Tviter.

Koristeći izabrani model za predviđanje emocionalnih kategorija bez prethodnog obučavanja (eng. *zero-shot*), uspešno je rešeno pitanje „hladnog starta“ za početnu anotaciju velikog skupa podataka na srpskom jeziku za koji još uvek nisu izgrađeni odgovarajući resursi ove vrste. Pomoću ovog modela, korišćenjem tehnike transfera znanja, napravljene su dve grupe obeležja:

- XLM-S: izabrana je emocionalna kategorija sa najvećom verovatnoćom;
- XLM-M: izabrane su sve emocionalne kategorije sa vrednostima verovatnoća > 0.3.

Kroz predložene načine konstrukcije XLM obeležja, svim strogo emocionalnim obeležjima, odnosno svim obeležjima izuzev „*others*“, je omogućeno da konkurišu za konačno obeležje. Minimalni prihvatljiv prag verovatnoće od 0.3 u drugoj grupi obeležja, je utvrđen eksperimentalno tokom predviđanja, u cilju da obuhvati emocije koje bi mogle biti prikrivene zbog grešaka koje mogu nastati prilikom transfera znanja između jezika i mogućeg pristrasnog ponašanja modela ka kategoriji „*others*“, na šta su autori ukazali u svom radu [207].

Tabela 8.7: *Primer teksta na srpskom jeziku obeleženog na prisustvo emocionalnih kategorija korišćenjem XLM unakrsno-jezičkog modela i pristupa zasnovanog na emocionalnom leksikonu*

Primer	XLM-S	XLM-M	*EmoLex.SR*
Nasukani otpad na nikad većim obalama reke, peščani sprudovi gde ih dugo nije bilo. Ovo je slika Dunava tokom jednog od najtoplijih leta u skorijoj prošlosti	sadness	sadness, suprise	anger, disgust, sadness, surprise

Sve prikupljene poruke u Tviter i Redit podkorpusima su dodatno obeležene korišćenjem emocionalnog rečnika *EmoLex.SR*, koji je specijalno dizajniran da reflektuje lingvističke i kulturološke karakteristike srpskog jezika (pogledati odeljak 8.2.3). Primer tekstualne sekvence na srpskom jeziku koja je obeležena na ovaj način je predstavljen u tabeli 8.7. Nakon automatskog predobeležavanja poruka, izvršen je odabir poruka prema sledećim kriterijumima:

- Objave imaju prepoznat emocionalni afekt korišćenjem bar jednog od XLM-M ili *EmoLex.tr* pristupa.
- Poruke imaju najmanje tri ispravne lema_{Sr}, uključujući i emotikone, kao specijalne znakove za predstavljanje emocionalnog afekta.
- Očuvanje konverzacione strukture poruka, odnosno zadržavanje svih komentara za izabrane inicijalne objave.

Tehnički nazivi obeležja su anonimizovani (*label_1*, *label_2*, *label_3*) kako bi se osiguralo da korišćena metodologija automatskog predobeležavanja, ni na koji način ne utiče na odluku anotatora prilikom provere. Nakon konačnog odabira, emocionalni korpus **`Social-Emo.SR`** sadrži ukupno 34,598 poruka, pri čemu podkorpusi **`Twiter-Emo.SR`** i **`Reddit-Emo.SR`** sadrže redom 16,669 i 17,929 poruka (pogledati tabelu 8.10).

Predobeležavanje u kategorije moralnih vrednosti

Poruke iz korpusa **Social-Emo.SR** su odabirane prema unapred definisanim kriterijumima kako bi se napravio korpus **Social-Mor.SR**. Odabir poruka iz emocionalnog korpusa je sproveden primenom pretrage po ključnim rečima koje su pažljivo odabrane da identifikuju poruke koje bi svojim sadržajem potencijalno mogle da imaju iskazivanje moralnih stavova ili da pokrenu moralne reakcije u komentarima. Ključne reči korišćene za pretragu su uključile izraze poput *„poštenje“*, *„pravda“*, *„saosećanje“*, *„zlo“*, *„dobro“*, *„odgovornost“*, *„dužnost“*, *„istina“*, i *„pravednost“*. Takođe, dodate su ključne reči specifične za društveno-politička i sociološka dešavanja na srpskom govornom području, kao što su *„korupcija“*, *„politika“*, *„sloboda“*, *„protest“*, *„nepravda“*, *„zdravstvo“*, *„obrazovanje“*, *„narod“*, *„lider“*, i *„institucije“*. Pretpostavka je bila da bi ove reči mogle pokrenuti diskusije u kojima korisnici izražavaju moralne stavove ili vrednosti. Primena pretrage po ključnim rečima omogućila je identifikaciju poruka koje imaju potencijal da izazovu diskusiju moralne prirode, s posebnim fokusom na relevantne teme koje se odnose na društvene i političke aspekte života. Nakon odabira, poruke su obeležene prema moralnim vrednostima koristeći tehniku prompt inženjeringa u kombinaciji sa *Falcon-7B-Instruct* LLM modelom. Pažljivom konstrukcijom upita prema LLM modelu, bez prethodnog doobučavanja, obezbeđeno je dobijanje željenih rezultata u odgovarajućem formatu (pogledati prilog C). U konačnoj instanci, korpus **Social-Mor.SR** sadrži ukupno 14,698 poruka (4,513 objava i 10,185 komentara), od čega 6,425 poruka (1,804 objava i 4,621 komentar) pripada **Twitter-Mor.SR** i 8,273 poruka (2,709 objava i 5,564 komentara) pripada korpusu **Reddit-Mor.SR**.

Važno je napomenuti da je u procesu obeležavanja tekstualnih sekvenci na srpskom jeziku na moralne vrednosti uočeno nekoliko značajnih izazova. Jedan od ključnih izazova je podrška dostupnih LLM za srpski jezik, jer iako je primetno da mnoge distribucije LLM razumeju srpski jezik, zvanična potvrda proizvođača o tome nije dostavljena. Razlog može da leži u činjenici da za jezike koji nisu široko zastupljeni u podacima za obučavanje proizvođač ne može da garantuje visoku efikasnost i preciznost u generisanom sadržaju. Dodatno, subjektivnost moralne interpretacije može biti poseban izazov, jer moralne vrednosti variraju u različitim kulturnim i društvenim kontekstima, što može rezultovati varijacijama u tačnosti obeležja [79]. Takođe, potencijalna pristrasnost modela, izazvana ograničenjima u podacima za obuku modela, koji često nisu jasno naznačeni, može izazvati netačno ili nejednako obeležavanje za određene vrste moralnih kategorija. Planirano je da se sve potencijalne nepravilnosti isprave kroz sveobuhvatnu ručnu proveru i detaljnu analizu LLM rezultata, kako bi se obezbedila tačnost i pouzdanost obeleženog korpusa.

8.3.3 Provera automatski dodeljenih obeležja

Sve prethodno automatski obeležene poruke dodatno je proverila grupa od šest anotatora sa srpskim kao maternjim jezikom. Njihov zadatak bio je da koriguju netačna i dvosmislena obeležja nastala automatskom obradom. Svaku poruku su nezavisno pregledala po dva anotatora, koji su prethodno bili detaljno upoznati sa anotacionim šemama predviđenim za svaki od obrađivanih jezičkih aspekata. (pogledati prilog B).

Stepen saglasnosti anotatora izračunat je korišćenjem Kohenovog kapa koeficijenta (eng. *Cohen's Kappa coefficient*, k) za svaki par anotacija i kategoriju obeležavanja. Za emocionalni korpus, u tabeli 8.8 prikazane su prosečne vrednosti koeficijenta k merene na skupu poruka koje su proveravali parovi anotatora. Dobijeni rezultati saglasnosti u ovom istraživanju, sa vrednostima k u rasponu od 0.28 do 0.47 (umerena do dobra saglasnost) za emocionalne kategorije *anger*, *joy*, *sadness* i *trust*, u rangu su sa stepenom saglasnosti dobijenim u srodnim istraživanjima. U istraživanjima kao što su [139, 175], prijavljene vrednosti

k za emocionalne kategorije se kreću u opsegu od 0.10 do 0.59, pri čemu se najviši nivo saglasnosti postiže za eksplicitnije emocionalne kategorije kao što su *joy* i *anger*, dok su emocionalne kategorije poput *disgust* ili *anticipation* često ocenjene sa nižim stepenom slaganja među anotatorima.

Pored toga, uvedene su dodatne mere slaganja *consensus*, *distinct* i *majority* saglasnosti za merenje prosečnog odnosa parova anotacija kada se anotatori potpuno slažu, uopšte ne slažu ili delimično slažu. Vrednosti za mere *consensus*, *distinct* i *majority* saglasnosti u tabeli 8.8 pokazuju da anotatori postižu nešto bolju usaglašenost pri obeležavanju Tviter poruka koje su generalno kraće, što može olakšati prepoznavanje emocionalnih signala. Drugi razlog se može nalaziti u činjenici da su, zbog prirode komunikacije na toj platformi, emocije eksplicitnije izražene, što dodatno olakšava prepoznavanje emocionalnih stanja. S druge strane, zbog mogućeg preplitanja različitih emocionalnih ekspresija u relativno kratkim tekstualnim sekvencama, na Tviteru je, takođe, primetan nešto niži stepen slaganja između anotatora, što bi zahtevalo dodatnu proveru obeležja u obeleženim korpusima.

Tabela 8.8: *Prosečna vrednost koeficijenta k slaganja između parova anotatora izračunata po emocionalnoj i moralnoj kategoriji u korpusima* **`Social-Emo.SR`** *i* **`Social-Mor.SR`**

Mera	Kategorija	`Twitter-Emo.SR`	`Reddit-Emo.SR`	Kategorija	`Twitter-Mor.SR`	`Reddit-Mor.SR`
k	*anger*	**0.47**	**0.46**	*authority*	0.16	0.07
	anticipation	0.10	0.17	*betrayal*	**0.28**	0.22
	disgust	0.22	0.24	*care*	0.19	0.10
	fear	0.15	0.21	*cheating*	0.25	**0.27**
	joy	**0.46**	**0.36**	*degradation*	**0.28**	0.16
	neutral	0.15	0.16	*fairness*	0.11	0.17
	sadness	0.28	0.32	*harm*	**0.33**	**0.26**
	surprise	0.18	0.27	*loyalty*	0.23	0.13
	trust	0.39	0.42	*non-moral*	0.02	0.06
				purity	0.13	0.13
				subversion	0.08	0.12
Avg k		0.27	0.29		0.19	0.15
consensus		0.18	0.25		0.09	0.08
distinct		0.38	0.46		0.41	0.57
majority		0.44	0.29		0.50	0.35

Algoritam za harmonizaciju obeležja koja su dodelili anotatori uključuje dva koraka. Najpre, za svaku poruku se sabiraju obeležja oba anotatora (pogledati jednačinu 8.10).

$$A = \sum_{n,L} Annotator\_labels \tag{8.10}$$

gde n označava broj anotatora ($n = 2$) i L je skup kategorija, $|L| = 9$ u emocionalnom korpusu[68], odnosno $|L| = 11$ u korpusu moralnosti[69]. Nadalje, harmonizovano obeležje (eng. *gold label*) se određuje uzimanjem najčešćeg obeležja $mode_m(A_l)$ iz skupa obeležja anotatora (pogledati jednačinu 8.11). Ukoliko postoji više nivoa prihvatljive saglasnosti ($m > 1$), formula određuje koji nivo odabrati na osnovu njihove učestalosti pojavljivanja i osigurava da ne dođe do preklapanja sa kategorijama odabranim u prethodnim nivoima ($m \leq t \leq k \leq L$).

$$Gold\_label = \sum_{m} mode_m(A_l), mode_m(A_l | l \in L, l \neq mode_t(A), m \leq t \leq n \leq L) \tag{8.11}$$

gde je m dogovoreni nivo slaganja anotatora ($m = 1$). U našim postavkama ($k = 2, m = 1$), opšta formula koja je predstavljena jednačinom 8.11 ima značenje da se potpuna i delimična

[68] L = [*anger*, *anticipation*, *disgust*, *fear*, *joy*, *neutral*, *sadness*, *surprise*, *trust*]
[69] L = [*authority*, *betrayal*, *care*, *cheating*, *degradation*, *fairness*, *harm*, *loyalty*, *non-moral*, *purity*, *subversion*]

slaganja između anotatora transformišu u presek njihovih obeležja, a potpuna neslaganja u uniju obeležja, odnosno kategorija u obeležjima. Isečak harmonizovanog korpusa u JSON formatu prikazan je sledećim fragmentom koda:

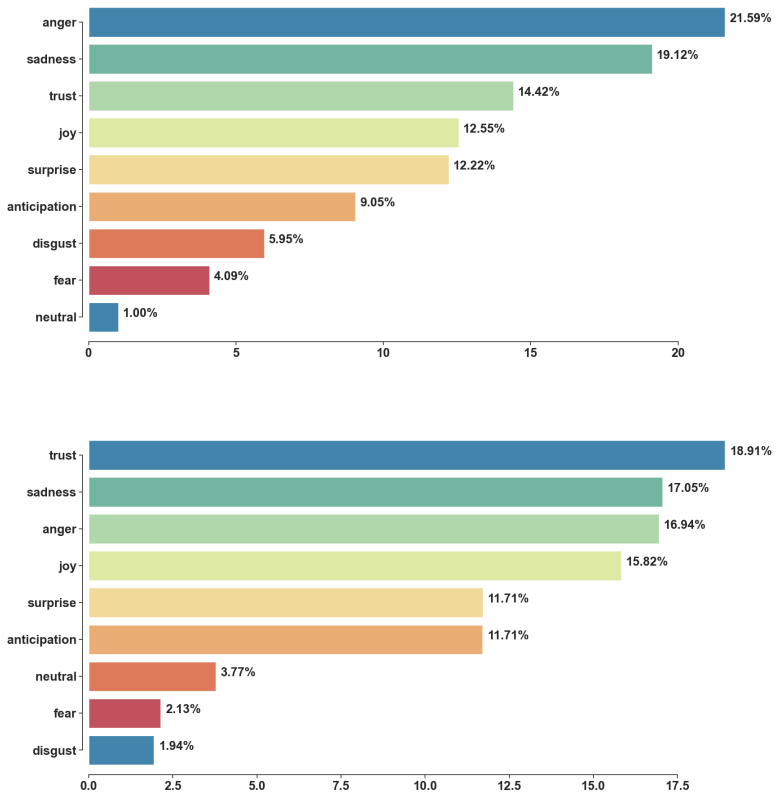
```
    "Messages": [
        {
            "id": 1555574367601299456,
            "annotations": [
                {
                    "annotator": "annotator-1",
                    "annotation": "care,loyalty"
                },
                {
                    "annotator": "annotator-2",
                    "annotation": "harm"
                },
                {
                    "annotator": "gold",
                    "annotation": "care,harm,loyalty"
                }
            ]
        },
        {
            "id": 1555883103251488769,
            "annotations": [
                {
                    "annotator": "annotator-1",
                    "annotation": "betrayal,cheating,harm"
                },
                {
                    "annotator": "annotator-2",
                    "annotation": "betrayal,cheating"
                },
                {
                    "annotator": "gold",
                    "annotation": "betrayal,cheating"
                }
            ]
        }, ...
]
```

Fragment koda 8.2: *Primeri poruka iz podkorpusa* **`Twitter-Mor.SR`** *sa dodeljenim obeležjima pojedinačnih anotatora i izračunatim harmonizovanim obeležjima*

8.4. Statističke karakteristike emocionalnog korpusa

U raspodeli emocionalnih kategorija u korpusu **`Social-Emo.SR`**, kao i njegovim podkorpusima **`Twitter-Emo.SR`** i **`Reddit-Emo.SR`**, primetna je najveća zastupljenost kategorija *anger* (21.59%) i *trust* (18.91%) u tom redosledu, dok kategorija *sadness* (19.12%/17.05%) uzima drugi najvažniji udeo u oba podkorpusa. Uočena je mala zastupljenost poruka sa dodeljenom *neutral* kategorijom (1%/3.77%), dok kategorije *fear* i *disgust* takođe imaju slabiju zastupljenost u podkorpusima sa 4.09%/2.13%, odnosno 5.95%/1.94%, sa nešto većim udelom u podkorpusu Tviter poruka. Preostale kategorije (*joy*, *surprise*, *anticipation*) imaju približno ujednačenu zastupljenost u podkorpusima koja varira od 9.05% do 15.02% (pogledati sliku 8.5).

Kroz primenu različitih statističkih mera nad obeleženim korpusima, pružili smo



Slika 8.5: *Vizuelni prikaz raspodele emocionalnih kategorija u podkorpusima* **Twitter-Emo.SR** *(gornja slika) i* **Reddit-Emo.SR** *(donja slika)*

sveobuhvatan pregled korpusa **Social-Emo.SR** i njegovih podkorpusa **Twitter-Emo.SR** i **Reddit-Emo.SR**. Podkorpusi pokazuju slične karakteristike u distribuciji višeznačnih obeležja (pogledati tabelu 8.9), pri čemu je u podkorpusu Tviter poruka primetna veća neuravnoteženost obeležja u odnosu na najzastupljenije obeležje (Mean Imbalance Ratio, MeanIR). Nasuprot tome, udeo poruka sa pridruženom jednom emocionalnom kategorijom u oznaci (*P_min*) od oko 60%, ukazuje na visok procenat ovakvih poruka u oba podkorpusa, što je direktna posledica uspostavljenih pravila anotacije i harmonizacije konačnih obeležja korišćenih u ovom istraživanju (pravilo 8.11).

Tabela 8.9: *Statistika višeznačnih emocionalnih obeležja u korpusu* **Social-Emo.SR**, *i njegovim podkorpusima* **Twitter-Emo.SR** *i* **Reddit-Emo.SR**

Korpus	Obeležja	*Dens*	*Card*	*avgIR*	*P_min*
Twitter-Emo.SR	201	0.19	1.68	4.45	0.59
Reddit-Emo.SR	153	0.17	1.50	3.47	0.62
Social-Emo.SR	204	0.18	1.59	3.03	0.61

Statističke mere o korpusima, prikazane u tabeli 8.10, uključuju sledeće podatke: broj poruka, broj konverzacija, prosečan broj tokena i ispravnih tokena po poruci, prosečnu dužinu poruke u karakterima, prosečan udeo emotikona u sadržaju poruke, kao i prosečan udeo poruka koje sadrže emotikone. Ove mere su izračunate kako za pojedinačne tipove poruka (objava ili komentar), tako i za sve poruke bez obzira na tip. Zapaženo je da poruke sa Redita imaju veću prosečnu dužinu, što se može pripisati ograničenju u maksimalnoj dužini poruke od 140 (kasnije 280) karaktera na Tviter platformi. Prosečan intenzitet

Tabela 8.10: *Statistika poruka u korpusu* **Social-Emo.SR** *obeleženog u emocionalne kategorije*

Korpus	Vrsta	Msg	Conv.	Tok.	Val	Chr	SRPOL	Emj	M-Emj
Twitter-Emo.SR	Objave	5884	4883	16.64	8.39	124.22	0.08	0.17	0.10
	Komentari	10785	1179	18.61	8.08	120.18	0.04	0.22	0.13
	Total	16669	4883	17.91	8.19	121.60	0.05	0.20	0.12
Reddit-Emo.SR	Objave	5322	5322	68.13	33.59	420.53	0.11	0.04	0.03
	Komentari	12607	521	46.34	22.00	266.37	0.07	0.05	0.04
	Total	17929	5322	52.81	25.44	312.13	0.08	0.05	0.03
Social-Emo.SR		34598	10205	36.00	17.13	220.33	0.07	0.12	0.08

sentimenta (SRPOL) poruka na Redit platformi pokazuje višu vrednost u poređenju sa sentimentom na Tviteru, kako za objave, tako i za komentare. S druge strane, Tviter poruke češće sadrže emotikone, kako u prosečnom broju njihovog pojavljivanja po poruci (Emj), tako i u učestalosti pojavljivanja emotikona u porukama (M-Emj). Procenat ispravnih tokena (Val) se smanjuje na oko 50% od ukupnog broja tokena (Tok) u poruci na obe platforme. Poslednje karakteristike ukazuju na neformalni stil i specifičnosti jezičkog izražavanja koji su karakteristični za svaku platformu.

Tabela 8.11: *Analiza zajedničkog pojavljivanja kategorija u konačnoj oznaci u podkorpusu* **Twitter-Emo.SR**

Category	*anger*	*anticip.*	*disgust*	*fear*	*joy*	*neutral*	*sadness*	*surprise*	*trust*
anger	6054 (13.4)	552	993	391	338	0	1714	1027	787
anticipation	1.22	2537 (5.62)	186	314	591	0	782	615	917
disgust	2.2	0.41	1669 (3.69)	157	161	0	585	365	247
fear	0.87	0.7	0.35	1148 (2.54)	123	0	541	293	340
joy	0.75	1.31	0.36	0.27	3519 (7.79)	0	514	693	923
neutral	0.0	0.0	0.0	0.0	0.0	280 (0.62)	0	0	0
sadness	3.79	1.73	1.3	1.2	1.14	0.0	5362 (11.87)	1149	1050
surprise	2.27	1.36	0.81	0.65	1.53	0.0	2.54	3425 (7.58)	785
trust	1.74	2.03	0.55	0.75	2.04	0.0	2.32	1.74	4043 (8.95)

Analiza zajedničkog pojavljivanja kategorija u konačnoj oznaci, prikazana u tabelama 8.11 i 8.12, pokazuje da se emocije poput *anger* i *sadness* često pojavljuju zajedno na obe platforme. Pozitivne emocije kao što su *joy*, *anticipation* i *trust* imaju tendenciju da se grupišu, dok se negativne emocije poput *anger* i *sadness* često povezuju sa *fear* i *disgust*. Veća zastupljenost kategorija *anticipation* i *trust* u porukama sa Redita i prisustvo *anger* u Tviter porukama ukazuje na specifične karakteristike svake platforme. Ovaj rezultat je delimično očekivan, jer Redit podstiče diskusije usmerene na pitanja i savete, što pogoduje izražavanju iščekivanja i poverenja, dok Tviter kroz svoj tematski kontekst i način komunikacije omogućava snažnije izražavanje negativnih emocija.

Dodatno, analizirano je prisustvo emocionalnih signala u svakom tipu poruke, kao i u različitim podkorpusima (pogledati tabelu 8.13). U Tviter objavama najdominantnija emocija je *trust*, dok se u komentarima najčešće pojavljuje kategorija *anger*. Suprotno tome, u Redit porukama *trust* i *anticipation* dominiraju u početnim objavama, dok su *sadness* i

Tabela 8.12: *Analiza zajedničkog pojavljivanja kategorija u konačnoj oznaci u podkorpusu* **Reddit-Emo.SR**

Category	*anger*	*anticip.*	*disgust*	*fear*	*joy*	*neutral*	*sadness*	*surprise*	*trust*
anger	4548 (11.89)	399	329	170	265	0	1189	640	529
anticipation	1.04	3143 (8.22)	43	95	545	0	563	390	1555
disgust	0.86	0.11	520 (1.36)	31	60	0	157	82	50
fear	0.44	0.25	0.08	573 (1.5)	54	0	200	103	160
joy	0.69	1.42	0.16	0.14	4247 (11.1)	0	470	474	986
neutral	0.0	0.0	0.0	0.0	0.0	1013 (2.65)	0	0	0
sadness	3.11	1.47	0.41	0.52	1.23	0.0	4576 (11.96)	677	729
surprise	1.67	1.02	0.21	0.27	1.24	0.0	1.77	3144 (8.22)	466
trust	1.38	4.07	0.13	0.42	2.58	0.0	1.91	1.22	5076 (13.27)

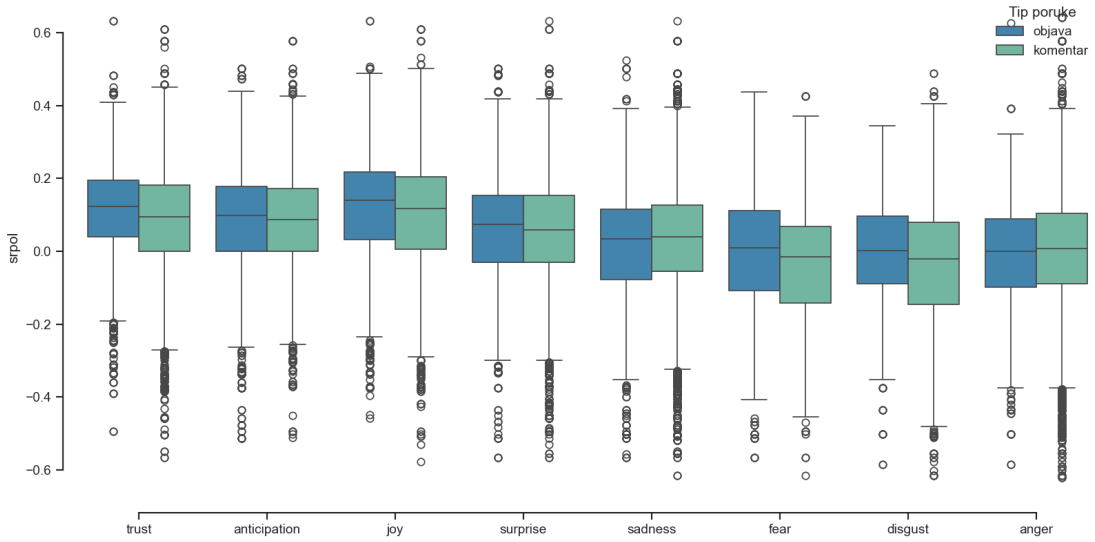
Tabela 8.13: *Relativna učestalost poruka prikazana po kategoriji emocija i tipu poruka*

Korpus	**Twitter-Emo.SR**		**Reddit-Emo.SR**	
Kategorija	Objave	Komentari	Objave	Komentari
anger	0.08	**0.29**	0.08	**0.20**
anticipation	0.14	0.06	**0.24**	0.07
disgust	0.05	0.07	0.01	0.02
fear	0.06	0.03	0.01	0.02
joy	0.15	0.11	0.08	**0.19**
neutral	0.02	0.00	0.05	0.03
sadness	**0.18**	**0.20**	0.07	**0.21**
surprise	0.13	0.12	0.07	0.13
trust	**0.19**	0.12	**0.38**	0.11

anger najzastupljenije emocije u komentarima. Emocije *anticipation* i *trust* su zastupljenije u početnim objavama na Reditu, što je delimično i očekivano, budući da korisnici na ovu platformu često dolaze radi saveta ili podrške od drugih korisnika. Inicijalne objave sa obeležjem *neutral* ipak podstiču emocionalno obojene odgovore. Premetno je, takođe, da početne objave na Reditu retko karakterišu emocije kao što su *fear* ili *disgust*, dok je njihovo prisustvo u početnim objavama nešto zastupljenije na Tviter platformi.

8.4.1 Emocije kao pokretači intenziteta sentimenta

Istražen je odnos između diskretnih emocionalnih kategorija, identifikovanih kroz konačna obeležja, i intenziteta sentimenta poruke, koji je izračunat pomoću SRPOL alata [184], u cilju utvrđivanja međusobnog uticaja prepoznatih emocionalnih stanja na sentiment iskaza. Na slici 8.6 je grafički prikazana zavisnost SRPOL intenziteta sentimenta i različitih emocionalnih kategorija i tipova poruka (objave, komentari). Pozitivne emocionalne kategorije kao što su *joy* i *anticipation* imaju tendenciju višeg pozitivnog sentimenta. Posebno, kategorija *anticipation* pokazuje dosledno pozitivan sentiment za objave i komentare, sa sličnim vrednostima medijane intenziteta sentimenta, što sugeriše da se ona uglavnom izražava u pozitivnom kontekstu. Poruke koje su kategorisane pomoću intuitivno negativnih emocionalnih kategorija, kao što su *anger*, *fear* i *disgust* pretežno potvrđuju negativan sentiment. Kod emocionalnih kategorija, kao što su *trust* i *surprise*, vrednosti sentimenta su



Slika 8.6: *Distribucija intenziteta sentimenta prikazana po emocionalnoj kategoriji i tipu poruke u podkorpusu* **`Twitter-Emo.SR`**

uglavnom pozitivne, sa uravnoteženim vrednostima između objava i komentara. Za sve emocionalne kategorije komentari pokazuju nešto niži sentiment u odnosu na inicijalne objave, sa širim opsegom vrednosti za određene kategorije, kao što su *joy*, *anger* i *sadness*. Promenljivost intenziteta sentimenta uočljiva u komentarima dodatno naglašava bogatstvo emocionalnog izražavanja u ovim vrstama poruka.

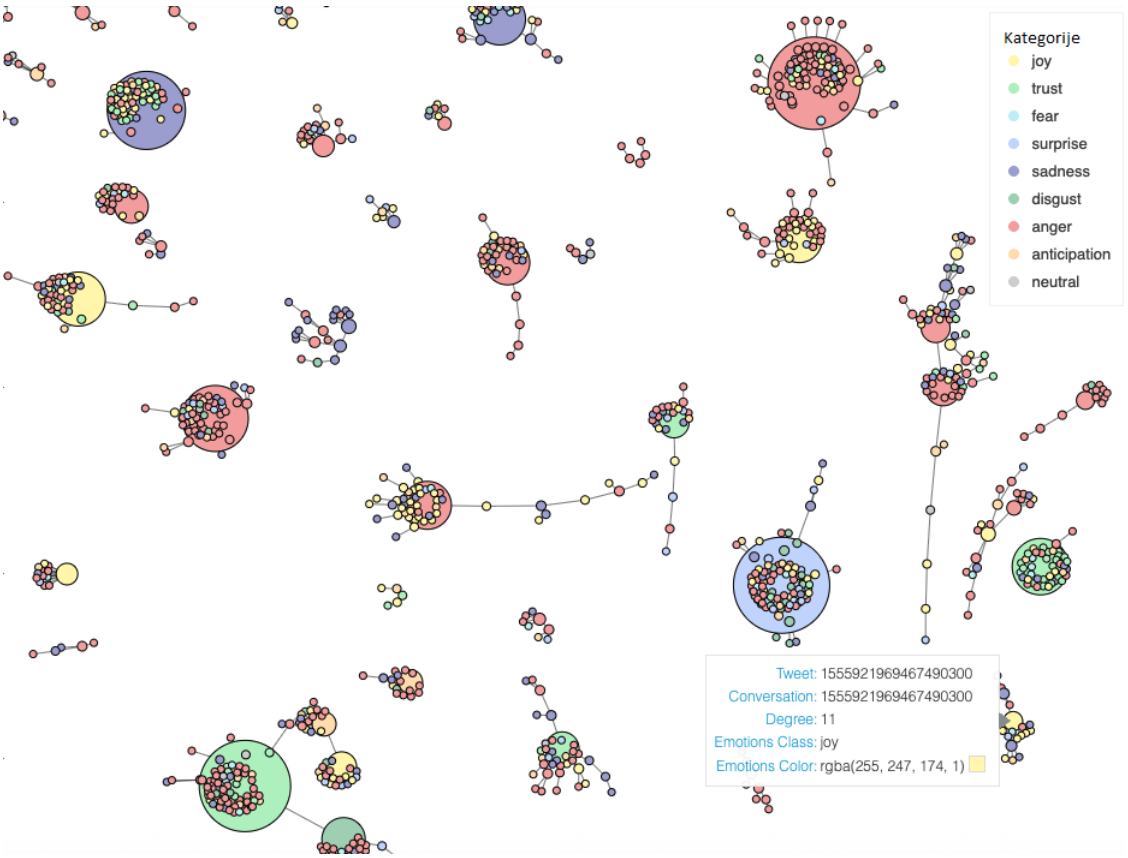
Pri tumačenju intenziteta polariteta iz višeznačnog skupa podataka, ključno je uzeti u obzir moguće greške koje mogu nastati iz nezavisnog posmatranja pojedinačnih kategorija u dodeljenim obeležjima. Takođe, greške u zaključivanju mogu nastati i usled nedoslednosti u obeležavanju i međusobnog preklapanja različitih emocija u izražavanjima složenih emocionalnih stanja. Uz oprez u tumačenju i svest o potencijalnim greškama, i dalje je moguće izvesti osnovne zaključke o emocionalnim trendovima unutar podataka.

8.4.2 Kvalifikacija odgovora na emocionalnu objavu

Reakcije na emocionalne objave na društvenim mrežama postale su centralna tačka za razumevanje dinamike digitalnih interakcija. Studije, kao što su one koje su sproveli autori u [48] i [196], istražuju obrasce i karakteristike korisničkih odgovora na emotivno obojeni sadržaj. Ova istraživanja pokazuju da emocionalne objave često generišu intenzivniju reakciju, izazivajući više komentara, sviđanja, deljenja i deljenja sa pridruženim komentarom (deljenja+) u poređenju sa neutralnim sadržajem. Rad [110] posebno naglašava ulogu reciprociteta i empatije u oblikovanju odgovora na emocionalne objave, osvetljavajući na taj način društvene mehanizme koji leže u osnovi digitalnih interakcija. Pored toga, istraživanje [115] ispituje uticaj iskazanog emocionalnog afekta u digitalnim zajednicama na opšti emocionalni ton u prostoru odgovora.

Tabela 8.14: *Statistika reakcija korisnika na emotivne objave po kategoriji emocija u korpusima* **Twitter-Emo.SR** *i* **Reddit-Emo.SR**

Korpus	**Twitter-Emo.SR**				**Reddit-Emo.SR**
Kategorija	sviđanja	komentara	deljenja	deljenja+	skor
anger	7.17	0.81	0.89	0.14	**6.36**
anticipation	7.82	0.88	0.47	0.17	3.37
disgust	**9.76**	**1.15**	0.68	0.22	**8.69**
fear	**9.68**	**1.09**	**1.90**	**0.36**	**8.61**
joy	**9.16**	0.71	0.50	0.12	**6.45**
neutral	5.82	0.69	0.48	0.13	4.56
sadness	6.58	0.85	0.74	0.16	**7.42**
surprise	7.87	0.85	0.87	**0.23**	**7.47**
trust	**9.40**	**1.04**	**1.46**	**0.27**	3.48
Avg	8.14	0.90	0.89	0.20	6.27



Slika 8.7: *Vizuelni prikaz konverzacionih nizova koji obuhvataju emocionalno obojene objave i pridružene odgovore izdvojene iz korpusa* **Twitter-Emo.SR**

U okviru korpusa **Social-Emo.SR**, emotivno obojene poruke su potvrdile veći intenzitet reakcija korisnika u odnosu na poruke kod kojih emocionalni signal nije prepoznat (kategorija *neutral*). Emocionalne objave, posebno one koje su obeležene kategorijama *disgust*, *fear* i *trust*, su privukle najveći stepen reakcija na Tviteru i Reditu, naročito u smislu sviđanja, komentarisanja i deljenja (pogledati tabelu 8.14). Sa druge strane, emocionalna kategorija *joy* je prepoznata sa visokim brojem sviđanja, ali sa manje interakcija u obliku komentarisanja ili deljenja. Nasuprot tome, objave koje izražavaju emocije kroz kategorije *neutral* ili *anticipation* su označene sa znatno manje reakcija, što ukazuje na to da su korisnici

skloniji interakciji sa objavama koje izazivaju snažne i intenzivne emocije. Kategorije *sadness* i *surprise* su postigle umerene nivoe reakcija korisnika, što sugeriše da, iako ove emocije korisnici prepoznaju, one ipak ne izazivaju tako snažnu reakciju kao, na primer, emocionalne kategorije *disgust* ili *fear*. Uopšteno, sadržaj sa intenzivnim emocijama ima tendenciju da podstiče veću reakciju korisnika, posebno u formi sviđanja na Tviteru i ukupnog skora na Reditu. Na slici 8.7 su vizuelno prikazane konverzacije sa emocionalno obojenim objavama i komentarima nastalim kao reakcije na takve inicijalne objave. Ovakva vrsta interaktivnog prikaza omogućava da se na brz i efikasan način uoče emocionalni pokretači i obrasci u reagovanju na emocionalno obojene sadržaje u konverzacijama na društvenim mrežama.

8.4.3 Izražavanje emocija kroz teme i kontekste

Emocije su u velikoj meri zavisne od subjektivnog doživljaja i konteksta u kojem se javljaju. Na društvenim mrežama, ispoljene emocije mogu varirati u zavisnosti od sadržaja, ličnih interesovanja i odnosa, ili aktuelnih događaja. Istraživanje tema i konteksta kao emocionalnih kontrasta predstavlja ključan aspekt analize emocija, otkrivajući složene načine na koje se emocije manifestuju u raznovrsnim temama i situacijama. Istraživanja u ovoj oblasti, poput rada [109], naglašavaju ulogu specifičnih tema u oblikovanju emocionalnog tona digitalnog diskursa. Pored toga, studija [148] istražuje uticaj kontekstualnih informacija na prepoznavanje emocija, ističući važnost uzimanja u obzir šireg narativa kako bi se tačno uočile emocionalne nijanse u datom tekstu.

Tabela 8.15: *Karakterističnije kategorije emocija na svakom od unapred definisanih konteksta u podkorpusima* **`Twitter-Emo.SR`** *i* **`Reddit-Emo.SR`**

		Objave		Komentari	
	Kontekst	Kategorija	SRPOL	Kategorija	SRPOL
`Twitter-Emo.SR`	muzika	*joy*	0.08	*joy*	0.08
	novosti	*sadness*	0.04	*anger*	0.02
	promocije	*joy,trust*	0.10	*anger*	0.04
	relaksacija	*joy*	0.09	*sadness*	0.04
	politika	*trust*	0.07	*anger*	0.04
	sarkazam	*joy,sadness*	0.12	*anger*	0.07
	komercijala	*joy,trust*	0.17	*anger,joy*	0.13
	obrazovanje	*joy*	0.14	*trust*	0.23
	institucije	*trust*	0.08	*anger*	0.06
	sport	*anticipation,joy*	0.12	*joy,sadness*	0.06
	Avg	*joy*	0.10	*anger*	0.07
`Reddit-Emo.SR`	ekologija	*anticipation*	0.12	*joy,trust*	0.13
	finansije	*trust*	0.12	*joy*	0.08
	pravni saveti	*anticipation,trust*	0.08	*anger*	0.05
	programiranje	*anticipation,trust*	0.13	*joy*	0.09
	istorija	*neutral*	0.11	*sadness*	0.09
	studentska pitanja	*joy,trust*	0.11	*joy*	0.10
	Avg	*anticipation,trust*	0.12	*joy*	0.09

Pristup analizi na osnovu različitih tematskih konteksta u okviru korpusa **`Social-Emo.SR`**, može pomoći da se razumeju načini iskazivanja emocija koji su povezani sa određenim temama, kao što je očekivano pokazivanje radosti u objavama vezanim za zabavu ili ljutnje u političkim diskusijama. Tabela 8.15 prikazuje različite unapred definisane kontekste u okviru korpusa **`Social-Emo.SR`**, kao što su vesti, muzika, politika, komercijalne usluge ili obrazovanje, u kojima su karakteristične emocionalne kategorije predstavljene

prema tipu poruke (objave i komentari) u podkorpusima **Twitter-Emo.SR** i **Reddit-Emo.SR**. Uočeno je da je, u kontekstima muzike i relaksacije, *joy* dominantna emocija koja predstavlja generalno pozitivan sentiment. Nasuprot tome, politika i vesti izazivaju više *anger* i *trust*, naročito u komentarima na Tviter platformi. Na Redit platformi, konteksti poput ekologije i finansija su dominantni u emocijama *anticipation* i *trust* u objavama, što potvrđuje očekivani ton usmeren ka budućnosti i neophodnim savetima.

Vrednosti intenziteta sentimenta izračunatog pomoću SRPOL alata u tabeli pružaju numerički prikaz jačine pozitivnih ili negativnih emocija u okviru svakog definisanog konteksta. U podkorpusu **Twitter-Emo.SR**, komercijalni i obrazovni konteksti beleže visok intenzitet sentimenta (0.17 i 0.14 za objave) sa dominantnim emocionalnim kategorijama *joy* i *trust*. S druge strane, u kontekstima kao što su novosti i politika, primetan je nešto niži intenzitet sentimenta, sa vrednostima od 0.02 do 0.04 i karakterističnom negativnom emocijom *anger* u komentarima, odnosno odgovorima korisnika. U podkorpusu **Reddit-Emo.SR**, intenzitet sentimenta u kontekstima kao što su ekologija, finansije i programiranje ima najveće vrednosti od $\sim$0.12 sa karakterističnim emocionalnim kategorijama kao što su *anticipation* i *trust*. Međutim, ozbiljniji konteksti kao što je istorija pokazuju neutralan sentiment u objavama (SRPOL od 0.11), ali naginju ka *sadness* (SRPOL od 0.09) u komentarima, što ukazuje na promenu intenziteta sentimenta u zavisnosti od konkretne teme i korisničkih doživljaja istorijskih događaja.

8.5. Statističke karakteristike korpusa moralnosti

Statistička analiza korpusa moralnosti **Social-Mor.SR** i njegovih podkorpusa **Twitter-Mor.SR** i **Reddit-Mor.SR** uključuje osnovne karakteristike kao što su ukupan broj poruka i konverzacija u korpusima. Dodatno, ova analiza je uključila i leksičke karakteristike korpusa kao što su prosečan broj i broj ispravnih tokena po poruci, prosečnu dužinu poruke izraženu u karakterima, prosečan udeo emotikona u sadržaju poruke, kao i prosečnu zastupljenost poruka sa emotikonima. Ovi parametri su analizirani na nivou tipa poruke (objava ili komentar), kao i na nivou svih poruka. Korpus moralnosti prati slične karakteristike kao i emocionalni korpus čiji je sastavni deo u pogledu dužina poruka u podkorpusima, intenziteta sentimenta, stilu pisanja i prisustvu emotikona (pogledati tabelu 8.16 i odeljak 8.4).

Tabela 8.16: *Leksičke i semantičke karakteristike korpusa* **Social-Mor.SR** *obeleženog u moralne kategorije*

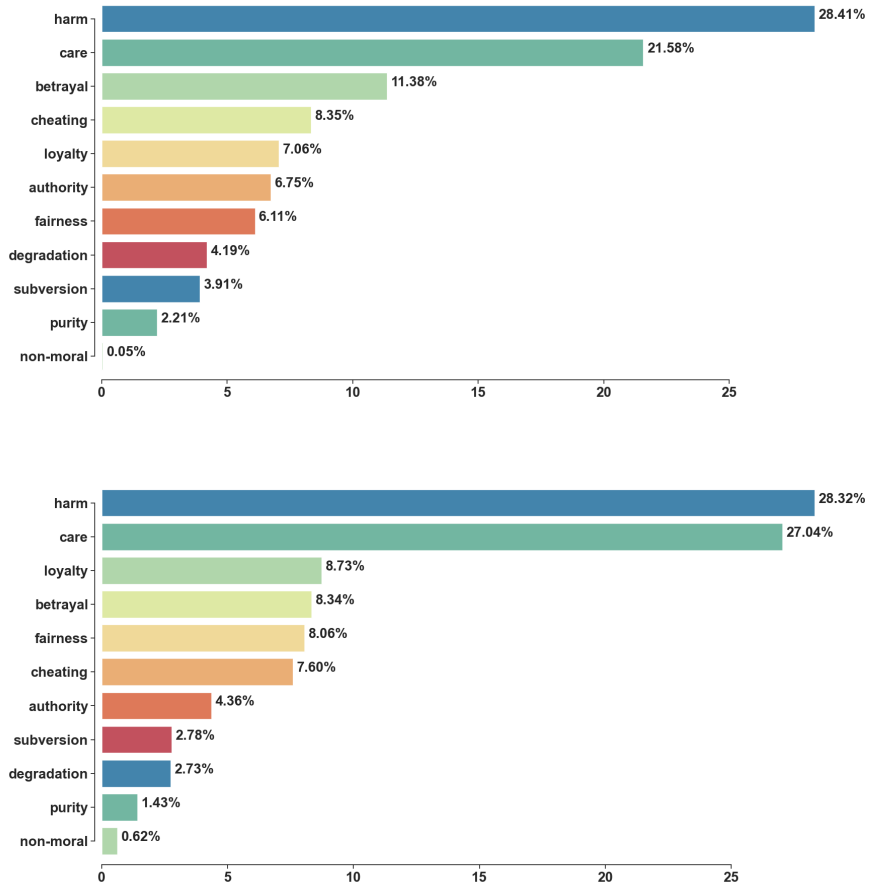
Korpus	Vrsta	Msg	Conv.	Tok.	Val	Chr	SRPOL	Emj	M-Emj
Twitter-Mor.SR	Objave	1472	1346	20.73	10.18	154.79	0.08	0.14	0.07
	Komentari	4621	338	19.48	8.49	125.77	0.03	0.20	0.11
	Total	6093	1346	19.78	8.89	132.78	0.04	0.19	0.10
Reddit-Mor.SR	Objave	1948	1948	107.75	53.79	664.74	0.10	0.03	0.02
	Komentari	5565	246	51.49	24.65	298.24	0.06	0.04	0.04
	Total	7513	1948	66.08	32.20	393.27	0.08	0.04	0.03
Social-Mor.SR		13606	3294	45.35	21.76	276.62	0.06	0.10	0.06

Podkorpusi korpusa **Social-Mor.SR** pokazuju slične karakteristike u distribuciji višeznačnih obeležja (pogledati tabelu 8.17), pri čemu je u podkorpusu Tviter poruka primetna veća neuravnoteženost u raspodeli obeležja (MeanIR) od 35.63. Udeo poruka sa pridruženom jednom moralnom kategorijom u oznaci (P_min) od oko 52%, ukazuje da

približno polovina poruka u celom korpusu je obeležena samo jednom moralnom kategorijom.

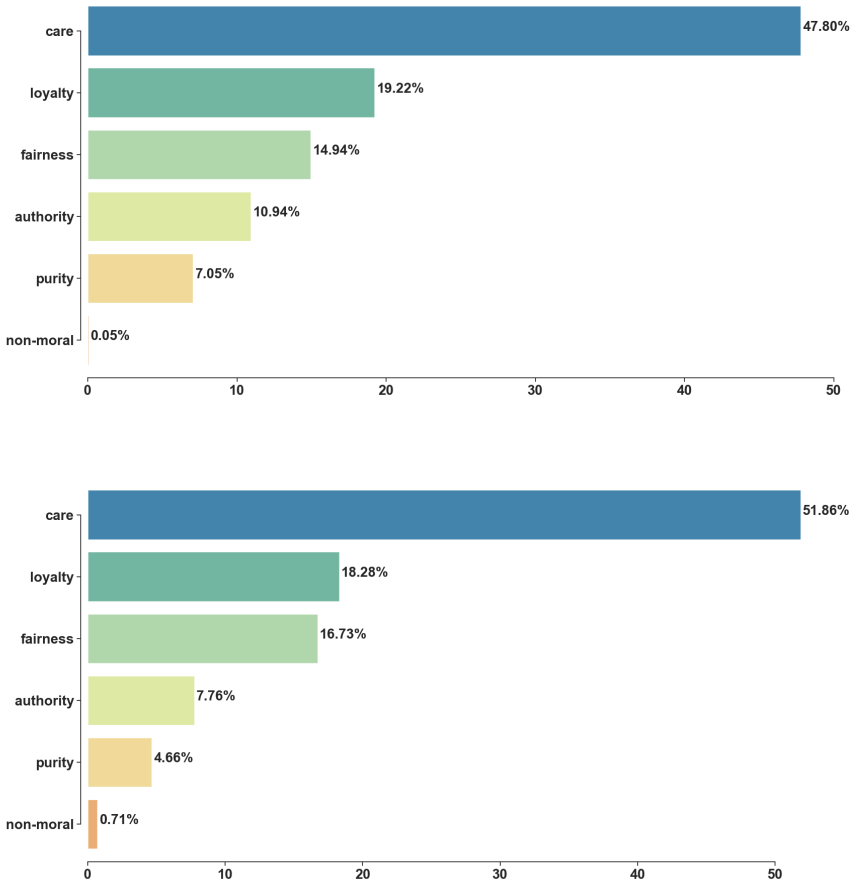
Tabela 8.17: *Statistika obeležja višeznačnog korpusa moralnosti* **Social-Mor.SR** *i njegovih podkorpusa* **Twitter-Mor.SR** *i* **Reddit-Mor.SR**

Korpus	Obeležja	*Dens*	*Card*	*avgIR*	*P_min*
Twitter-Mor.SR	328	0.16	1.80	35.63	0.55
Reddit-Mor.SR	303	0.16	1.79	8.82	0.50
Social-Mor.SR	383	0.16	1.80	10.23	0.52



Slika 8.8: *Vizuelni prikaz raspodele kategorija moralnog sentimenta u podkorpusima* **Twitter-Mor.SR** *(gornja slika) i* **Reddit-Mor.SR** *(donja slika)*

U raspodeli kategorija moralnog sentimenta u korpusu **Social-Mor.SR**, kao i njegovim podkorpusima **Twitter-Mor.SR** i **Reddit-Mor.SR**, primetna je najveća zastupljenost kategorije *harm* (28.41%/28.32%), dok kategorija *care* (21.58%/27.04%) uzima drugi najvažniji udeo u oba podkorpusa. Uočena je mala zastupljenost poruka sa dodeljenom *non-moral* kategorijom (0.05%/0.62%), dok kategorije *subversion*, *degradation* i *purity* takođe imaju slabiju zastupljenost u podkorpusima sa 3.91%/2.78%, odnosno 4.19%/2.73% i 2.21%/1.43%, sa nešto većim udelom u podkorpusu Tviter poruka. Preostale kategorije (*betrayal*, *cheating*, *loyalty*, *authority*, *fairness*) imaju približno ujednačenu zastupljenost sa drugačijim raspodelama u podkorpusima koja varira od 4.36% do 11.38% (pogledati sliku 8.8). Kategorije



Slika 8.9: *Vizuelni prikaz raspodele kategorija osnovnih moralnih vrednosti u podkorpusima* **`Twitter-Mor.SR`** *(gornja slika) i* **`Reddit-Mor.SR`** *(donja slika)*

Tabela 8.18: *Analiza zajedničkog pojavljivanja moralnih kategorija u konačnim obeležjima podkorpusa* **`Twitter-Mor.SR`**

Category	*authority*	*betrayal*	*care*	*cheating*	*degrad.*	*fairness*	*harm*	*loyalty*	*non-moral*	*purity*	*subver.*
authority	740 (3.95)	131	332	111	36	169	253	147	0	87	93
betrayal	0.7	1253 (6.69)	300	251	120	120	702	126	0	27	133
care	1.77	1.6	2380 (12.71)	248	98	329	769	403	0	132	162
cheating	0.59	1.34	1.32	873 (4.66)	87	103	454	78	0	23	102
degradation	0.19	0.64	0.52	0.46	501 (2.68)	38	287	39	0	12	52
fairness	0.9	0.64	1.76	0.55	0.2	632 (3.38)	250	142	0	43	79
harm	1.35	3.75	4.11	2.42	1.53	1.34	3102 (16.57)	219	0	65	301
loyalty	0.79	0.67	2.15	0.42	0.21	0.76	1.17	773 (4.13)	0	71	49
non-moral	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9 (0.05)	0	0
purity	0.46	0.14	0.71	0.12	0.06	0.23	0.35	0.38	0.0	253 (1.35)	12
subversion	0.5	0.71	0.87	0.54	0.28	0.42	1.61	0.26	0.0	0.06	421 (2.25)

Tabela 8.19: *Analiza zajedničkog pojavljivanja moralnih kategorija u konačnim obeležjima podkorpusa* **Reddit-Mor.SR**

Category	*authority*	*betrayal*	*care*	*cheating*	*degrad.*	*fairness*	*harm*	*loyalty*	*non-moral*	*purity*	*subver.*
authority	603 (2.7)	98	276	80	35	155	215	169	0	45	52
betrayal	0.44	1125 (5.04)	323	245	90	137	702	133	0	15	114
care	1.24	1.45	3662 (16.4)	317	100	528	1307	728	0	98	127
cheating	0.36	1.1	1.42	1013 (4.54)	62	138	516	97	0	7	56
degradation	0.16	0.4	0.45	0.28	347 (1.55)	37	238	40	0	12	40
fairness	0.69	0.61	2.36	0.62	0.17	1078 (4.83)	372	265	0	24	66
harm	0.96	3.14	5.85	2.31	1.07	1.67	3784 (16.95)	378	0	53	241
loyalty	0.76	0.6	3.26	0.43	0.18	1.19	1.69	1196 (5.36)	0	59	46
non-moral	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	107 (0.48)	0	0
purity	0.2	0.07	0.44	0.03	0.05	0.11	0.24	0.26	0.0	206 (0.92)	3
subversion	0.23	0.51	0.57	0.25	0.18	0.3	1.08	0.21	0.0	0.01	368 (1.65)

osnovnih moralnih vrednosti pokazuju približno ujednačenu raspodelu u oba podkorpusa, sa najdominantnijom kategorijom *care* sa udelom od 47.80%/51.86% u podkorpusima Tviter i Redit poruka, u tom redosledu (slika 8.9).

Analiza zajedničkog pojavljivanja kategorija u konačnoj oznaci, prikazana u tabelama 8.18 i 8.19, pokazuje da su moralne kategorije *care* i *harm* dominantne u oba podkorpusa, što ukazuje na njihovu centralnu ulogu u iskazivanju moralnih stavova na društvenim mrežama. Kategorije poput *betrayal*, *cheating*, *fairness* i *loyalty* takođe pokazuju značajnu prisutnost, ali sa nešto nižim vrednostima, dok je kategorija *non-moral* marginalno zastupljena, naročito na podkorpusu **Twitter-Mor.SR**. Razlike između platformi uključuju veću povezanost kategorija na Reditu, dok su na Tviteru kategorije slabije povezane (*purity* i *degradation*) što ukazuje da su diskusije na Reditu složenije, dok su Tviter poruke jednostavnije u pogledu iskazivanja moralnih stavova.

Tabela 8.20: *Relativna učestalost poruka prikazana po moralnoj kategoriji i tipu poruka*

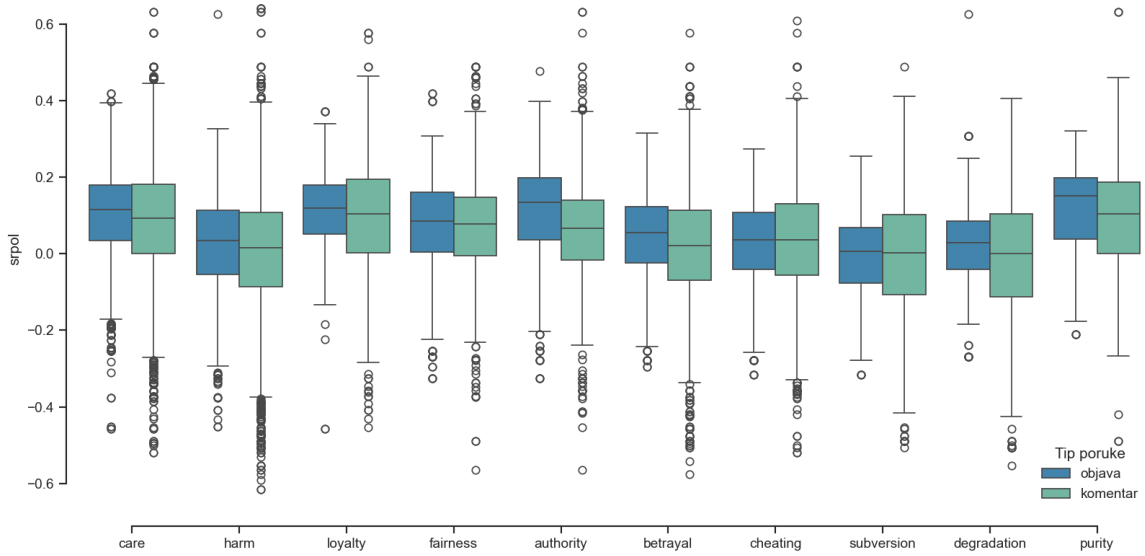
Korpus	**Twitter-Mor.SR**		**Reddit-Mor.SR**	
Kategorija	Objave	Komentari	Objave	Komentari
authority	0.05	0.12	0.04	0.05
betrayal	0.12	0.08	0.09	0.07
care	**0.21**	**0.26**	**0.26**	**0.32**
cheating	0.09	0.05	0.07	0.08
degradation	0.05	0.03	0.03	0.02
fairness	0.05	0.07	0.07	0.10
harm	**0.30**	**0.24**	**0.31**	**0.20**
loyalty	0.06	0.09	0.07	0.13
non-moral	0.00	0.00	0.01	0.01
purity	0.02	0.04	0.01	0.02
subversion	0.04	0.02	0.03	0.02

Relativna učestalost poruka po moralnim kategorijama i tipovima poruka u podkorpusima **Twitter-Mor.SR** i **Reddit-Mor.SR** pokazuje da kategorije *care* i *harm* dominiraju u oba podkorpusa, bez obzira na tip poruke. Na Tviteru, kategorija *care* obuhvata 21% objava i 26% komentara, dok na Reditu ova kategorija pokriva 26% objava i 32% komentara.

Slično, kategorija *harm* ima značajnu zastupljenost, sa najvišom vrednošću u objavama (30% na Tviteru i 31% na Reditu), dok je nešto niža u komentarima (24% na Tviteru i 20% na Reditu). Kategorije poput *authority*, *betrayal*, i *loyalty* imaju umerenu zastupljenost, pri čemu *loyalty* pokazuje veću prisutnost u Redit komentarima (13%) nego na Tviteru (9%). Negativno orijentisane kategorije, kao što su *degradation* i *subversion*, beleže nisku učestalost u oba korpusa, sa najvećim vrednostima od ∼5% poruka prisutnih u Tviter objavama.

8.5.1 Sentiment moralnih vrednosti

Prema MFT, osnovne moralne kategorije *authority*, *care*, *fairness*, *loyalty* i *purity* se na osnovu sentimenta mogu podeliti na dihotomne parove prema sentimentu, odnosno vrlinama i manama, koji nose: *authority* ↔ *subversion*, *care* ↔ *harm*, *fairness* ↔ *cheating*, *loyalty* ↔ *betrayal* i *purity* ↔ *degradation*. Svaka od dihotomnih kategorija može biti povezana s različitim vrednostima sentimenta i njegovim intenzitetom. Na primer, kategorija *care* obično izaziva pozitivan sentiment empatije i saosećanja kada je vrednost ispoštovana, i negativan sentiment kada je prekršena. Kategorija *purity* može proizvesti pozitivan osećaj poštovanja i uzvišenosti pri očuvanju svetih normi, ali intenzivno negodovanje prilikom njegovog kršenja. Uporedna analiza diskretnih moralnih kategorija, identifikovanih kroz konačna obeležja, i intenziteta sentimenta poruke, koji je izračunat pomoću SRPOL alata, potvrđuje ispravnost obeležavanja preko poklapanja očekivanih i dobijenih vrednosti sentimenta za svaku od moralnih kategorija (pogledati sliku 8.10).



Slika 8.10: *Distribucija intenziteta sentimenta prikazana po moralnoj kategoriji i tipu poruke u podkorpusu* **`Twitter-Mor.SR`**

U korpusu **`Twitter-Mor.SR`**, moralne kategorije koje predstavljaju izražavanje moralnih vrlina, kao što su *authority*, *care*, *fairness*, *loyalty* i *purity* imaju tendenciju višeg pozitivnog sentimenta, nasuprot kategorijama kojima se izražavaju mane ovih vrednosti (pogledati sliku 8.10). Posebno, kategorije *care* i *purity* pokazuju izrazito pozitivan sentiment za objave i komentare, sa sličnim srednjim vrednostima, što sugeriše da se one uglavnom koriste u pozitivnom kontekstu. Kod moralnih kategorija kao što su *harm* i *degradation*, vrednost intenziteta sentimenta je izraženo negativna, sa varijabilnim vrednostima između objava i komentara. Za sve moralne kategorije komentari imaju nižu vrednost medijane intenziteta sentimenta u odnosu na inicijalne objave, sa ekstremnijim vrednostima i širem opsegu

vrednosti za određene emocije kao što su *betrayal*, *care* i *harm*. Mogući uzrok može ležati u metodi ekspanzije višeznačnih obeležja, kojom se svaka pripadajuća kategorija distribuirano mapira na posebne vrste u tabelarnom prikazu, čime se potencijalno narušava značenje inicijalnih obeležja u daljoj analizi. Pri tumačenju intenziteta sentimenta iz višeznačnog skupa podataka, važno je uzeti u obzir moguće greške koje mogu nastati iz nezavisnog posmatranja pojedinačnih kategorija u dodeljenim obeležjima. Takođe, greške u zaključivanju mogu nastati i usled nedoslednosti u obeležavanju i međusobnog preklapanja različitih moralnih kategorija u izražavanjima složenih moralnih stavova.

8.5.2 Kvalifikacija odgovora na iskazan moralni stav

Moralnost i reakcije na društvenim mrežama su teme koje sve više privlače pažnju istraživača, jer ovi digitalni prostori otvaraju novi prostor za izražavanje moralnih stavova i interakcije među ljudima koje se brzo šire [216]. Istraživanja ukazuju na to da je verovatnoća deljenja sadržaja koji izražavaju kršenje moralnih vrednosti ili nepoštovanja normi veća u odnosu na verovatnoće deljenja drugih vrsta sadržaja [30]. Takođe, moralni stavovi izraženi kroz viralne kampanje često izazivaju masovne reakcije, bilo pozitivne (solidarnost) [193] ili negativne (panika i osuda) [163], u zavisnosti od njihovog usklađivanja s vrednostima određene zajednice [203]. Društvene mreže tako postaju arene za moralnu debatu u kojima se formira javno mnjenje i u kojima moralne norme evoluiraju kroz neprestanu i brzu digitalnu interakciju između korisnika.

Tabela 8.21: *Reakcije korisnika na objave po kategoriji moralne vrednosti u podkorpusima* **`Twitter-Mor.SR`** *i* **`Reddit-Mor.SR`**

Korpus			**`Twitter-Mor.SR`**		**`Reddit-Mor.SR`**
Kategorija	sviđanja	komentara	deljenja	deljenja+	skor
authority	6.22	**1.09**	0.61	**0.25**	4.73
betrayal	8.23	**1.05**	0.57	0.13	4.50
care	5.56	0.76	0.43	0.10	3.94
cheating	7.23	0.89	0.49	0.14	**5.47**
degradation	5.40	0.59	0.37	0.06	**5.05**
harm	6.58	0.90	0.46	0.14	4.86
fairness	6.63	0.88	0.54	0.14	3.78
loyalty	**10.77**	**1.39**	**2.49**	**0.43**	3.89
non-moral	1.22	0.22	0.00	0.00	3.12
purity	**10.86**	**1.07**	0.68	**0.41**	4.45
subversion	4.39	0.80	0.35	0.12	**5.40**
Avg	6.64	0.88	0.63	0.17	4.47

Reakcije korisnika na objave u korpusu **`Social-Mor.SR`** otkrivaju značajne razlike u angažmanu korisnika između moralnih kategorija i platformi (pogledati tabelu 8.21). Na Tviteru su kategorije *purity* i *loyalty* najpopularnije, sa najviše sviđanja (10.86 i 10.77), dok je *loyalty* takođe dominantna u broju komentara (1.39) i deljenja (2.49), što ukazuje na sklonost korisnika ka izražavanju podrške i identifikaciji sa temama lojalnosti i moralne čistoće. Na Reditu, najviše vrednosti reakcija korisnika, prikazane ukupnim skorom, zabeležene su za kategorije poput *cheating* (5.47), *degradation* (5.05), i *subversion* (5.40), što ukazuje na veći interes korisnika za kontroverzne i etički izazovne teme. Prosečna vrednost reakcija korisnika na Reditu (4.47) odražava generalno umeren nivo angažmana korisnika, dok niska vrednost za *non-moral* kategoriju (3.12) potvrđuje slabiji interes za teme koje nisu jasno moralno definisane. Ovi rezultati ukazuju da korisnici Tvitera naginju ka temama koje

evociraju podršku i empatiju, dok korisnici Redita gravitiraju ka temama koje izazivaju intelektualnu raspravu i polemiku. Razlike u angažmanu između platformi reflektuju razlike u njihovim kulturama komunikacije, gde je Tviter više fokusiran na emocionalni sadržaj, dok Redit podstiče dublje diskusije o etičkim pitanjima.

8.5.3 Karakteristične moralne vrednosti kroz teme i kontekste

U tabeli 8.22 su prikazane kategorije moralnih vrednosti karakteristične u različitim kontekstima identifikovanim u podkorpusima **`Twitter-Mor.SR`** i **`Reddit-Mor.SR`**. Na Tviteru, kategorija *care* je dominantna u gotovo svim temama, posebno u kontekstima kao što su muzika, promocije, relaksacija, i sarkazam, dok se u sportskim i obrazovnim temama često povezuje sa *loyalty*, čime se naglašava značaj brige i lojalnosti u ovoj tematskoj celini. Kategorija *harm* je prisutna u političkim i institucionalnim diskusijama, što ukazuje na teme koje evociraju štetu ili pretnju. Kategorija *cheating* dolazi do izražaja u komercijalnim kontekstima, što odražava percepciju moralnih odstupanja od očekivanja korisnika. Na Redit platformi, kategorija *care* takođe pokazuje svoju dominantnu karakteristiku, posebno u temama kao što su ekologija, finansije, i studentska pitanja, što ukazuje na fokus korisnika na brigu i podršku u ovim temama. U istorijskim diskusijama, karakteristične kategorije *authority* i *purity* naglašavaju tradicionalne vrednosti i autoritet. Kategorije *harm* i *cheating* su značajne u pravnim savetima, što reflektuje specifičan interes za pravdu i moralna odstupanja.

Tabela 8.22: *Karakteristične moralne kategorije na svakom od unapred definisanih konteksta u podkorpusima* **`Twitter-Mor.SR`** *i* **`Reddit-Mor.SR`**

		Objave		Komentari	
	Kontekst	Kategorija	SRPOL	Kategorija	SRPOL
`Twitter-Mor.SR`	muzika	*care,loyalty*	0.00	*care,harm*	0.09
	novosti	*harm*	0.03	*harm*	0.03
	promocije	*care*	0.08	*care,harm*	0.07
	relaksacija	*care*	0.06	*harm*	0.01
	politika	*harm*	0.05	*betrayal,harm*	0.04
	sarkazam	*care*	0.14	*harm*	0.08
	komercijala	*care*	0.20	*cheating,harm*	0.19
	obrazovanje	*loyalty,purity*	0.11	*care*	0.13
	institucije	*authority,care*	0.12	*harm*	0.11
	sport	*loyalty,care*	0.14	*care*	0.09
	Avg	*care*	0.09	*harm*	0.07
`Reddit-Mor.SR`	ekologija	*care*	0.11	*care,loyalty*	0.09
	finansije	*care*	0.12	*betrayal,care*	0.08
	pravni saveti	*authority,harm*	0.07	*cheating,harm*	0.05
	programiranje	*care*	0.13	*care*	0.10
	istorija	*authority,purity*	0.08	*authority,care*	0.07
	studentska pitanja	*care*	0.11	*care*	0.10
	Avg	*care*	0.10	*care*	0.08

8.6. Semantički leksikon moralnosti – MFD.SR

Savremena istraživanja sve češće potvrđuju efikasnost automatskih tehnika u izgradnji leksikona, koje mogu delimično ili potpuno zameniti tradicionalne, vremenski zahtevne i subjektivne ručne pristupe [82]. Napredne metode uključuju proširivanje postojećih leksikona primenom ML algoritama i Embd reprezentacija reči za pronalaženje relevantnih

izraza i njihovih semantičkih težina [12]. Leksikon moralnih reči za srpski jezik, pod dodeljenim nazivom *MFD.SR*, razvijen je u skladu sa savremenim istraživačkim pristupima, uz korišćenje dostupnih NLP tehnika i resursa specifičnih za srpski jezik. Rečnik je izgrađen kombinovanjem automatskih metoda za prepoznavanje značajnih moralno obojenih reči u obeleženim tekstualnim sekvencama na srpskom jeziku i ručne provere podataka. U procesu izgradnje, izdvajaju su sledeći glavni koraci:

- **Prikupljanje i provera obeleženih podataka** – za formiranje leksikona korišćeni su korpusi obeleženih tekstova prema 5 osnovnih moralnih vrednosti koje prepoznaje MFT, i to:
  - Prvi izvor tekstualnih podataka jeste korpus **`Social-Mor.SR`** sa ∼13.6k konverzacionih poruka na srpskom jeziku koji je obeležen prema prisustvu moralnih vrednosti u 11 kategorija moralnog sentimenta. Ove kategorije su grupisane u 5 (+ *non-moral*) kategorija osnovnih moralnih vrednosti prema kojima je vršena kategorizacija reči u leksikonu.
  - Pored podataka prikupljenih sa društvenih mreža, za konstrukciju leksikona korišćeni su i anotirani tekstovi koji su prethodno služili kao osnova za izgradnju *eMFD* leksikona na engleskom jeziku (pogledati odeljak 6.3.2). Konkretno, korišćen je opsežan korpus od 8,276 novinskih članaka, dok je za anotaciju izabrano 2,995, od kojih je 1,010 bilo ručno obeležio najmanje jednan anotator. Za potrebe konstrukcije leksikona, proizvedeno je 63,958 obeleženih tekstualnih fragmenata (eng. *highlights*), koji su korišćeni za identifikovanje izraza sa izraženim moralnim značenjem i njihovu kategorizaciju po moralnim osnovama *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion*, *sanctity/degradation* i *non-moral*.

    Za potrebe konstrukcije srpskog leksikona, obeleženi tekstovi su najpre automatski prevedeni na srpski jezik korišćenjem GT alata, a zatim pažljivo provereni i prilagođeni ručnim putem kako bi odgovarali specifičnostima moralnog izražavanja u okviru srpskog jezika i kulture. Ovi tekstovi su poslužili kao dodatni izvor termina i fraza povezanih sa moralnim vrednostima, čime je unapređena raznovrsnost i semantička pokrivenost leksikona u kontekstu srpskog jezika.

- **Predobrada i analiza teksta** – prikupljeni podaci su obrađeni korišćenjem standardnih NLP tehnika za procesiranje teksta, kao što su:
  - Tokenizacija, određivanje vrste reči i lematizacija – tekstualne sekvence se tokenizuju, određuje vrsta reči za svaki token i pronalazi njen osnovni oblik korišćenjem ML modela izgrađenih za srpski jezik [191]. U ovom koraku, svaki token se transformiše u uređenu trojku (token, lema_{Sr}, PoS). Na primer, reči kao što su „*vozio*“ ili „*vozićе*“ postaju uređene trojke („*vozio*“, „*voziti*“, VERB) i („*vozićе*“, „*voziti*“, VERB), čime se one standardizuju u isti oblik prilikom upotrebe lema_{Sr} u narednim koracima. Za dalju obradu izdvojeni su tokeni koji pripadaju vrstama reči: imenica (NOUN), glagol (VERB), pridev (ADJ) i prilog (ADV), budući da tokeni sa ovim PoS obeležjima predstavljaju najrelevantnije kandidate za izražavanje semantičkih aspekata moralnosti u jeziku.
  - Uklanjanje imenovanih entiteta i specijalnih karaktera – eliminacija reči koje nemaju značajnu semantičku vrednost za formiranje leksikona kao što su lična imena, imena organizacija ili lokacija, kao i znakovi interpunkcije, nevidljivi znakovi ili tokeni koji sadrže brojeve[70].

[70] Funkcionalne reči (najčešće zamenice, predlozi, veznici, rečce i druge) koje, takođe, ne nose značenje

- **Primena c-Tf-Idf algoritma** – nakon što su svi podaci predprocesirani, primenjena je modifikovana Tf-Idf tehnika za izračunavanje značajnosti svake reči, odnosno leme, u okviru pojedinačnih klasifikacionih kategorija [71]. Ova metoda je primenjena za efikasno pronalaženje reči koje su karakteristične za određene moralne kategorije iz prethodno obeleženih tekstualnih korpusa. Za primenu metode, neophodno je prethodno grupisanje svih tekstualnih sekvenci iste kategorizacije u jedan dokument nad kojim se primenjuje standardna Tf-Idf tehnika vektorizacije teksta. c-Tf-Idf težine u vektorskoj reprezentaciji kategorije, odnosno klase, izračunavaju se korišćenjem formule predstavljene jednačinom 8.12:

$$\begin{aligned} cf_{t,C} &= |\{c \in C : t \in c\}| \\ Icf_{t,C} &= \log\left(\frac{N}{cf_{t,C}}\right) \\ c\text{-}Tf\text{-}Idf_{t,c,C} &= Tf_{t,c} \times Icf_{t,C} \end{aligned} \tag{8.12}$$

  gde je:

  - C skup klasa u korpusu,
  - $N = |C|$ broj klasa u korpusu,
  - $cf_{t,C}$ je frekvencija klasa u korpusu koje sadrže termin t,
  - $Icf_{t,C}$ inverzna frekvencija klasa u korpusu koje sadrže termin t
  - $Tf_{t,c}$ je frekvencija termina t u klasi c.

- **Izračunavanje sentimenta reči** – svakoj reči je pridružena vrednost intenziteta sentimenta korišćenjem *SentiWords.SR* leksikona [184]. Za lema_{Sr}-PoS parove u leksikonu *MFD.SR* za koje nisu postojali odgovarajući ekvivalenti u *SentiWords.SR* leksikonu, vrednost sentimenta je određena na osnovu sličnosti između reči u neposrednom okruženju u alfabetno uređenom *MFD.SR* leksikonu. Sličnost između dve reči izračunata je korišćenjem kosinusne sličnosti (eng. *cosine similarity*, *cossim*) nad njihovim vektorskim reprezentacijama (A, B), koje odražavaju učestalost pojavljivanja karaktera u svakoj od reči (pogledati jednačinu 8.13). Ovakav pristup omogućio je dodeljivanje vrednosti intenziteta sentimenta rečima na osnovu zajedničkih karakterističnih obrazaca i sličnosti u frekvenciji pojavljivanja karaktera.

$$cossim = \frac{A \cdot B}{\|A\|\|B\|} \tag{8.13}$$

  gde je:

$$A \cdot B = \sum_{i=1}^{n} A_i B_i \tag{8.14}$$

  skalarni proizvod vektora, a:

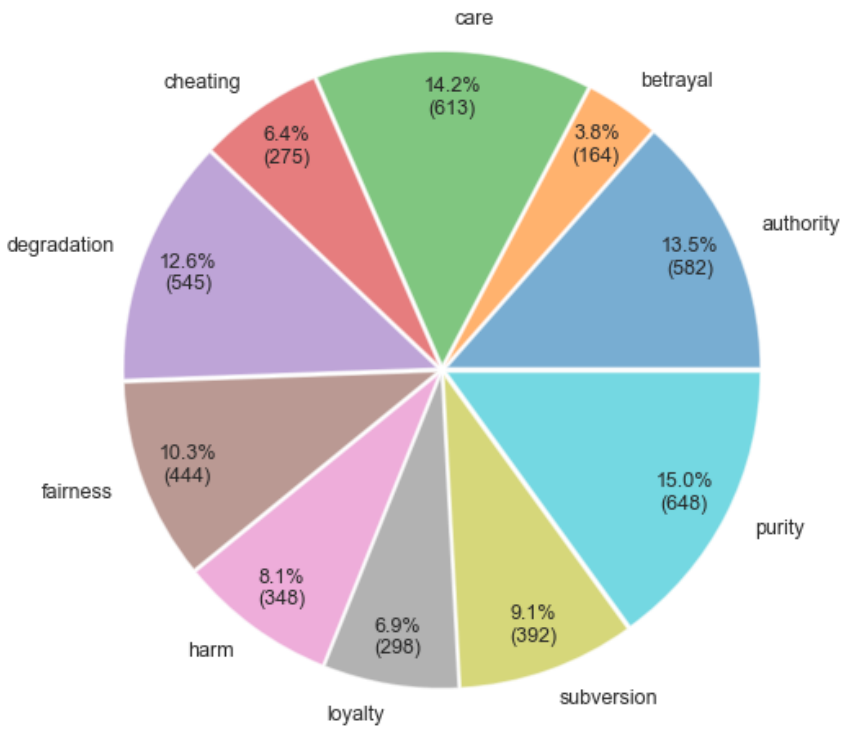
$$\|A\| = \sqrt{\sum_{i=1}^{n} A_i^2}, \|B\| = \sqrt{\sum_{i=1}^{n} B_i^2} \tag{8.15}$$

  norme vektora. Kosinusna udaljenost (eng. *cosine distance*, *cosdis*) se na osnovu prethodnih formula računa kao:

$$cosdis = 1 - cossim \tag{8.16}$$

će biti uklonjene primenom algoritma za pronalaženje značajnih reči u narednom koraku, umesto ručnog pravljenja liste reči koja bi bila odgovarajuća za ovaj zadatak.

Kao prihvatljiva minimalna granica sličnosti, postavljena je vrednost $cossim \geq 0.85$ ili analogno $cosdis \leq 0.15$. Nakon primene algoritma, ukupno 1,648 reči je dobilo pridruženu vrednost intenziteta sentimenta, koja ranije nije postojala, a koja je određena na osnovu semantičke povezanosti između dve reči. Primeri ovakvih reči su derivacioni oblici kao na primer („*agresivan*“, „*agresivno*“), („*civilizovan*“, „*civilizovati*“), („*ekstreman*“, „*ekstremista*“, „*ekstremistički*“, „*ekstremizam*“) ili („*žrtva*“, „*žrtvovati*“). Na ovaj način moguće je proširiti *SentiWords.SR* leksikon identifikovanim derivacionim i varijacionim oblicima koji ovim leksikonom još uvek nisu obuhvaćeni.



Slika 8.11: *Raspodela kategorija moralnog sentimenta u MFD.SR leksikonu*

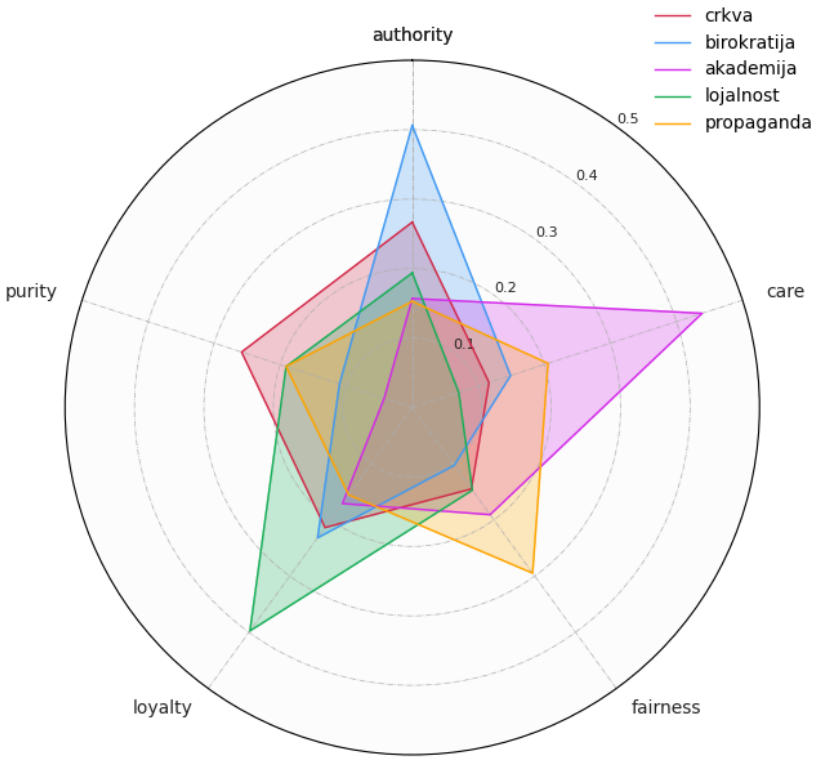
Konstruisanje leksikona moralnih reči

Na osnovu izračunatih c-Tf-Idf težina, konstruisan je leksikon moralnih reči, pod nazivom *MFD.SR*, koji uključuje karakteristične reči za svaku od moralnih vrednosti. Leksikon sadrži reči sa njihovim pripadajućim težinama, koje su zasnovane na učestalosti njihovog pojavljivanja u tekstovima obeleženim na moralne vrednosti. Za svaku reč uspostavljen je kriterijum selekcije koji podrazumeva da c-Tf-Idf vrednost u najmanje jednoj moralnoj kategoriji premašuje prag od 5×10^{-5}, uz dodatni uslov da reč sadrži više od dva karaktera. Vrednosti težina za preostale reči su zatim normalizovane tako da njihova suma bude jednaka jedinici. Svi ponovljeni $lema_{Sr}$-PoS parovi, kao i parovi koji nisu imali dodeljenu vrednost sentimenta su uklonjeni iz leksikona. Za svaki $lema_{Sr}$-PoS par zatim je izvršena kategorizacija prema maksimalnoj vrednosti težina u moralnim kategorijama i vrednostima sentimenta ($\leq 0 \rightarrow$ *vice*, $\geq 0 \rightarrow$ *virtue*) u obeležja moralnih vrednosti i moralnog sentimenta, kao što je prikazano u sledećem primeru:

```
    "MFD.SR": [
        {
            "lemma": diskriminacija,
            "PoS": NOUN,
            "weights":
                {
                    "authority": 0.18,
                    "care": 0.16,
                    "fairness": 0.25,
                    "loyalty": 0.14,
                    "purity": 0.27
                },
            "sentiment": -0.64,
            "moral_value": purity,
            "moral_sentiment": purity.vice,
            "moral_category": degradation
        },
        ...
]
```

Fragment koda 8.3: *Primer jednog $lema_{Sr}$-PoS para iz MFD.SR leksikona sa uključenim c-Tf-Idf težinama, intenzitetom sentimenta i moralnim obeležjima*

Izračunate težine lema iz leksikona se mogu vizuelno predstaviti na radijalnom grafikonu sa osama koje predstavljaju moralne kategorije. Na slici 8.12 je predstavljeno pet primera lema iz leksikona *„crkva“*, *„birokratija“*, *„akademija“*, *„lojalnost“* i *„propaganda“* na osnovu vrednosti koje su $lema_{Sr}$ dobile za svaku od pet kategorija osnovnih moralnih vrednosti nakon primene c-Tf-Idf algoritma korišćenog za konstrukciju leksikona.



Slika 8.12: *Vizuelni prikaz c-Tf-Idf težina na primeru pet lema iz MFD.SR leksikona*

Tabela 8.23: *Statistika MFD.SR leksikona prema vrsti reči i kategorijama moralnih vrednosti i moralnog sentimenta*

Kategorija	Pridevi	Prilozi	Imenice	Glagoli	Total
authority	138	23	285	136	582
subversion	84	13	188	107	392
	222	36	473	243	**974**
care	119	55	315	124	613
harm	70	35	156	87	348
	189	90	471	211	**961**
fairness	95	16	235	98	444
cheating	60	13	133	69	275
	155	29	368	167	**719**
loyalty	69	24	134	71	298
betrayal	35	10	73	46	164
	104	34	207	117	**462**
purity	153	31	352	112	648
degradation	134	15	281	115	545
	287	46	633	227	**1193**
Total	957	235	2152	965	**4309**

MFD.SR leksikon sadrži 4,309 jedinstvenih lema$_{Sr}$-PoS parova, sa statistikom pojavljivanja po svakoj moralnoj kategoriji i vrsti reči predstavljenom u tabeli 8.23. Prema vrsti reči, leksikon sadrži 2,152 (∼50%) imenica, 957 (∼22.2%) prideva, 965 (∼22.4%) glagola i 235 (∼5.5%) priloga koji su obeleženi na moralnu vrednost i moralni sentiment. Kao što je prikazano na slici 8.11, leksikon ima približno ravnomernu raspodelu između kategorija moralnog sentimenta, pri čemu kategorija *purity* (15.0%) ima najveću zastupljenost, dok su kategorije *betrayal* (3.8%) i *cheating* (6.4%) najmanje zastupljene u leksikonu. U pogledu kategorizacije na moralnu vrednost, kategorije imaju približno ravnomernu raspodelu, pri čemu kategorija *purity* sadrži najveći broj (1,193, 27.7%), a kategorija *loyalty* najmanji broj (462, 10.7%) pridruženih lema$_{Sr}$-PoS parova.

8.7. Izgradnja modela za prepoznavanje emocionalnog afekta i moralne vrednosti

Ispravnost obeležja u korpusima **`Social-Emo.SR`** i **`Social-Mor.SR`**, je izvršena formiranjem ML modela sa zadatkom višeznačne klasifikacije emocija i moralne vrednosti, tim redosledom. Formirani ML modeli su doobučeni modeli transformer arhitekture, korišćenjem naprednih tehnika kao što su transferno učenje na enkoder modelima ili prompt inženjering na dekoder modelima. Korišćeni enkoder modeli predstavljaju različite BERT modele, koji po svojoj arhitekturi (XLM) ili principima obuke i korišćenim podacima podržavaju srpski jezik. Za generativni model izabran je LLaMA-3.2-3B-Instruct model iz grupe LLaMA-3 modela, koji su optimizovani za upotrebu u višejezičnim dijalozima i prilagođeni za rad sa instrukcijama [68]. Prilagođavanje podrazumeva obuku sloja za klasifikaciju na vrhu unapred obučenih BERT ili LLaMA modela, koristeći obeležene podatke u odgovarajućem formatu kako bi se težine modela prilagodile zadatku prepoznavanja emocija (pogledati tabelu 8.13). Modeli koji su korišćeni kao osnova za doobučavanje, sa glavnim karakteristikama prikazanim u tabeli 8.24, su:

- **XLM-R$_{Base}$**[71]/**XLM-R$_{Large}$**[72] – pretreniran na višejezičnom korpusu i podržava preko 100 jezika, uključujući srpski. Njegova arhitektura omogućava razumevanje teksta na više jezika, što je korisno za rad sa tekstovima na srpskom jeziku koji sadrže unakrsne jezičke elemente (na primer anglicizme). Upotreba ovog modela za doobučavanje na specifičnom korpusu kao što je Tviter može poboljšati razumevanje slenga, skraćenica i specifičnih izraza.
- **Twitter-XLM-R$_{Base}$**[73]/**Twitter-XLM-R$_{Large}$**[74] – prilagođen specifično za sadržaje sa Tviter društvene mreže što ga čini izuzetno pogodnim za obradu kratkih, neformalnih tekstova koji sarže emotikone i skraćenice. Model podžava više jezika, uključujući i srpski, a doobučavanje na srpskom korpusu može dodatno prilagoditi model za razumevanje lokalnih izraza i karakterističnog jezika na srpskom.
- **Jerteh-81**[75]**/Jerteh-355**[76] – specifično trenirani na velikom korpusu srpskog jezika i prilagođeni za obradu srpskih tekstova na oba pisma (ćirilica i latinica). Njihova specijalizacija za srpski jezik pruža odličnu osnovu za doobučavanje na specifičnim domenima, što može poboljšati performanse modela u prepoznavanju konteksta i lokalnog žargona [181].
- **BERTić**[77] – treniran na bosanskom, hrvatskom, crnogorskom i srpskom jeziku, što mu daje širinu u razumevanju regionalnih razlika i varijacija jezika. Pogodan je za doobučavanje korišćenjem obeleženih korpusa na srpskom jeziku jer već poseduje osnovne jezičke strukture i rečnik, pri čemu se može efikasno prilagoditi za specifične izraze i kontekst platforme [124].
- **LLaMA-3.2-3B-Instruct**[78] – je LLaMA model sa smanjenim brojem parametara (3B) u odnosu na prethodne modele ove arhitekture, koji je optimizovan za rad sa tekstualnim podacima i doobučen za prepoznavanje instrukcija. Korišćenje ove arhitekture modela moglo bi doprineti poboljšanju u prepoznavanju emocionalnog afekta i moralnih vrednosti u tekstovima, u poređenju sa prethodno korišćenim transformer arhitekturama (BERT).

Tabela 8.24: *Osnovni modeli izabrani za doobučavanje na zadatku klasifikacije tekstova na srpskom jeziku u emocionalne kategorije*

Arh.	Model	#J	Podržani jezici	T	V	#Param
BERT	XLM-R$_{Base}$	100+	višejezični	SP	32k	270M
	XLM-R$_{Large}$	100+	višejezični	SP	32k	550M
	Twitter-XLM-R$_{Base}$	100+	višejezični	SP	32k	270M
	Twitter-XLM-R$_{Large}$	100+	višejezični	SP	32k	550M
	Jerteh-81	4	balkanski	BPE	50k	81M
	Jerteh-355	4	balkanski	BPE	50k	355M
	BERTić	4	balkanski	BPE	32k	110M
LLM	LLaMA-3.2-3B-Instruct	8+	višejezični	SP	128k	3B

[71] https://huggingface.co/FacebookAI/xlm-roberta-base
[72] https://huggingface.co/FacebookAI/xlm-roberta-large
[73] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base
[74] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-large-2022
[75] https://huggingface.co/jerteh/Jerteh-81
[76] https://huggingface.co/jerteh/Jerteh-355
[77] https://huggingface.co/classla/bcms-bertic
[78] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Za potrebe obučavanja modela, podaci su najpre podeljeni u trening, validacioni i test skup korišćenjem metode stratifikacije višeznačnih obeležja radi očuvanja ravnomerne raspodele emocionalnih kategorija u svakom od kreiranih skupova za obučavanje i proveru modela [177]. Iz svakog korpusa su odabrane poruke koje imaju najviše tri ($\leq$3) kategorije i čija dužina ne prevazilazi 512 tokena. Obeležja su zatim enkodirana u vektorski prostor čija je dimenzija jednaka broju kategorija klasifikacije. Ovakva struktura zavisnog atributa je neophodna za primenu metode višeznačne klasifikacije korišćenjem DL algoritma za doobučavanje modela. Iz razloga nedovoljne zastupljenosti neutralnih kategorija (*neutral*, odnosno *non-moral*) u korpusima za obučavanje (pogledati odeljak 8.4 i 8.5), poruke sa ovim obeležjima su eliminisane iz skupova podataka, nakon izvršene provere u jednom od eksperimenata. U pogledu izabranih tehnika za procesiranje i ulaznih podataka za obučavanje, na zadatku klasifikacije emocionalnog afekta, izvršeni su eksperimenti na **`Twitter-Emo.SR`** podkorpusu, sa sledećim značenjem:

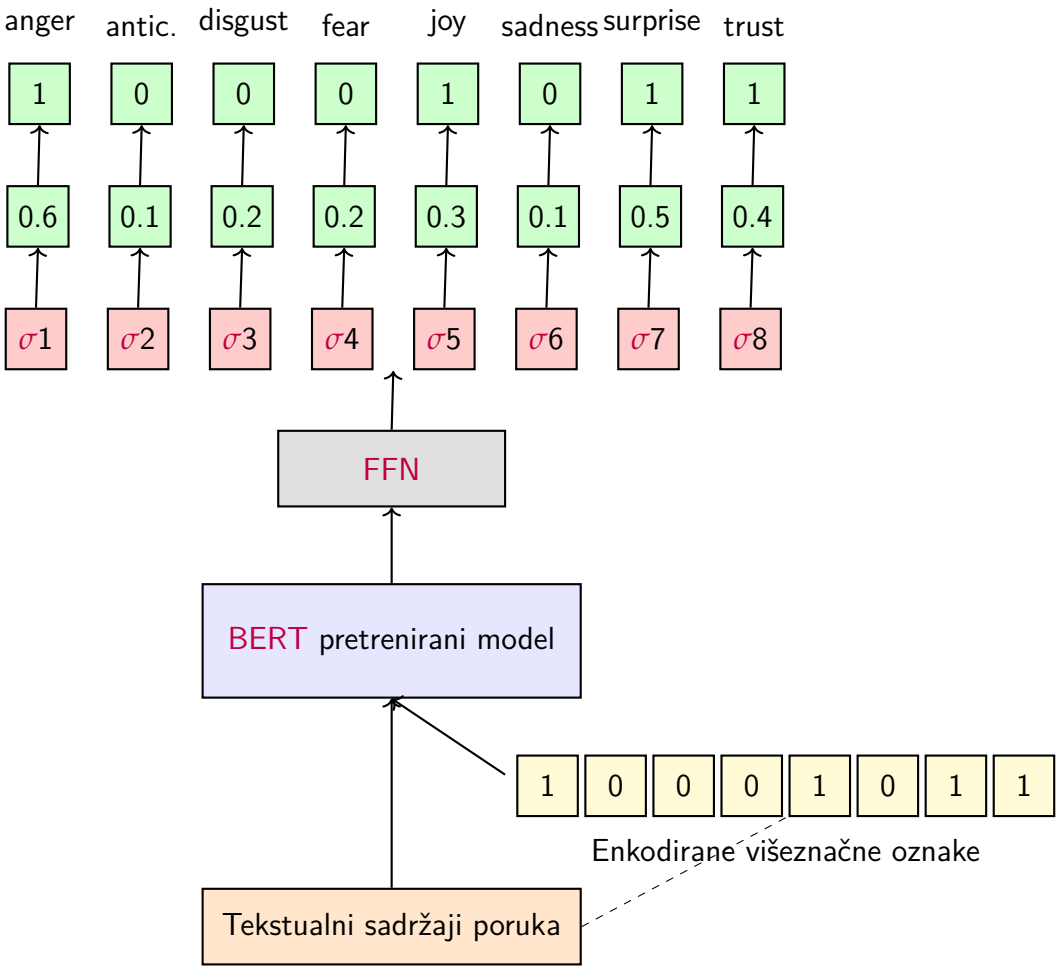
- *Orig* – bez procesiranja ulazne tekstualne sekvence;
- *Masked* – maskiranjem heš oznaka (*#hash*), korisničkih imena (*@user*) i linkova (*http*), koji se mogu pronaći u porukama preuzetim sa društvenih mreža;
- *Masked+Emo* – korišćenje samo striktno emocionalnih obeležja (*Gold_label* $\neq$ „*neutral*“) uz maskiranje iz prethodnog koraka.

Najbolji rezultati iz prethodnog eksperimenta su poslužili za izbor tehnike za obradu ulaznih sekvenci u glavnom eksperimentu za klasifikaciju emocionalnog afekta nad celim korpusom **`Social-Emo.SR`** i njegovim podkorpusima. Na zadatku klasifikacije moralnih vrednosti, odnosno provere obeležavanja korpusa **`Social-Mor.SR`**, za ulazne sekvence se koristi isti princip obrade ulaznih podataka, dok klasifikacija obuhvata:

- Klasifikaciju osnovnih moralnih vrednosti – prema MFT koja identifikuje 5 osnovnih moralnih vrednosti, odnosno kategorija klasifikacije: *authority*, *care*, *fairness*, *loyalty*, *purity*.
- Klasifikaciju moralnih vrednosti prema sentimentu – svaka primarna moralna vrednost je prema moralnom sentimentu (vrlina, mana) podeljena na dihotomne parove: *authority* $\rightarrow$ *authority/subversion*, *care* $\rightarrow$ *care/harm*, *fairness* $\rightarrow$ *fairness/cheating*, *loyalty* $\rightarrow$ *loyalty/betrayal*, *purity* $\rightarrow$ *purity/degradation*, koje čine ukupno 10 klasifikacionih kategorija u ovom eksperimentu.

Za svaki od izabranih modela, eksperimentalnim putem su izabrani optimalni hiperparametri za obučavanje modela koji uključuju stopu učenja od 2×10^{-5}, veličinu grupe za obuku od 32 poruke, težinsku dezintegraciju od 0.01 i maksimalnih 15 prolazaka kroz skup za obučavanje. Algoritam takođe koristi rano zaustavljanje kako bi zaustavio obuku kada nema poboljšanja u greškama nad podacima za treniranje i evaluaciju, dok se sveobuhvatna evaluacija modela izvršava na testnom skupu koji nije korišćen u toku obuke. Ovakav pristup obučavanju, sa optimalnim izborom hiperparametara, omogućava efikasno i precizno obučavanje modela koji je sposoban da generalizuje naučene zavisnosti iz podataka na novim konverzacionim porukama na srpskom jeziku.

Eksperimenti sa izgradnjom modela za klasifikaciju emocionalnog afekta su izvršeni nad celim korpusom **`Social-Emo.SR`**, kao i nad pojedinačnim podkorpusima **`Twitter-Emo.SR`** i **`Reddit-Emo.SR`**. Analogno, eksperimenti sa izgradnjom modela za klasifikaciju moralne vrednosti su izvršeni nad celim korpusom **`Social-Mor.SR`**, kao i nad pojedinačnim podkorpusima **`Twitter-Mor.SR`** i **`Reddit-Mor.SR`**. Ovakav pristup je omogućio analizu performansi modela nad različitim delovima izgrađenih korpusa i njihovih speci-



Slika 8.13: *Arhitektura algoritma doobučavanja BERT modela nad korpusom konverzacionih poruka sa dodeljenim višeznačnim obeležjima prikazana na primeru klasifikacije emocionalnog afekta*

fičnosti, kao što su različiti stilovi, tematske varijacije ili distribucija obeležja. Doobučavanje je iskorišćeno sa ciljem da se maksimalno iskoriste mogućnosti različitih BERT arhitektura za razumevanje konteksta i semantičkih odnosa unutar tekstova. Takođe, korišćenjem ovog pristupa je moguće dobijanje uvida u efikasnost klasifikacionih modela i otkrivanje potencijalnih nedostataka obeleženih podataka.

Za eksperiment sa LLaMA-3.2-3B-Instruct modelom, ulazi su specijalno dizajnirane instrukcije za model koje sadrže kontekstualne primere iz obeleženih podataka čime se instrukcija dodatno obogaćuje za pravilno učenje i davanje odgovora. Tokom doobučavanja nad ovako dizajniranim skupom instrukcija, osnovni model dobija sposobnost prepoznavanja emocionalnih i moralnih signala i razumevanja njihovog kontekstualnog značenja u rečenicama. Format instrukcije, koja je data na engleskom jeziku da bi je model bolje razumeo [103], je predstavljen sledećim iskazom:

LLaMA Prompt: *Classify the text based on the emotional tone into one or multiple (maximum 3) emotional categories from the following list:*

[anger, anticipation, disgust, fear, joy, neutral, sadness, surprise, trust].

Return the answer as the emotional label containing assigned categories separated by commas.

text: data [„text“]

emotional label: data [„label“]

Ovakav pristup obučavanja modela omogućava LLM modelu da nauči relevantne karakteristike značajne za klasifikaciju emocija i poboljša preciznost u prepoznavanju ovih aspekata u novim tekstovima, pri čemu doobučeni LLaMA modeli postaju prilagođeniji specifičnim zahtevima zadatka kategorizacije teksta (pogledati prilog C). Proces obučavanja LLaMA modela najpre uključuje predprocesiranje tekstualnih sadržaja maskiranjem heš oznaka, korisničkih imena i internet adresa, kao i podelu na trening, evaluacioni i test skup koristeći stratifikovanu podelu za višeznačnu klasifikaciju, na isti način kao kod doobučavanja BERT modela. Model koristi konfiguraciju za kvantizaciju u 4-bitnom formatu radi smanjenja memorijskih zahteva i *LoRA* (eng. *Low-Rank Adaptation*) tehniku koja omogućava efikasno doobučavanje velikih modela sa smanjenim brojem parametara, odnosno neophodnim računarskim resursima. Tokom treninga, korišćena je stopa učenja jednaka vrednosti 2×10^{-4} i veličina grupe za obuku od 4 instrukcije.

9. Evaluacija rezultata

9.1. Evaluacija izgrađenih semantičkih resursa

9.1.1 SentiWords.SR

Kolekcije podataka obeležene sentimentom

Za proveru ispravnosti formiranog SRPOL alata, neophodno je imati kolekciju tekstova obeleženih prema intenzitetu ili kategorijama sentimenta, što omogućava upoređivanje stvarnih obeležja sa rezultatima dobijenim pomoću SRPOL. U tu svrhu, istovremeno sa razvojem SRPOL, kreirane su dve testne kolekcije iz različitih tematskih oblasti: *Tweets.SR* i *RTS.SR*. Prva kolekcija, *Tweets.SR*, obuhvata poruke sa društvenih mreža na srpskom jeziku, prikupljene sa unapred odabranih Tviter naloga koji pišu na srpskom jeziku tokom jednog dana, i sadrži ukupno 7,668 poruka. Druga kolekcija, nazvana *RTS.SR*, sastoji se od nasumično odabranih rečenica iz novinskih tekstova *Leipzig Corpora Collection*[79] za srpski jezik objavljenih u toku 2019. godine. Ova kolekcija uključuje rečenice iz objava na glavnom nacionalnom informativnom portalu *www.rts.rs*, koji u svojim objavama pokriva širok spektar oblasti i tema – od opštih vesti do politike, kulture i sporta – pri čemu su sve rečenice napisane na standardnom savremenom srpskom jeziku. Ukupan broj rečenica u kolekciji *RTS.SR* iznosi 7,197. Tri anotatora, raznovrsnih starosnih dobi, pola i zanimanja, izvršili su obeležavanje ovih kolekcija prema pozitivnom, negativnom i neutralnom sentimentu koji je u tekstualnim sadržajima prepoznat.

Treća kolekcija obuhvata korisničke recenzije o filmovima koje su često korišćene za evaluaciju alata za analizu sentimenta. Za analizu SRPOL izdvojeno je ukupno 3,490 filmskih recenzija iz javno objavljenog i obeleženog sentimentalnog korpusa na srpskom jeziku *SentiComments.SR*. Korpus *SentiComments.SR* je anotiran prema šemi sa šest kategorija sentimenta: obeležja ± 1 označavaju pretežno pozitivan/negativan sentiment, obeležja $\pm$M označavaju dvosmislen ili mešovit sentiment sa blagom tendencijom ka pozitivnom/negativnom, dok obeležja $\pm$NS predstavljaju uglavnom neutralan sentiment sa blagom tendencijom ka pozitivnom ili negativnom sentimentu [19]. Za potrebe evaluacije SRPOL alata, korpus *SentiComments.SR* je zatim, za potrebe ovog istraživanja, pregrupisan u nove testne kolekcije na sledeći način:

- ***SC.SR I*** – uključuje sve recenzije koje su kategorisane kao pozitivne ili negativne na osnovu obeležja sentimenta. Ovaj korpus sadrži ukupno 3,490 recenzija.
- ***SC.SR II*** – uključuje recenzije obeležene kao ± 1 i $\pm$M, koje su kategorizovane kao pozitivne ili negativne (-1 i -M kao negativne, +1 i +M kao pozitivne). Ovaj korpus sadrži ukupno 2,871 recenziju.
- ***SC.SR III*** – uključuje recenzije obeležene kao ± 1 i $\pm$NS, koje su kategorizovane kao pozitivne ili negativne (-1 i -NS kao negativne, +1 i +NS kao pozitivne). Ovaj korpus sadrži ukupno 2,876 recenzija.
- ***SC.SR IV*** – uključuje samo recenzije obeležene kao ± 1, koje su kategorizovane kao pozitivne ili negativne (-1 kao negativne, +1 kao pozitivne). Ovaj korpus sadrži ukupno 2,257 recenzija.

[79] https://wortschatz.uni-leipzig.de/en/download/Serbian

- **SC.SR V** – uključuje sve recenzije, koje su kategorizovane kao pozitivne, negativne i neutralne (-1 i -M kao negativne, +1 i +M kao pozitivne, ±NS kao neutralne).

Za statističko testiranje SRPOL alata, koristićemo *SentiPol.SR* leksikon srpskih reči sa diskretnim obeležjima sentimenta, kao referentnu vrednost za poređenje ispravnosti pristupa u razvijenom leksikonu i pridruženom algoritmu. Leksikon *SentiPol.SR* je razvijen na osnovu automatskih prevoda tri engleska leksikona: NRC [139], AFFIN i Bing, koje su pratile opsežne ručne jezičke korekcije. Analiza sentimenta sprovedena je na skupu rečenica izvučenih iz 120 romana dostupnih u kolekciji SrpELTeC, koji su bili obeleženi sentimentalnim kategorijama. U poređenju sa drugim metodama, tehnika zasnovana na leksikonu pokazala se kao stabilnija za primenu u različitim skupovima tekstualnih podataka [189].

Provera leksikona

U prvom eksperimentu, izvršeno je poređenje SRPOL sa *SentiPol.SR* u sposobnosti predviđanja sentimenta recenzija u *SC.SR I* koristeći LR kao klasifikacionu metodu koja se, zbog svoje jednostavnosti, često koristi za testiranje alata za analizu sentimenta [87]. U zadatku klasifikacije, obeležje sentimenta recenzija se koristi kao zavisna binarna promenljiva, dok su vrednosti sentimenata izračunate pomoću SRPOL i *SentiPol.SR* korišćene kao prediktorske promenljive. Obe verzije modela, odnosno leksikona na koje se oslanjaju, su trenirane na 80% i testirane na preostalih 20% podataka.

Tabela 9.1: *Statistički značaj modela i zavisnih atributa u klasifikaciji sentimenta korišćenjem modela logičke regresije. Oznake: Koef. - slobodan koeficijent, SG - standardna greška, OV - odnos verovatnoća*

Model	χ^2	Koef.	SG	z	P> |z|	OV
SRPOL	813.5	8.15	0.03	28.5	0.001	3464.2
SentiPol.SR	339.8	0.95	0.05	18.4	0.001	2.58

Rezultati u tabeli 9.1 pokazuju da model SRPOL ima značajno veću Hi-kvadrat (eng. *Chi-square*, χ^2) vrednost (813.5) u poređenju sa *SentiPol.SR* (339.8), što ukazuje na bolju usklađenost modela u odnosu na osnovni model bez prediktora. Statistika χ^2 meri koliko je model sa prediktorima bolji u objašnjavanju varijanse zavisne promenljive u poređenju sa osnovnim modelom. Koeficijent za SRPOL je 8.15, sa vrlo niskom standardnom greškom (0.03), što naglašava pozitivan efekat na predviđanje pozitivnog sentimenta, dok je za *SentiPol.SR* koeficijent 0.95 sa nešto višom standardnom greškom (0.05). Oba modela imaju z-vrednosti koje ukazuju na statistički značajan uticaj prediktora (28.5 za SRPOL i 18.4 za *SentiPol.SR*), pri čemu je p-vrednost u oba slučaja 0.001, što potvrđuje značajnost prediktora u oba modela za predviđanje sentimenta. Dodatno, SRPOL pokazuje znatno veći odnos verovatnoća (3,464.2) u poređenju sa *SentiPol.SR* (2.58), što ukazuje na snažan uticaj prediktora u modelu SRPOL, čime se potvrđuje njegova superiornost u klasifikaciji sentimenta.

U drugom eksperimentu, izgrađeni su modeli binarne klasifikacije nad *Tweets.SR*, *RTS.RS* i *SC.SR I/II/III/IV* kolekcijama podataka koristeći LR kao klasifikacionu metodu. Modeli SRPOL i *SentiPol.SR* su evaluirani na osnovu tačnosti u kategorizaciji 20% podataka koji nisu bili uključeni za obuku modela. SRPOL je pokazao superiornost u tačnosti predviđanja sentimenta na svim kolekcijama. U proseku, SRPOL je tačno predvideo polaritet za 79% tekstova, dok je *SentiPol.SR* tačno predvideo 70% tekstova (pogledati tabelu 9.2). Rezultati prikazani u tabelama 9.1 i 9.2 otkrivaju nekoliko vrednih činjenica. U zadatku binarne klasifikacije, SRPOL pokazuje doslednu tačnost u rangu od 76% do 81% na testiranim kolekcijama. Posebno u skupu *SentiComments.SR*, rezultati SRPOL su u opsegu

rezultata koje su autori kolekcije prijavili kao najbolje za binarnu klasifikaciju [19]. Dodatno, u *SentiComments.SR* skupu podataka, uključenje obeležja ±M u *SC.SR II* smanjuje tačnost za -4.7% u poređenju sa *SC.SR IV*, gde su uključeni samo jasno izraženi sentimenti (±1). Zanimljivo je da je sličan efekat zabeležen sa obeležjima ±NS u *SC.SR III*, sa smanjenom tačnošću od -2.3%. Ove razlike mogu biti povezane sa anotacionom šemom koja je korišćena za ±M obeležja, koja po dizajnu sadrži dvosmislenosti u sentimentima, što su i autori kolekcije napomenuli u svom istraživanju [19].

Tabela 9.2: *Poređenje rezultata binarne klasifikacije sentimenta (pozitivna/negativna) korišćenjem LR modela nad SRPOL and SentiPol.SR leksikonima i kolekcijama podataka iz različitih domena*

	SRPOL					SentiPol.SR				
Kolekcija	Acc	F_1	Prec	Rec	ρ	Acc	F_1	Prec	Rec	ρ
Tweets.SR	0.81	0.76	0.76	0.77	**0.56**	0.71	0.68	0.67	0.70	0.30
RTS.SR	**0.81**	0.79	0.81	0.78	**0.58**	0.75	0.73	0.75	0.73	0.42
SC.SR I	0.76	0.70	0.72	0.70	0.46	0.67	0.57	0.64	0.58	0.29
SC.SR II	0.76	0.71	0.73	0.71	0.49	0.67	0.58	0.64	0.59	0.31
SC.SR III	0.78	0.72	0.75	0.71	0.48	0.71	0.61	0.67	0.61	0.33
SC.SR IV	**0.81**	0.77	0.79	0.76	**0.53**	0.71	0.62	0.68	0.61	0.38
Total Avg	**0.79**	0.74	0.76	0.74	0.52	**0.70**	0.63	0.67	0.64	0.34

Treći eksperiment obuhvatio je testiranje mogućnosti predviđanja sentimenta uključivanjem neutralne kategorije. Rezultati dobijeni u ovom eksperimentu, prikazani u tabeli 9.3, izvršeni su nad *Tweets.SR*, *RTS.RS* i *SC.SR V* kolekcijama podataka koristeći LR kao klasifikacionu metodu. Rezultati su potvrdili dobre performanse SRPOL alata, sa prosečnom postignutom tačnošću od 59.6% na zadatku višeklasne klasifikacije ($n = 3$) nad korišćenim kolekcijama.

Tabela 9.3: *Poređenje rezultata višeklasne klasifikacije sentimenta (pozitivna/negativna/neutralna) korišćenjem LR modela nad SRPOL and SentiPol.SR leksikonima i kolekcijama podataka iz različitih domena*

	SRPOL					SentiPol.SR				
Kolekcija	Acc	F_1	Prec	Rec	ρ	Acc	F_1	Prec	Rec	ρ
Tweets.SR	0.59	0.43	0.50	0.44	0.51	0.58	0.40	0.39	0.43	0.23
RTS.SR	0.60	0.48	0.53	0.46	0.52	0.57	0.39	0.45	0.42	0.31
SC.SR V	0.60	0.49	0.56	0.51	0.50	0.53	0.35	0.34	0.40	0.32
Total Avg	**0.60**	0.47	0.53	0.47		**0.56**	0.51	0.38	0.42	0.29

Dodatno, u okviru svih eksperimenata, izračunat je koeficijent korelacije ρ između tačnih obeležja sentimenta i prosečnih vrednosti sentimenta u tekstualnim sekvencama izračunatih pomoću SRPOL i *SentiPol.SR* modela. Za izračunavanje ρ u *SentiComments.SR* kolekcijama, obeležja su mapirana u numeričke vrednosti koje oslikavaju stepen sentimenta prema načinu obeležavanja (±NS $\rightarrow$ ±0.3, ±M $\rightarrow$ ±0.7). Prema standardnom pristupu za interpretaciju koeficijenata korelacije, koji preporučuju autori u [174], SRPOL, sa ρ vrednostima u rangu [0.46 - 0.58] na zadatku binarne klasifikacije, i [0.50 - 0.52] na zadatku klasifikacije sa tri klase, pokazuje umerenu korelaciju i nadmašuje *SentiPol.SR* na svim kolekcijama podataka. Autori u [87] sproveli su slične testove korelacije nad podacima sa društvenih mreža, novinskim tekstovima i filmskim recenzijama, koji su donekle uporedivi sa skupovima *Tweets.SR*, *RTS.SR* i *SentiComments.SR* korišćenim u ovom istraživanju. Sprovedeni testovi korelacija nad najpoznatijim razvijenim alatima i leksikonima za određivanje sentimenta za engleski jezik su u rangu sa ρ vrednostima dobijenim u okviru ovog istraživanja.

9.1.2 EmoLex.SR

Kolekcije podataka obeležene emocionalnim afektivnim kategorijama

Da bismo efikasno proverili ispravnost leksikona formiranog za srpski jezik, koristili smo dve različite obeležene kolekcije podataka na emocionalni afekt. Prvi je *XED*[80] višejezični skup podataka koji je obeležen pomoću osam osnovnih Plutčikovih emocionalnih kategorija [145]. Kolekcija *XED*, izgrađena nad paralelnim korpusom *OpenSubtitles2016*[81] [118], donosi emotivne projekcije filmskih titlova na 30 različitih jezika, uključujući srpski. *XED* skup podataka je prepoznat kao odgovarajući resurs za proveru valjanosti konstruisanog leksikona, zbog korišćenja istog skupa Plutčikovih emocionalnih kategorija i tekstova koji nude kontekst određenog jezičkog stila. Rečenice iz *XED* projekcija na srpskom jeziku poravnate su sa engleskim paralelnim rečenicama kako bi se otklonila potencijalna neslaganja u emocionalnim obeležjima koja mogu nastati iz tog razloga. Dodatno, duplirane stavke su eliminisane, a svi srpski tekstovi su normalizovani na srpsku latinicu i sačuvani u *UTF-8* (*Unicode*) formatu, čime su izbegnuti problemi mešanja 8-bitnih kodnih strana. Kolekcija podataka koja je nastala na ovaj način sadrži 4,317 paralelnih srpsko-engleskih rečenica dostupnih za analizu, nazvana *XED-Emo.SR* u daljem tekstu, i predstavlja provereni srpsko-engleski skup paralelnih rečenica sa emocionalnim obeležjima iz osam Plutčikovih emocionalnih kategorija.

Drugi skup podataka, nazvan *LLM-Emo.SR*, sadrži 407 paralelnih rečenica na engleskom i srpskom jeziku koje su napravljene pomoću Čet-GPT alata i klasifikovane u 8 emocionalnih i 2 sentimentalne kategorije (pogledati Prilog C). Kroz nekoliko nezavisnih iteracija, Čet-GPT model je dobio instrukcije da napravi rečenice na engleskom jeziku, prevede ih na srpski jezik i kategorizuje svaki par rečenica u jednu ili više Plutčikovih osnovnih kategorija. Parametar temperature je postavljen na 0.7 da bi se postigao viši nivo kreativnosti u generisanim sadržajima. Svi leksički problemi su ispravljeni, što je rezultovalo sintetički generisanim skupom podataka *LLM-Emo.SR*, koji je obeležen emocionalnim afektivnim kategorijama. Primeri za svaku emocionalnu kategoriju iz kolekcije paralelnih rečenica *LLM-Emo.SR* predstavljeni su u tabeli 9.4.

U tabeli 9.5 prikazana je statistika višeznačno obeleženih kolekcija podataka *XED-Emo.SR* i *LLM-Emo.SR*, koja uključuje broj poravnatih instanci, broj različitih kombinacija obeležja (eng. *Labelsets*), raznolikost (eng. *Diversity*, *Div*), gustinu (eng. *Density*, *Dens*), kardinalnost (eng. *Cardinality*, *Card*), prosečan odnos disbalansa po oznaci (*avgIR*) i procenat instanci u kolekciji sa samo jednom kategorijom u obeležju (*P_min*).

Provera leksikona

Za evaluaciju razvijenog EmoLex.SR leksikona, kao osnovu za poređenje je korišćen NRC.EmoInt leksikon koji je automatskim alatima preveden na srpski jezik. Leksikoni su evaluirani na tekstualnim sadržajima iz *LLM-Emo.SR* i *XED-Emo.SR* kolekcija podataka. Nakon tokenizacije tekstualnih sekvenci (jednačina 9.1, $t_i, i = 1, ..., m$), tokeni su obeleženi sa PoS obeležjima i lemmatizovani (jednačina 9.1, $l_i, i = 1, ..., n$) koristeći modele za određivanje vrste reči i lematizaciju napravljene za srpski jezik [191, 192]. Koristeći BoW tehniku, svaki lema_{Sr}-PoS se upoređivao sa unosima u leksikonima emocija, i ukoliko je lema_{Sr}-PoS pronađen (jednačina 9.1, $e_i, i = 1, ..., k$), odgovarajuća vrednost je dodata u listu vrednosti za svaku od kategorija. Da bi se utvrdile emocionalne kategorije teksta, prosek emocionalnih vrednosti je izračunat za svaku od kategorija. Kategorije sa najvećim vred-

[80] https://huggingface.co/datasets/xed_en_fi
[81] https://www.opensubtitles.org/

Tabela 9.4: *Primeri iz LLM-Emo.SR paralelne kolekcije podataka kreirane i kategorisane u Plutčikove emocionalne kategorije pomoću Čet-GPT alata*

Emocija	Jezik	Primer
joy	En	Sarah smiled as she opened the present. She felt delighted and grateful.
	Sr	Sara se nasmejala dok je otvarala poklon. Osećala se oduševljeno i zahvalno.
anticipation	En	As the plane took off, Emily felt a rush of excitement. She was looking forward to her vacation in Hawaii.
	Sr	Kada je avion poleteo, Ema je osetila uzbuđenje. Jedva je čekala odmor u Havajima.
anger	En	After finding out about the betrayal, Mark's face turned red with anger. He couldn't believe his friend had lied to him.
	Sr	Nakon saznanja o izdaji, Markovo lice se crvenelo od besa. Nije mogao da veruje da ga je prijatelj lagao.
disgust	En	The smell from the garbage made Jane's stomach turn. She couldn't stand the sight of the rotting food.
	Sr	Miris iz smeća je izazvao mučninu kod Jane. Nije mogla da podnese prizor trule hrane.
sadness	En	The weight of her sorrow is crushing her heart, and she felt like she'll never be able to escape the endless cycle of pain and emptiness.
	Sr	Težina njene tuge pritiskala joj srce i činilo joj se kao da nikada neće moći izaći iz beskrajnog ciklusa bola i praznine.
surprise	En	When she opened the gift, Maria's eyes widened with surprise. She never expected to receive such a thoughtful present.
	Sr	Kada je otvorila poklon, Marijine oči su se raširile od iznenadjenja. Nikada nije očekivala da će dobiti tako pažljivo odabran poklon.
fear	En	As the storm approached, John's heart raced with fear. He was terrified of the lightning and thunder.
	Sr	Kada se oluja približila, srce Jovana je ubrzano kucalo od straha. Bio je prestrašen od munja i gromova.
trust	En	As she hugged her best friend, Maria felt a sense of comfort and safety. She knew she could always count on her.
	Sr	Dok je zagrljajem pozdravljala najbolju drugaricu, Marija je osetila osećaj udobnosti i sigurnosti. Znala je da joj uvek može verovati.

Tabela 9.5: *Statistika višeznačno obeleženih kolekcija emocionalnog afekta*

Kolekcija	Instanci	*Labelsets*	*Div*	*Dens*	*Card*	*avgIR*	*P_min*
XED-Emo.SR	4,317	112	0.44	0.18	1.47	1.51	0.63
LLM-Emo.SR	407	25	0.09	0.15	1.27	3.62	0.73

nostima određene su kao konačne emocionalne kategorije, za dati leksikon i tekst koji se razmatra (jednačina 9.2).

$$text = \{t_1, t_2, ..., t_m\} = \{l_1, l_2, ..., l_n\} = \{e_1, e_2, ..., e_k\}, m \leq n, k \leq n \tag{9.1}$$

$$Labels^L_{text} = \underset{i(1,2)}{\text{argmax}} \frac{\sum_{j=1}^{k} ascore^i_e}{|k|}, i = 1, 2, ..., L, |L| = 8 \tag{9.2}$$

Razvijene verzije leksikona emocija *EmoLex.SR-v(1,2)* upoređivane su sa osnovnim NRC.EmoLex.tr u rečenicama napisanim na srpskom jeziku iz paralelnih *LLM-Emo.SR* i *XED-Emo.SR* kolekcija tekstualnih podataka. Kategorija *neutral* je uklonjena iz razmatranja u slučajevima kada postoji kao kategorija u leksikonima *EmoLex.SR-v(1,2)*. Tabela 9.6 predstavlja makro F_1, Prec, Rec po svakoj od kategorija, kao i njihovu srednju vrednost, dobijenih prilikom poređenja pravih obeležja sa obeležjima identifikovanim pomoću leksikona (jednačina 9.2). Dobijeni rezultati potvrđuju potrebu za razvojem leksikona prilagođenim

srpskom jeziku, što dokazuju i poboljšanja u odnosu na osnovni model, koja se mogu primetiti na svim metrikama uporedo sa unapređenjem leksikona (v(1,2)).

U rezultatima za *LLM-Emo.SR* kolekciju podataka može se primetiti da samo za kategoriju *anger* osnovni model nije poboljšan, za kategorije *anticipation* i *sadness* prva verzija leksikona je dala najbolju F_1 vrednost, dok je za ostalih pet kategorija najbolja F_1 vrednost dobijena sa proširenom verzijom leksikona. Prosečna vrednost makro F_1 vrednosti povećana je za 7.1%, sa 69.1% na 76.2%. Na konverzacionoj kolekciji podataka *XED-Emo.SR*, uočljive su niže vrednosti na svim merikama, što može biti uzrokovano specifičnostima neformalnog jezika i potencijalnim nepoklapanjima koja mogu postojati u anotacionim šemama između kolekcije podataka i leksikona (pogledati odeljak 9.1.2). Veći skok u rezultujućim vrednostima za verzije leksikona *v1* i *v2* može se pripisati uključivanju sinonima i korišćenim tehnikama za izračunavanje emocionalnog skora (normalizacija i agregacija). Leksikon *v1* je pokazao optimalne performanse za emocionalne kategorije *anger* i *fear*, dok se leksikon *v2* pokazao superiornijim u identifikovanju preostalih kategorija. Srednja vrednost makro F_1 mere na *v2* modelu povećana je za 7.5%, sa 15.6% na 23.1%, u odnosu na osnovni model.

Tabela 9.6: *Poređenje rezultata višeznačne klasifikacije emocija korišćenjem NRC.EmoInt.tr (osnova) i EmoLex.SR-(v1, v2) leksikona nad LLM-Emo.SR i XED-Emo.SR kolekcijama tekstualnih podataka*

	NRC.EmoInt.tr (osnova)			EmoLex.SR-v1			EmoLex.SR-v2		
LLM-Emo.SR	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec
anger	**0.852**	0.958	0.767	0.784	0.833	0.741	0.809	0.792	0.826
anticipation	0.740	0.626	0.905	**0.783**	0.714	0.867	0.723	0.615	0.875
disgust	0.732	0.577	1.000	0.800	0.692	0.947	**0.808**	0.731	0.905
fear	0.831	0.822	0.841	0.819	0.756	0.895	**0.874**	0.844	0.905
joy	0.728	0.639	0.848	0.720	0.651	0.806	**0.745**	0.669	0.841
sadness	0.700	0.636	0.778	**0.876**	0.836	0.920	0.830	0.800	0.863
surprise	0.500	0.346	0.900	0.638	0.577	0.714	**0.644**	0.538	0.800
trust	0.442	0.358	0.576	0.586	0.641	0.539	**0.661**	0.679	0.643
Macro Avg	0.691	0.620	0.827	**0.751**	0.792	0.804	**0.762**	0.708	0.832
XED-Emo.SR	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec
anger	0.104	0.060	0.392	**0.131**	0.078	0.416	0.114	0.065	0.453
anticipation	0.151	0.094	0.383	0.207	0.138	0.413	**0.281**	0.222	0.381
disgust	0.112	0.064	0.432	0.123	0.078	0.285	**0.159**	0.114	0.260
fear	0.187	0.133	0.318	**0.191**	0.138	0.313	0.181	0.126	0.317
joy	0.282	0.212	0.422	0.369	0.324	0.431	**0.418**	0.449	0.391
sadness	0.196	0.135	0.357	0.232	0.170	0.365	**0.255**	0.209	0.327
surprise	0.052	0.029	0.259	0.060	0.0334	0.293	**0.148**	0.099	0.295
trust	0.167	0.129	0.230	0.266	0.301	0.238	**0.295**	0.370	0.245
Macro Avg	0.156	0.107	0.349	**0.197**	0.157	0.344	**0.231**	0.207	0.334

U dodatnoj analizi predstavljenoj u tabeli 9.7, rezultati sugerišu da *EmoLex.SR-v2* nadmašuje druge leksikone preko drugih uključenih metrika, kao što su Pokrivanje leksikona (eng. *Lexicon Coverage, LexCvg*), EMR, Acc, HS, HL, kao i makro F_1, Prec i Rec, što dodatno potvrđuje efikasnost leksikona u višeznačnoj klasifikaciji emocija izvršenoj nad *LLM-Emo.SR* kolekcijom podataka. HS odražava sličnost između predviđenih i istinitih etiketa, dok HL meri deo obeležja koje su netačno predviđene. Leksikon *EmoLex.SR-v2* ostvaruje najviši HS sa 72.3% i najniži od HL od 7.2%, što ukazuje na bolje ukupne performanse u predviđanju emocija u poređenju sa drugim leksikonima. Pored toga, pažljivim LC prilagođavanjem srpskih leksikona nadmašuje se učinak engleskog leksikona NRC.EmoInt

na svim merama, kada se ovi leksikoni primene na paralelne jezičke tekstove dostupne u *LLM-Emo.SR* kolekciji podataka.

Tabela 9.7: *Poređenje rezultata klasifikacije emocionalnog afekta dobijenih pomoću engleskog i srpskih leksikona nad LLM-Emo.SR kolekcijom paralelnih tekstova*

Leksikon	LexCvg	EMR	Acc	F_1	Prec	Rec	HS	HL
NRC.EmoInt	0.950	0.534	0.912	0.694	0.618	0.802	0.649	0.087
NRC.EmoInt.tr	0.928	0.535	0.917	0.691	0.620	0.827	0.652	0.083
EmoLex.SR-v1	0.975	0.560	0.922	0.751	0.713	0.804	0.706	0.077
EmoLex.SR-v2	0.995	0.585	0.927	0.762	0.708	0.832	0.723	0.072

Provera Čet-GPT rezultata

Čet-GPT *gpt-3.5-turbo* model korišćen je za rešavanje zadataka T1–T4 postavljenih u toku konstrukcije emocionalnog rečnika na srpskom jeziku. U okviru ovog odeljka ćemo pokušati da kvantifikujemo ukupan doprinos ovog LLM modela u rešavanju ovih zadataka.

Na zadatku prevođenja pojedinačne reči (T1), Čet-GPT je uspešno uklonio uočena ograničenja sa GT API, što je rezultiralo značajnim poboljšanjem ukupne preciznosti, sa 62.5% na 85.0%. Model Čet-GPT je uspeo da ispravi 27.1% svih prevoda (72.4% nekorektnih GT prevoda) dobijenih pomoću GT alata. Pored toga, GT alat je proizveo bolje rezultate prevođenja u 4.7% svih prevoda (31.9% nekorektnih Čet-GPT prevoda). Konačno, najbolji rezultati su postignuti korišćenjem kombinacije oba alata za automatsko prevođenje, što je doprinelo ukupnoj tačnosti od 89.6% (pogledati tabelu 9.8).

Tabela 9.8: *Tačnost GT i Čet-GPT alata na zadatku prevođenja pojedinačne reči*

	Nekorektni		Korektni		
Korektni	Čet-GPT	GT	Čet-GPT	GT	Total
GT	0.047 (0.319)	-	0.578 (0.680)	-	0.625
Čet-GPT	-	0.271 (0.724)	-	0.578 (0.925)	0.850
Total	0.047	0.271	0.578		0.896

Na zadatku obeležavanja pojedinačnih reči u kategorije afekata (T2), kategorije generisane pomoću Čet-GPT alata imaju 47.0% makro F_1 u predviđanju konačnog obeležja u NRC.EN leksikonu (EN-Gold) i nadmašuju WNA leksikon koji ima F_1 od 40.0% koji se koristio za prilagođavanje NRC.EN obeležja. U predviđanju konačnih labela u srpskom leksikonu (SR-Gold), brojevi se povećavaju u korist Čet-GPT. Posmatrano ponašanje je uzrokovano LC između jezika, ali i načinom izgradnje leksikona NRC.EN, korišćenih šema za obeležavanje i potencijalnih grešaka koje mogu nastati prilikom dodeljivanja emocionalnih obeležja (pogledati tabelu 9.9). Rezultati takođe pokazuju da su konačna obeležja EmoLex.SR-Gold i nakon prilagođavanja, uglavnom zasnovana na NRC.EN-Gold obeležjima (EN-Gold $\rightarrow$ SR-Gold) sa vrednošću od 80.0% makro F_1 mere (EN-Gold $\rightarrow$ SR-Gold).

Evaluacija uspešnosti generisanja sinonima na zadatku (T3), predstavljena u tabeli 9.10, podrazumeva procenjivanje tačnosti i zastupljenosti Čet-GPT sinonima prikupljenih u Total i Gold listama sinonima. Model Čet-GPT je u mogućnosti da generiše sinonime za 87.5% od reči iz leksikona, u poređenju sa 22.7% reči za koje su sinonimi pronađeni koristeći SWN leksički resurs, čime se značajno proširuju postojeće mogućnosti pronalaženja sinonima. Sinonimi zasnovani na Čet-GPT pokazuju 0.3 F_k i 60.26% poklapanja sa Inc sinonimima. Prethodni brojevi sugerišu da je Čet-GPT u stanju da generiše sinonime u većoj meri od

Tabela 9.9: *Stepen uticaja Čet-GPT i WNA emocionalnih obeležja na EN/SR-Gold obeležja*

	EN-Gold		⟹	SR-Gold		
Metrika	Čet-GPT	WNA.SR		Čet-GPT	EN-Gold	WNA.SR
Prec	0.500	0.395		0.588	0.758	0.399
Rec	0.529	0.456		0.738	0.890	0.495
F_1	0.470	0.400		0.608	0.800	0.418
HS	0.413	0.386		0.608	0.745	0.436
F_k	0.183	0.156		0.401	0.663	0.184

standardnih resursa, ali ovi LLM izlazi zahtevaju pažljivu ručnu proveru njihove ispravnosti. Sinonimi zasnovani na Čet-GPT čine značajan deo Gold liste sinonima, preovlađujući nad SWN i Man izrađenim sinonimima, sa 0.47 F_k i 74.4% svoje pokrivenosti.

Tabela 9.10: *Merenje uključenosti Čet-GPT i SWN lista sinonima u Total i Gold listama sinonima*

Syn	Syn	LexCvg	F_k	HS	F_1	Prec	Rec
Čet-GPT	SWN	0.227	0.04	0.877	0.059	0.065	0.058
	Inc	0.236	0.31	0.949	0.609	0.603	0.629
Total	Čet-GPT	0.875	0.63	0.844	0.773	0.758	0.804
	SWN	0.227	0.02	0.454	0.464	0.449	0.496
	Inc	0.236	0.19	0.502	0.437	0.427	0.458
Gold	Čet-GPT	0.875	0.48	0.784	0.745	0.744	0.765
	SWN	0.227	0.04	0.433	0.447	0.442	0.470
	Man	0.205	0.25	0.526	0.424	0.415	0.445

U okviru zadatka za generisanje paralelnih rečenica obeleženih emocionalnim afektivnim kategorijama (T4), model Čet-GPT je pokazao izuzetan doprinos. Lingvistički eksperti su izvršili ručnu proveru 1,000 rečenica koje su napravljene pomoću Čet-GPT modela. U tom postupku je identifikovano 5.4% rečenica sa nepravilnom strukturom. Uočene nepravilnosti su se odnosile na odstupanja u korišćenju odgovarajućeg oblika reči, očekivanim redosledom reči u rečenici i uočljivim prisustvom leksički i semantički sličnih rečenica. Pored toga, 2.7% rečenica je imalo netačna obeležja emocionalnih kategorija, odnosno kategorije nisu bile deo unapred definisane liste emocionalnih kategorija, kao što su *pride* (ponos) ili *determination* (odlučnost). Semantički slične rečenice, njih 59.3% od inicijalne kolekcije, identifikovane su i uklonjene korišćenjem višejezičnog transformer modela [219] koji podržava srpski jezik[82]. Uprkos uočenim nepravilnostima, kao što su semantički slične rečenice ili prisustvo kategorija koje ne pripadaju unapred definisanom skupu emotivnih kategorija, konačna prečišćena kolekcija paralelnih rečenica predstavlja vredan resurs za analizu emocija na srpskom jeziku. Primenom uspostavljene procedure za kreiranje i proveru, *LLM-Emo.SR* kolekcija tekstualnih podataka se može dodatno kvantitativno i kvalitativno unaprediti ponavljanjem postupka i korišćenjem naprednijih verzija Čet-GPT modela.

9.1.3 MFD.SR

Leksikon *MFD.SR* je proveren upoređivanjem prosečne učestalosti reči iz leksikona u odgovorima na pitanja o pet moralnih vrednosti koja su bila uključena u anketi o razumevanju moralnih vrednosti na srpskom jeziku (pogledati prilog D). Anketa je organizovana

[82] https://huggingface.co/sentence-transformers/use-cmlm-multilingual

putem internet društvenih grupa i obuhvatila je 3 grupe sa ukupno 51 pitanjem.U anketi je učestvovalo 370 ispitanika, od kojih je 260 studenata i 110 anonimnih korisnika društvenih mreža. U okviru druge grupe pitanja, od ispitanika je zatraženo da u obliku deskriptivnih tekstualnih odgovora opišu pet moralnih vrednosti predloženih MFT teorijom na način kako ih oni doživljavaju, kao i da prilože primere njihovog poštovanja i kršenja.

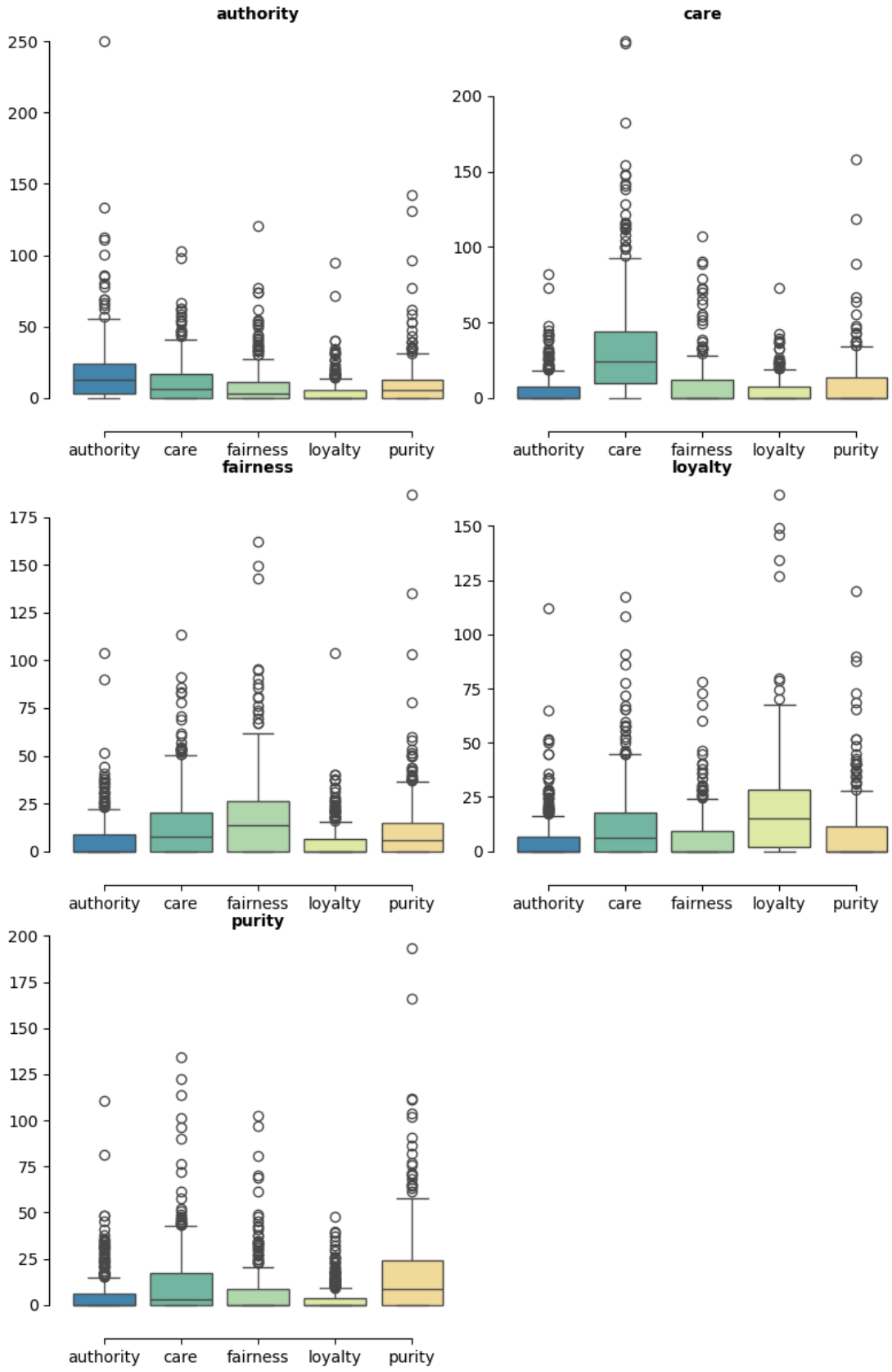
Tekstualne sekvence u odgovorima ispitanika su obrađene korišćenjem jezičkih resursa razvijenih za srpski jezik. Iz skupa odgovora su najpre uklonjeni odgovori koji nisu bili razumljivi ili nisu bili relevantni za postavljeno pitanje, a zatim je nad preostalim tekstualnim sekvencama primenjena restauracija dijakritika [123]. Svaki tekstualni opis moralne vrednosti (kombinovanjem odgovora za vrlinu i manu) je tokenizovan, određene su vrste reči, kao i osnovni oblici reči, odnosno leme prema morfološkim pravilima srpskog jezika [191]. Nakon toga su konstruisani skupovi moralno relevantnih lema_{Sr}-PoS parova za svaku od moralnih vrednosti. Na primer, za skup reči proizašao iz konteksta povezanog sa moralnom osnovom briga (kategorija *care*) u opisima svakog učesnika, posebno je izračunat broj pojavljivanja lema_{Sr}-PoS parova povezanih sa svim moralnim kategorijama koje prepoznaje *MFD.SR* leksikon (kategorije *authority*, *care*, *fairness*, *loyalty*, *purity*). Najzad, brojevi pojavljivanja lema_{Sr}-PoS parova u svakom od oformljenih skupova su normalizovani ukupnim brojem lema_{Sr}-PoS parova iz leksikona povezanih sa nekom moralnom kategorijom u okviru moralne osnove i ukupnim brojem lema_{Sr}-PoS parova za datu moralnu osnovu.

Na slici 9.1 prikazan je prosečan odnos učestalosti lema_{Sr}-PoS parova za svaku moralnu osnovu prepoznatih korišćenjem *MFD.SR* leksikona. Statistika pojavljivanja ukazuje na značajno veću učestalost lema_{Sr}-PoS parova iz moralne kategorije koja je povezana sa moralnom osnovom u poređenju sa lema_{Sr}-PoS parovima iz drugih kategorija. Statistička mera F_s izvršena za skupove pojavljivanja lema_{Sr}-PoS parova u svakoj moralnoj vrednosti, pokazuje statistički značaj ovih skupova vrednosti u svakoj od grupa (*authority*: F_s=34.4, p$\leq$0.05; *care*: F_s=138.9, p$\leq$0.05; *fairness*: F_s=40.9, p$\leq$0.05; *loyalty*: F_s=55.5, p$\leq$0.05; *purity*: F_s=39.4, p$\leq$0.05), čime je potvrđena ispravnost *MFD.SR* leksikona.

Dodatna provera ispravnosti leksikona je izvršena korišćenjem treće grupe pitanja iz sprovedene ankete koja predstavljaju srpski prevod Upitnika o moralnim osnovama (eng. *Moral Foundations Questionnaire, MFQ*). Ovaj upitnik omogućava procenu vrednovanja sklonosti ispitanika ka određenim moralnim vrednostima. U okviru ankete je korišćena verzija MFQ sa 32 pitanja, koji je preveden na srpski, potvrđen i objavljen[83] uz pridruženu skalu i sistem za vrednovanje pitanja (pogledati prilog D). Provera leksikona pomoću MFQ je izvršena ispitivanjem odnosa između specifičnih atributa tekstualnih opisa i MFQ vrednosti za svakog ispitanika. Pretpostavka od koje se polazi je da ispitanici koji imaju veće MFQ vrednosti za određenu moralnu osnovu su u mogućnosti da tu moralnu osnovu bolje jezički opišu korišćenjem brojnijih i raznovrsnijih moralnih jezičkih fraza i termina. U tom cilju je za svaki tekstualni odgovor koji je povezan sa nekom moralnom osnovom izračunat udeo moralnih lema_{Sr}-PoS parova, odnosno lema_{Sr}-PoS parova koji su ispravni u srpskom jeziku i prepoznate pomoću *MFD.SR* leksikona za datu moralnu osnovu. Pirsonov koeficijent korelacije (r) između MFQ vrednosti i udela *MFD.SR* lema_{Sr} koje su ispitanici koristili u svojim opisima je pokazao postojanje korelacije za moralne osnove *care* (r=0.13, p=0.01) i *loyalty* (r=0.11, p=0.03), dok nije dobijena značajna korelacija za preostale tri moralne osnove (*fairness*: r=0.06, p=0.23; *authority*: r=0.07, p=0.20; *purity*: r=0.04, p=0.47).

Leksikon *MFD.SR* omogućava pravilnu kategorizaciju moralno relevantnih situacija u tekstovima na srpskom jeziku prema odgovarajućim moralnim osnovama, što predstavlja potvrdu ispravnosti izgrađenog leksikona. Dodatno, analiza korelacija sa MFQ vrednosti-

[83] https://moralfoundations.org/questionnaires/



Slika 9.1: *Raspodela kategorija moralnog sentimenta u MFD.SR leksikonu*

ma je pokazala postojanje bolje razvijenih jezičkih sistema za moralne kategorije *care* i *loyalty*, dok takvo pravilo nije uočeno za preostale moralne kategorije. Delimično objašnjenje za ove rezultate je da su moralne osnove *care* i *loyalty* bolje kvantifikovane u MFQ, dok su moralne osnove *fairness*, *authority* i *purity* više zavisne od kulture i jezika, što implicira potrebu za prilagođavanjem MFQ upitnika i načina vrednovanja koji bi odgovarali srpskoj kulturi. Ove interpretacije su delimično u skladu sa prethodnim istraživanjima, koja pokazuju da su vrednosti *care* i *fairness* centralne za moralno prosuđivanje u različitim kulturama [73], dok su vrednosti *loyalty*, *authority* i *purity* podložnije političkoj ideologiji, kulturi i religioznosti [67]. Drugi razlog se može pronaći u mogućoj nedovoljnoj pokrivenosti rečnika odgovarajućim lema_{Sr} za kategoriju *fairness*, što bi zahtevalo dalje obogaćivanje leksikona.

Važno je napomenuti nešto veću zastupljenost mlađe populacije u uzorku ispitanika, koju čine 65.2% mladih od 18-30 godina, kao i 62% žena. Prethodna istraživanja sa međunarodno raznolikim uzorcima pokazala su da žene češće pokazuju više vrednosti na moralnim temeljima *care*, *fairness* i *purity* nego muškarci, dok muškarci postižu više vrednosti od žena na osnovama *loyalty* i *authority* u MFQ [66]. U našem uzorku, žene su postigle više vrednosti u svim moralnim kategorijama, pri čemu su dostigle približno jednake vrednosti u kategoriji *authority* (pogledati tabelu D.2 u prilogu D). Nedosledni obrasci u kategoriji *fairness* iz analiza korelacija MFQ moralnih atributa potencijalno su uzrokovani nedovoljnim brojem ispitanika muškog pola, starijih, kao i drugih socio-demografskih grupa, u korišćenom uzorku. Iako je ovom analizom potvrđena ispravnost *MFD.SR* leksikona za istraživanje moralnosti u srpskom jeziku, ona takođe naglašava potrebu za produbljivanjem istraživanja kako bi se ispitali uzročni odnosi između MFQ vrednosti i upotrebe moralno relevantnih reči na osnovu *MFD.SR* leksikona.

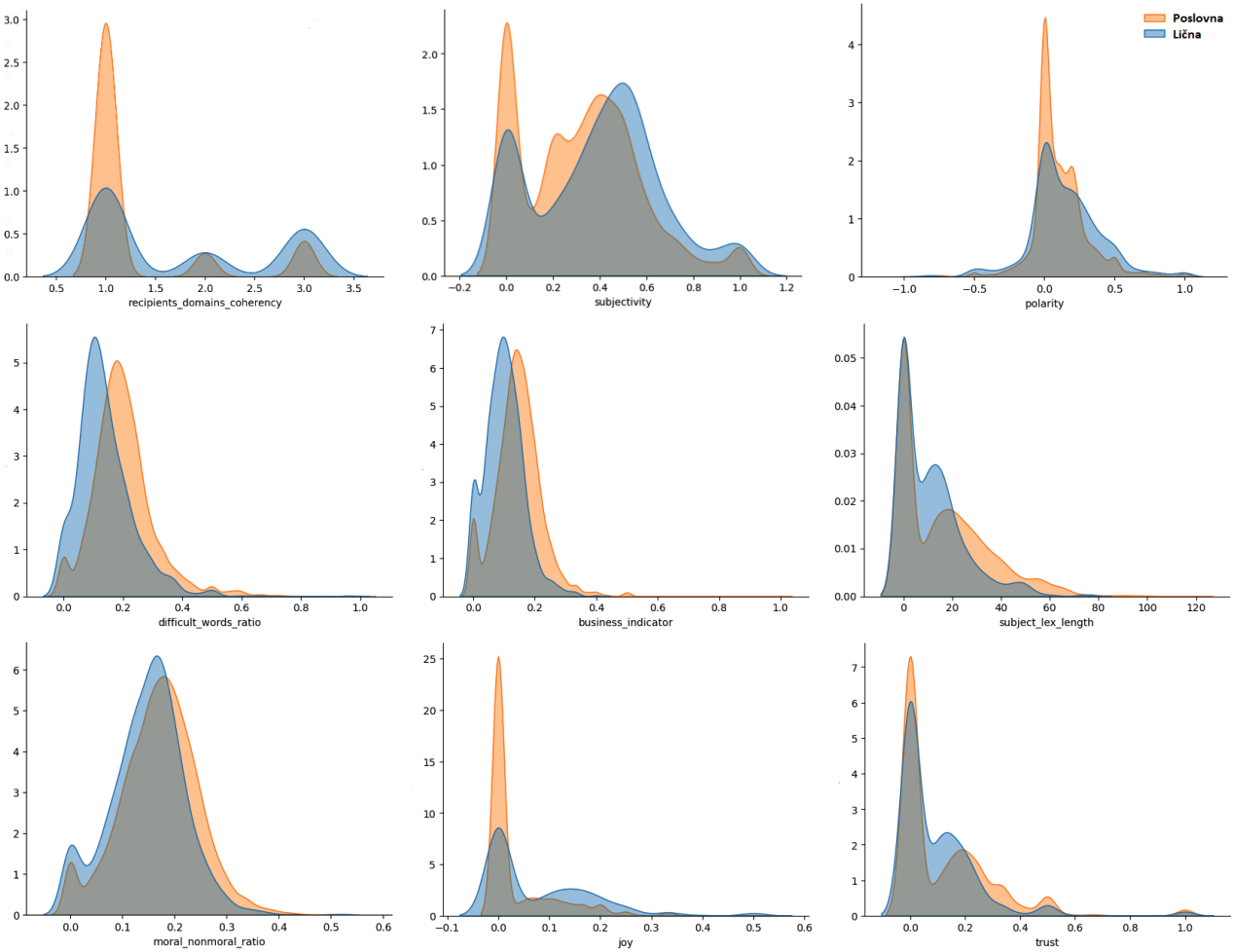
9.2. Emocionalni i moralni atributi kao nezavisne promenljive

9.2.1 Poslovna nasuprot ličnoj poruci

Značajni atributi klasifikacije

Za analizu značaja izračunatih Meta atributa najpre je korišćen pristup jedne promenljive zasnovan na F_s, kojim se meri razlika između grupa u zavisnom atributu u odnosu na pojedinačne nezavisne atribute (pogledati prilog A, tabelu A.2). Dodatno, izvršena je procena gustine kernelom (eng. *Kernel Density Estimation, KDE*) koja omogućava vizualizaciju raspodelu vrednosti atributa u različitim klasama. Grafikon KDE na slici 9.2 pokazuje kako se raspodela vrednosti pojedinih atributa razlikuje u njihovim raspodelama u klasama **`Poslovna`** i **`Lična`**. Dužina naslova poruke, odnos složenih i moralnih reči pokazuju veće vrednosti u klasi **`Poslovna`**. Klasa **`Lična`** ima veće vrednosti subjektivnosti i polariteta, kao i broj reči koje se pripisuju emocionalnom stanju radosti. Atribut koji meri koherentnost domena primalaca pokazuje veće vrednosti u klasi **`Lična`**, što ukazuje i na očekivanu veću raznovrsnost domena primalaca u privatnim porukama i potencijalno visok značaj ovog atributa za zadatak **`PL`** klasifikacije.

Multikolinearni atributi iz skupa Meta atributa su uklonjeni primenom hijerarhijskog grupisanja na ρ, sa pragom od 0.7, pri čemu je zadržan jedan atribut iz svake grupe kolinearnih atributa. Na dobijenoj listi nekolinearnih atributa, primenjeni su algoritmi permutacije vrednosti i izbacivanja jednog atributa (pogledati odeljak 4.1.3) da bi se izmerio uticaj svakog atributa na varijabilnost tačnosti modela i koristili ovu meru za izbor konačnog



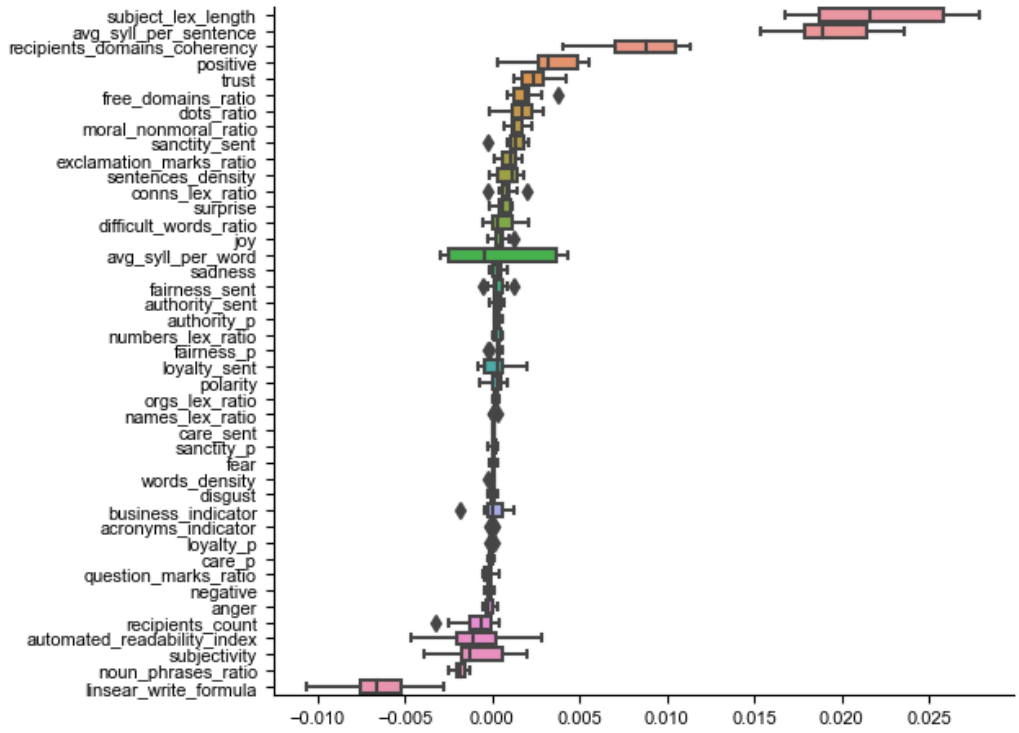
Slika 9.2: *Prikaz varijacija u raspodeli vrednosti atributa u klasama* **`Poslovna`** *i* **`Lična`**

skupa nekolinearnih i najvažnijih Meta atributa za svaki pristup.

Poređenjem rezultata metoda permutacije vrednosti atributa (slika 9.3) i isključivanja atributa (slika 9.4) korišćenih za procenu važnosti atributa u klasifikaciji poruka na klase **`Poslovna`** i **`Lična`**, primećuje se značajna sličnost u ključnim atributima, ali i određene razlike u prioritetima. Oba pristupa identifikuju sledeće atribute:

- *joy*, *fear* i *trust* (EmoAtr),
- *fairness_sent*, *authority_sent*, *loyalty_sent* i *moral_nonmoral_ratio* (MorAtr),
- *recipients_domains_coherency*, *free_domains_ratio* i *recipients_count* (ConAtr),
- *subjectivity*, *polarity*, *FRES*, *LWM* i *ARI* (ExpAtr),
- *business_indicator*, *ASPW*, *avg_word_length*, *subject_length*, *word_density*, *sentence_density* (LexAtr),
- *names_ratio*, *numbers_ratio* i *connectors_ratio* (NERAtr),
- *exclamations_ratio* i *dots_ratio* (PncAtr).

kao važne za razlikovanje poslovnih i ličnih poruka, što ukazuje na to da moralni i emocionalni ton značajno doprinose klasifikaciji. Ipak, metoda permutacija dodatno ističe *recipients_count* i *subject_lex_length* kao najvažnije, dok su ovi atributi u metodi isključivanja rangirani niže, što sugeriše da permutacije daju veći značaj konverzacionim i leksičkim



Slika 9.3: *Stepen značaja atributa u klasifikaciji* **Poslovna** *i* **Lična** *korišćenjem metode permutacije vrednosti atributa*

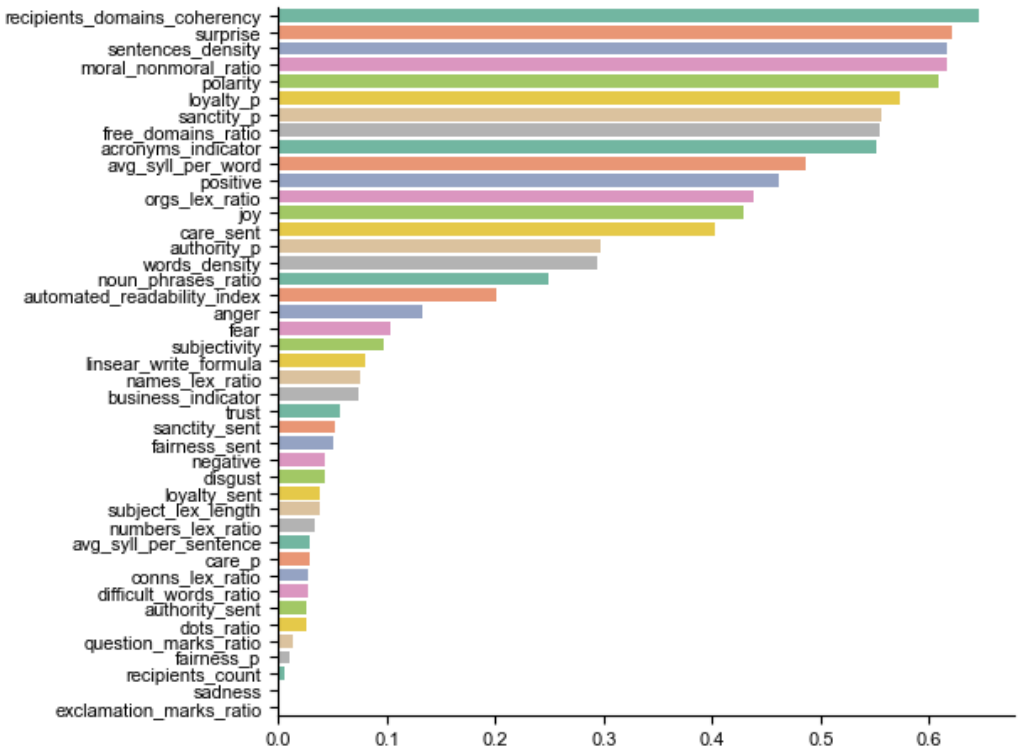
atributima. Takođe, metoda isključivanja naglašava važnost moralnih atributa, poput *moral_nonmoral_ratio* i *loyalty_p*, dok metoda permutacija vrednosti daju veću težinu atributima za merenje čitljivosti teksta, kao što su *linsear_write_formula* i *avg_word_length*. Ovo sugeriše da metoda isključivanja atributa bolje identifikuje kontekstualne i moralne aspekte, dok metoda permutacije bolje reflektuje strukturne karakteristike i formalnost teksta. Na osnovu rezultata dobijenih u dve različite metode, kao i važnosti pojedinačnih atributa izračunatih primenom pristupa jedne promenljive na ovom zadatku (prilog A, tabela A.2), izračunat je konačan stepen važnosti svakog atributa koji je predstavljen sledećim fragmentom koda:

```
FUNCTION select_features(data, f_stats_file, perm_importance_file,
    drop_one_file, p_thresh=0.005, importance_thresh=0.3, corr_thresh=0.7):

    // Korak 1: Učitavanje stepena važnosti dobijenih različitim pristupima
    f_stats = LOAD(f_stats) WHERE p_value >= p_thresh
    perm_importance = LOAD(perm_importance)
    drop_one = LOAD(drop_one)

    // Korak 2: Normalizacija stepena značajnosti atributa u [0,1] opseg
    FOR EACH metric IN [f_stats, perm_importance, drop_one]:
        min_val = MIN(metric.importance_values)
        max_val = MAX(metric.importance_values)
        metric.normalized = (metric.importance - min_val)/(max_val - min_val)

    // Korak 3: Izračunavanje konačnog stepena značajnosti atributa
    combined_importances = []
    FOR EACH feature IN UNION(features_from_all_metrics):
        importance = (GET_IMPORTANCE(f_stats, feature) +
                GET_IMPORTANCE(perm_importance, feature) +
                GET_IMPORTANCE(drop_one, feature)) / 3
        combined_importances.APPEND((feature, importance))
```



Slika 9.4: *Stepen značaja atributa u klasifikaciji* **Poslovna** *i* **Lična** *korišćenjem metode isključivanja jednog atributa*

```
// Korak 4: Sortiranje i odabir atributa
combined_importances.SORT_BY(importance DESCENDING)
selected_features = [feature FOR (feature, importance) IN
combined_importances
                    WHERE importance > importance_thresh]

// Korak 5: Dodatno uklanjanje potencijalnih visoko korelisanih atributa
final_features = []
FOR i, feature1 IN selected_features.ENUMERATE():
    keep_feature = TRUE
    FOR j, feature2 IN selected_features.ENUMERATE():
        IF i != j AND CORRELATION(data[feature1], data[feature2]) >
corr_thresh:
            IF combined_importances[i].importance <= combined_importances[j
].importance:
                keep_feature = FALSE
                BREAK
    IF keep_feature:
        final_features.APPEND(feature1)

RETURN final_features
```

Fragment koda 9.1: *Metoda odabira konačnog skupa značajnih atributa korišćenjem pristupa jedne i više promenljivih*

Koristeći prethodni algoritam za kombinovanje tri pristupa i postavljanjem granica za prihvatljiv stepen značajnosti (p_thresh=0.005, $importance\_thresh$=0.3) i koeficijent korelacije između atributa ($corr\_thresh$=0.7), formirana je konačna lista značajnih atributa za klasifikaciju poruka u klase **Poslovna** i **Lična** (**PL**):

```
značajni_atributi = [
    "recipients_count",
    "subject_lex_length",
    "recipients_domains_coherency",
    "positive",
    "joy",
    "loyalty_p",
    "moral_nonmoral_ratio",
    "polarity",
    "trust",
    "linsear_write_formula",
    "avg_word_length",
    "free_domains_ratio",
    "sanctity_p",
    "business_indicator",
    "surprise",
    "fairness_p"
]
```

Fragment koda 9.2: *Lista značajnih Meta atributa u klasifikaciji poruka elektronske pošte u klase* **Poslovna** *i* **Lična** *(***PL***)*

koji se mogu iskoristiti kako bi se optimizovala izgradnja modela i poboljšale njegove performanse. Izabrana metoda za procenu značajnosti atributa je efikasna jer kombinuje statističku analizu (F_s) sa dve tehnike nezavisne od algoritma. Na ovaj način procena značajnosti atributa se vrši kroz različite metodološke aspekte čime se smanjuje rizik od potencijalnih i neželjenih arhitekturalnih zavisnosti prilikom odabira atributa. Ovako odabrani atributi omogućavaju formiranje efikasnijih i interpretabilnijih modela kroz smanjenje dimenzionalnosti celokupnog skupa Meta atributa, što je posebno korisno za tradicionalne ML algoritme kao što je *SGDClassifier*. Za DNN modele, odabrani atributi služe kao početna osnova za izbor atributa uz korišćenje dopunskih tehnika kao što je L2 regularizacija kako bi se identifikovale potencijalno važne nelinearne zavisnosti u ulaznim podacima. Dodatno, odabrani atributi olakšavaju interpretabilnost modela kroz alate kao što je SHAP, dok upotreba ovih atributa smanjuje računsko opterećenje i omogućava efikasnije praćenje performansi modela u toku obučavanja. Ovaj pristup, iako robustan, zahteva pažljivu proveru kako bi se osiguralo da se ne izgube ključne informacije, posebno pri radu sa kompleksnim modelima koji mogu prepoznati složenije obrasce u podacima.

Rezultati klasifikacije

Za procenu performansi tehnika koristimo standardne mere evaluacije za binarnu klasifikaciju koje potiču iz teorije o pretraživanju informacija – Prec, Rec i F_1, Acc i Acc_{Bal}. Cilj nam je da poboljšamo opštu i izbalansiranu tačnost modela klasifikacije, kao i F_1 na manje zastupljenoj klasi **Lična**. Rezultati poređenja BoW, Tf-Idf i BERT Embd sa i bez uključenih Meta atributa u SGD-SVM i ERT algoritmima u Msg-Ext eksperimentu su predstavljeni u tabeli 9.11. Dobijeni rezultati pokazuju da Tf-Idf (Unigram), Tf-Idf-Ngram, $n \in [1,2]$, i Tf-Idf-Ngram-Chr, $n \in [1,4]$, kao načinima predstavljanja teksta značajno poboljšavaju performanse modela u poređenju sa BoW (Unigram). Dodatno, Tf-Idf-Ngram težine generalno daju najbolje performanse u svim eksperimentima i korišćenim merama. Pored toga, tradicionalni algoritmi sa Tf-Idf težinama primenjeni u Msg-Ext eksperimentu imaju uporedive metričke vrednosti sa algoritmima DL i čak ih prevazilaze u tačnosti, dok DL daju bolju izbalansiranu tačnost i F_1 rezultat na manjinskoj **Lična** klasi. Algoritam ERT pokazao je nešto niže vrednosti za sve mere u poređenju sa SGD-SVM algoritmom klasifikacije. Iz prikazanih rezultata takođe možemo primetiti da BiLSTM+Att je pokazao bolje performanse

u poređenju sa BiLSTM bez Att. Najvažnije, svi modeli sa pridruženim Meta atributima su pokazali bolje rezultate za najmanje 0.1% u svakom od eksperimenata.

Tabela 9.11: *Poređenje tradicionalnih (SGD-SVM, ERT) i DL algoritama (BiLSTM, BiLSTM+Att) za različite tehnike predstaljanja sadržaja elektronske poruke, sa i bez uključivanja Meta atributa u Msg-Ext eksperimentu*

				Poslovna			Lična		
Algoritam	Atributi	Acc	Acc_{Bal}	Prec	Rec	F_1	Prec	Rec	F_1
ERT	BoW	0.901	0.689	0.914	0.978	0.945	0.739	0.399	0.519
	BoW + Meta	0.907	0.705	0.918	0.980	0.948	0.769	0.429	0.551
	Tf-Idf	0.902	0.695	0.915	0.978	0.946	0.741	0.411	0.529
	Tf-Idf + Meta	0.906	0.705	0.918	0.978	0.947	0.754	0.432	0.550
	Tf-Idf-Ngram	0.901	0.672	0.909	0.984	0.945	0.774	0.360	0.492
	Tf-Idf-Ngram + Meta	0.903	0.689	0.914	0.981	0.946	0.763	0.396	0.522
	Tf-Idf-Ngram-Chr	0.898	0.662	0.906	0.984	0.944	0.769	0.339	0.471
	Tf-Idf-Ngram-Chr + Meta	0.901	0.676	0.910	0.983	0.945	0.769	0.369	0.499
SGD-SVM	BoW	0.900	0.799	0.947	0.937	0.942	0.620	0.660	0.640
	BoW + Meta	0.902	0.802	0.947	0.940	0.943	0.630	0.664	0.646
	Tf-Idf	0.920	0.826	0.953	0.955	0.954	0.707	0.698	0.702
	Tf-Idf + Meta	0.925	0.821	0.951	0.964	0.957	0.744	0.679	0.710
	Tf-Idf-Ngram	**0.928**	0.832	0.954	0.964	**0.959**	0.750	0.701	0.725
	Tf-Idf-Ngram + Meta	**0.929**	**0.838**	0.956	0.963	**0.959**	0.748	0.713	**0.730**
	Tf-Idf-Ngram-Chr	0.923	0.823	0.952	0.960	0.956	0.726	0.685	0.705
	Tf-Idf-Ngram-Chr + Meta	0.922	0.830	0.954	0.956	0.955	0.711	0.704	0.707
BiLSTM	BERT-Embd	0.914	0.811	0.945	0.956	0.950	0.715	0.667	0.690
	BERT-Embd + Meta	0.915	0.818	0.953	0.949	0.951	0.670	0.686	0.678
BiLSTM+Att	BERT-Embd	0.921	0.834	0.959	0.951	0.955	0.676	0.717	0.696
	BERT-Embd + Meta	0.923	0.834	0.957	0.954	0.955	0.703	0.713	0.708

Tabela 9.12: *Eksperimenti nad različitim ML arhitekturama i reprezentacijama teksta sa uključenim Meta atributima*

				Poslovna			Lična		
Eksp.	Alg./Atributi	Acc	Acc_{Bal}	Prec	Rec	F_1	Prec	Rec	F_1
Msg	BiLSTM/Embd	0.923	**0.860**	0.975	0.938	0.956	0.583	0.782	0.668
	SGD-SVM/Tf-Idf	**0.926**	0.817	0.949	0.967	**0.958**	**0.756**	0.667	**0.709**
Msg-Ext	BiLSTM/Embd	0.925	0.834	0.957	0.954	0.955	0.703	0.713	0.708
	SGD-SVM/Tf-Idf	**0.929**	**0.838**	0.956	0.963	**0.959**	0.748	0.713	**0.730**
Brch	BiLSTM/Embd	0.941	0.872	0.966	0.966	0.966	0.778	0.778	0.778
	SGD-SVM/Tf-Idf	**0.957**	**0.903**	0.974	0.977	**0.975**	0.850	0.829	**0.839**
Brch-Ext	BiLSTM/Embd	0.940	0.877	0.970	0.960	0.965	0.739	0.794	0.765
	SGD-SVM/Tf-Idf	**0.958**	**0.913**	0.977	0.975	**0.976**	0.840	0.850	**0.845**
Brch*	BiLSTM/Embd	0.939	0.860	0.955	0.974	0.964	0.838	0.745	0.789
	SGD-SVM/Tf-Idf	**0.952**	**0.902**	0.974	0.971	**0.973**	0.819	0.832	**0.825**

Sledeći nivo eksperimenata obuhvatio je poređenje rezultata primenom različitih ML algoritama i vektorskih reprezentacija sadržaja poruka. Eksperimenti Msg-Ext i Brch-Ext su uključivanjem internet domena adresa primalaca (Ext) u sadržaju prikupili dodatno znanje o svakoj poruci, tako da je ceo sistem neznatno poboljšao tačnost u poređenju redosledno sa eksperimentima Msg i Brch. Eksperiment Brch-Ext je u potpunosti iskoristio povezanost

poruka u svakoj grani (odgovori i domeni adresa primaoca) što je doprinelo poboljšanoj tačnosti za 2.7% i 0.1% u poređenju sa osnovnim eksperimentima Msg i Brch*. Arhitekture nad celokupnom granom poruka Brch je doprinela poboljšanju tačnosti u proseku za +0.7% u odnosu na Msg, ali niže u proseku za -0.2% u odnosu na Brch-Ext eksperiment. Razlog može da leži u očekivanoj kontektualnoj povezanosti između neposrednih poruka u posmatranom skupu podataka. Dodavanje izabranih Meta atributa je u svakom eksperimentu doprinelo povećanju ukupne tačnosti od +0.3% do +1.3%, što ukazuje na značaj dodavanja pravilno odabranih semantičkih i sintaktičkih atributa za poboljšanje performansi klasifikacije na ovom zadatku, kao što je predstavljeno u tabeli 9.12.

Rezultati iz drugih istraživanja

U do sada objavljenim naučnim radovima predložene su tri različite metode kojima se poruke elektronske pošte klasifikuju u kategorije **`Poslovna`** i **`Lična`**. Sve metode koriste različite distribucije ručno obeleženih korpusa *Enron* poruka, koje su proverene pomoću različitih strategija klasifikacije. Rezultati predstavljeni u radu [89] zasnovani su na autorski obeleženom skupu podataka *Enron*, koji se u istraživačkim radovima označava kao skup podataka *Sheffield Enron*. Iako u radu [89] nije nedvosmisleno naznačena struktura poruka, kao i odnos podele podataka u koracima obučavanja i testiranja modela, objavljeni rezultati se mogu iskoristiti za poređenje na najopštenijem nivou.

Tabela 9.13: *Poređenje rezultata dobijenih u predloženoj metodologiji sa najboljim rezultatima objavljenim u naučnim radovima na istom zadatku*

			`Poslovna`			**`Lična`**		
Rad	Eksp.	Acc	Prec	Rec	F_1	Prec	Rec	F_1
[89]		0.930	0.920	0.990	0.950	0.950	0.690	0.800
[8]		0.912	0.967	0.921	0.944	0.735	0.875	0.799
[9]		0.910	0.966	0.929	0.947	0.634	0.793	0.705
	Msg	0.926	0.949	0.967	0.958	0.756	0.667	0.709
	Msg-Ext	0.929	0.956	0.963	0.959	0.748	0.713	0.730
	Brch	0.957	0.974	0.977	0.975	0.850	0.829	0.839
	Brch-Ext	**0.958**	0.977	0.961	**0.976**	0.840	0.850	**0.845**
	Brch*	0.952	0.974	0.971	0.973	0.819	0.832	0.825

Rezultati dobijeni nakon primene modela iz pristupa razvijenog u okviru ovog istraživanja nadmašuju do sada prijavljene rezultate u ukupnoj Prec, Rec i F_1 vrednosti na manjinskoj, **`Lična`** klasi, u Brch, Brch-Ext i Brch* eksperimentima. S druge strane, u našem radu je korišćen obeležen skup podataka predstavljen u [8]. Rezultati dobijeni u Msg osnovnom eksperimentu nadmašuju rezultate objavljene u radovima [8] i [9] u ukupnoj Acc vrednosti (+1.4%/+1.6%). Vrednost F_1 na manjinskoj **`Lična`** klasi u eksperimentima Msg i Msg-Ext je bolja (+0.4% i +2.5% respektivno) od vrednosti predstavljene u [9]. Upoređivanjem drugih mera, dolazi se do zaključka da i one nadmašuju rezultate prikazane u oba ova rada u ukupnoj vrednosti Acc (+4.6%), Rec (+4.9%) i F_1 (+2.9%) na klasi **`Poslovna`**, i F_1 (+6.4%) na klasi **`Lična`**, u eksperimentima Brch, Brch-Ext i Brch* (pogledati tabelu 9.13). Iako u radu nije jednoznačno navedeno da li su autori posmatrali samo poslednju poruku elektronske pošte (Msg) ili celu konverzacionu granu poruka (Brch*), dobijene vrednosti potvrđuju uspešnost predložene metode u oba ova slučaja.

9.2.2 Istinitost glasine i tip delovanja na objavljenu glasinu

Značajni atributi klasifikacije

Na zadacima prepoznavanja istinitosti glasine (**IG**) i tipa delovanja u komentaru na glasinu (**TD**) primenjen je isti pristup za pronalaženje preliminarne liste značajnih atributa, kao i u slučaju **PL** zadatka. Najpre su iz skupa Meta atributa identifikovani multikolinearni atributi primenom hijerarhijskog grupisanja na ρ, sa postavljenim pragom od 0.7, pri čemu je zadržan jedan atribut iz svake grupe korelisanih atributa. Na dobijenoj listi nezavisnih atributa, primenjeni su algoritmi permutacije vrednosti i izbacivanja jednog atributa (pogledati odeljak 4.1.3) da bi se različitim pristupima izmerio uticaj svakog atributa na varijabilnost tačnosti modela. Ova mera je korišćena za izbor konačnog skupa nekolinearnih i najvažnijih Meta atributa u svakom pristupu.

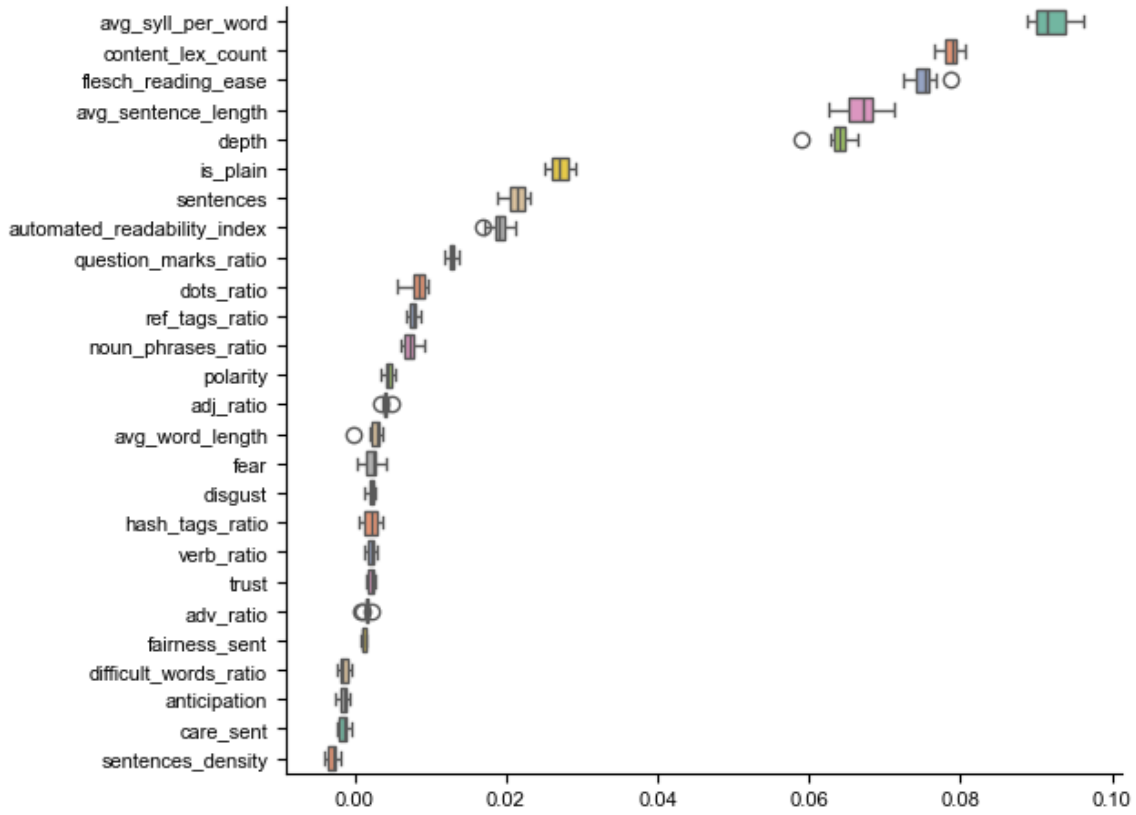
Na zadatku **TD**, poređenjem rezultata metoda isključivanja atributa (slika 9.6) i permutacije vrednosti atributa (slika 9.5), uočavaju se različiti aspekti važnosti pojedinačnih karakteristika. Metoda isključivanja atributa pokazuje da emocionalni i moralni atributi, kao što su *anticipation*, *loyalty_sent*, *trust* i `fairness_p`, imaju visok doprinos u tačnoj klasifikaciji poruka. Pored toga, važnu ulogu imaju i leksičke karakteristike kao što su *noun_phrases_ratio* ili *content_lex_count*. Sa druge strane, metoda permutacije vrednosti atributa jasno ističe da su leksičke i čitljivosti karakteristike dominantne: *ASPW*, *content_lex_count*, *FRES*, *ARI* i *avg_sentence_length* su među najuticajnijim atributima, dok emocionalni i moralni atributi imaju znatno manji značaj. Kombinovanjem rezultata iz obe metode, uz uključivanje statističke provere značaja atributa pristupom jedne promenljive (prilog A, tabela A.2) kao što je prikazano u fragmentu koda 9.2.1, formirana je konačna lista značajnih atributa za klasifikaciju poruka na zadatku **TD**:

```
značajni_atributi = [
    "noun_phrases_ratio",
    "content_lex_count",
    "ASPW",
    "FRES",
    "avg_sentence_length",
    "anticipation",
    "trust",
    "loyalty_sent",
    "fairness_p",
    "ARI",
    "polarity",
    "joy",
    "adj_ratio"
]
```

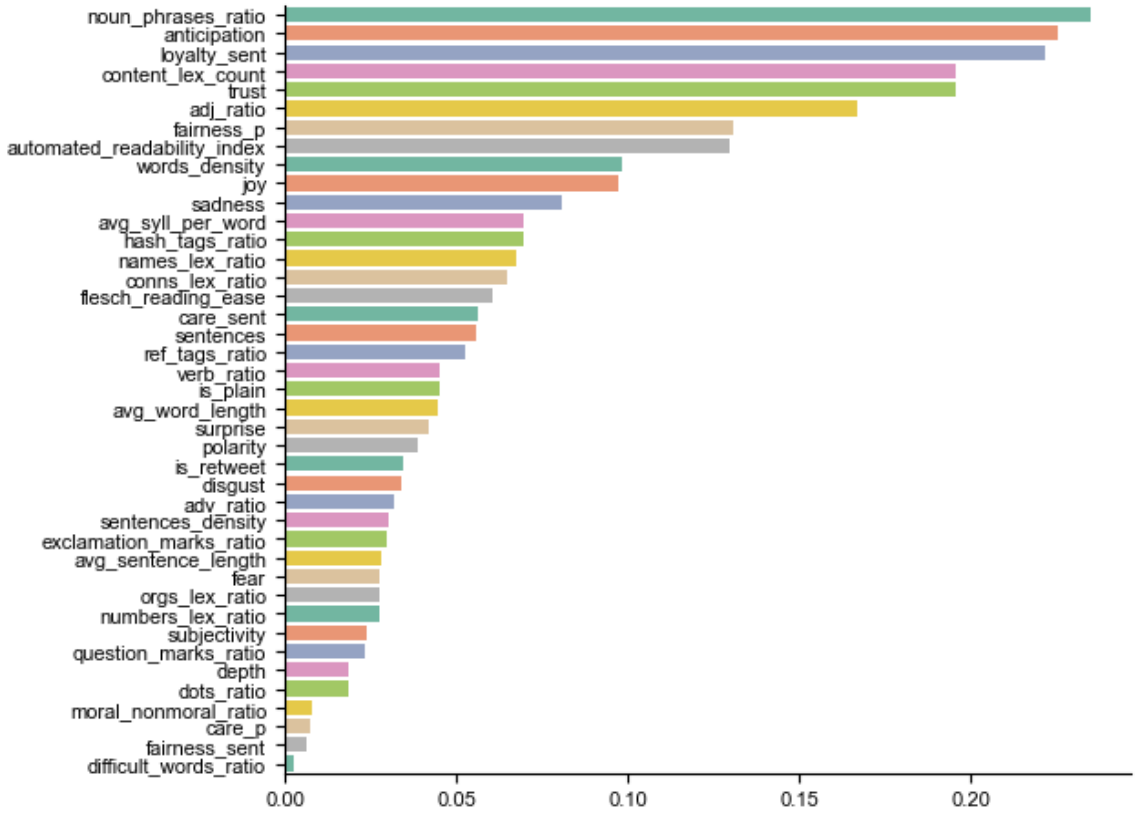
Fragment koda 9.3: *Lista značajnih Meta atributa u klasifikaciji poruka sa društvenih mreža prema tipu delovanja na objavljenu glasinu (**TD**)*

Na zadatku **IG**, poređenjem rezultata metoda isključivanja atributa (slika 9.8) i permutacije vrednosti atributa (slika 9.7), uočavaju se različiti aspekti važnosti pojedinačnih karakteristika. Na osnovu rezultata iz dve metode ocenjivanja značaja atributa identifikovani su atributi koji konzistentno doprinose klasifikacionom zadatku u oba slučaja. Ovi atributi se mogu smatrati najpouzdanijim indikatorima karakteristika koje definišu izražavanje u porukama na društvenim mrežama. Oba pristupa identifikuju sledeće atribute:

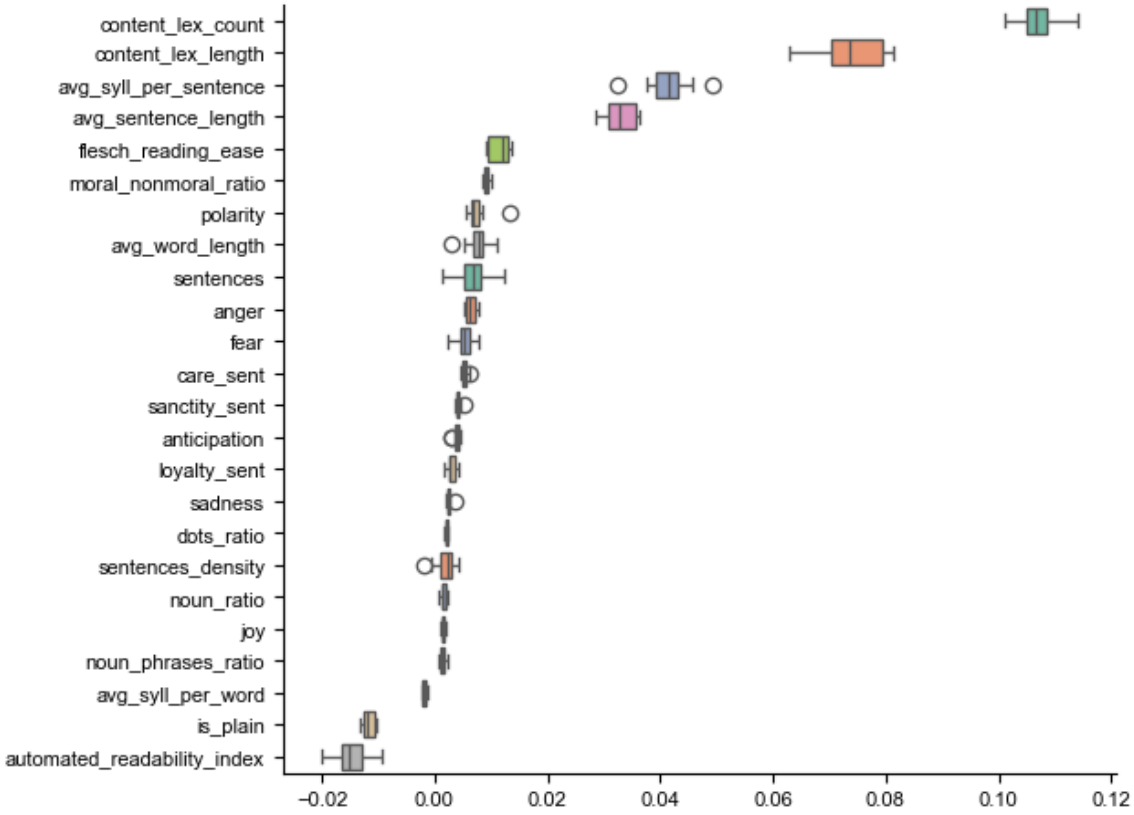
- *content_lex_count*, *content_lex_length*, *avg_word_length*, *avg_sentence_length*, *ASPS* (LexAtr),
- *anger*, *joy*, *fear* i *anticipation* (EmoAtr),



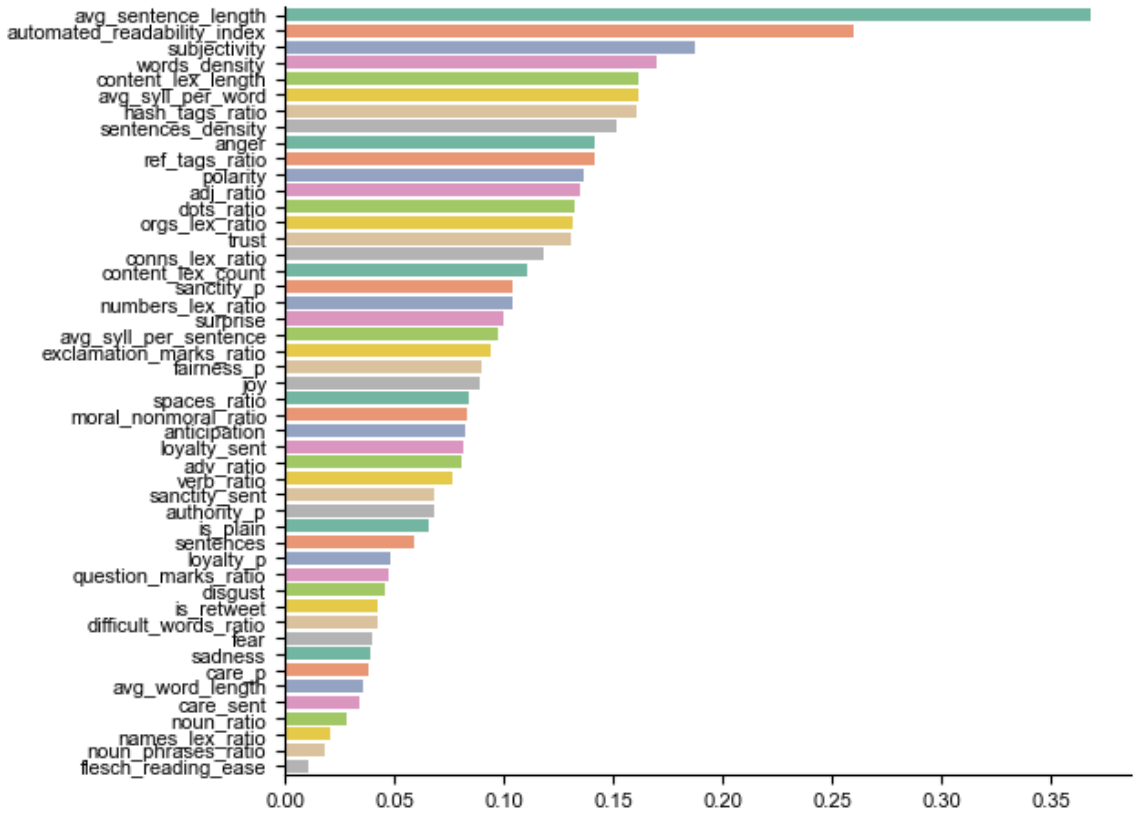
Slika 9.5: *Stepen značaja atributa na zadatku* **IG** *izračunat metodom permutacije vrednosti atributa*



Slika 9.6: *Stepen značaja atributa na zadatku* **IG** *izračunat metodom isključivanja jednog atributa*



Slika 9.7: *Stepen značaja atributa na zadatku* **IG** *izračunat metodom permutacije vrednosti atributa*



Slika 9.8: *Stepen značaja atributa na zadatku* **IG** *izračunat metodom isključivanja jednog atributa*

- *sanctity_sent*, *care_sent*, *loyalty_sent* i *moral_nonmoral_ratio* (MorAtr),
- *subjectivity*, *polarity* (ExpAtr).

Kombinovanjem rezultata iz obe metode, uz uključivanje statističke provere značaja atributa pristupom jedne promenljive (prilog A, tabela A.2) kao što je prikazano u fragmentu koda 9.2.1, formirana je konačna lista značajnih atributa za klasifikaciju poruka na zadatku **IG**:

```
značajni_atributi = [
    "avg_word_length",
    "avg_sentence_length",
    "content_lex_count",
    "content_lex_length",
    "ASPS",
    "anger",
    "fear",
    "joy",
    "anticipation",
    "moral_nonmoral_ratio",
    "loyalty_sent",
    "care_sent",
    "sanctity_sent",
    "polarity",
    "subjectivity"
]
```

Fragment koda 9.4: *Lista značajnih Meta atributa u klasifikaciji objava na društvenim mrežama prema svojoj istinitosti (**IG**)*

Na osnovu preliminarne liste najznačajnijih atributa može se zaključiti da su za glasine na društvenim mrežama karakteristične sledeće osobine:

- Stil i složenost izraza poruke – što ukazuje da su glasine stilski jednostavnije ili, nasuprot tome složenije u cilju postizanja uverljivosti, autoritativnosti i pouzdanosti kod čitalaca.
- Sentimentalni, emocionalni i moralni ton poruke – glasine su emocionalno obojene i pozivaju se na vrednosti koje ciljna publika podržava što može podstaknuti emocionalne reakcije i uticati na veći stepen deljenja sadržaja.

Rezultati iz drugih istraživanja

Najuspešniji pristupi u okviru *SemVal2019* takmičenja na zadacima klasifikacije tipa delovanja na objavu (zadatak **A.**, **TD**) i istinitosti glasine (zadatak **B.**, **IG**) oslanjali su se na kombinaciju bogatih jezičkih reprezentacija poruka (BERT) i modelovanja strukture konverzacije. Sekvencijalni pristup za obradu grane konverzacije podrazumeva da se svaki komentar unutar grane posmatra kao deo niza neposrednih i međusobno zavisnih poruka kako bi se poboljšalo prepoznavanje stava prema glasinama i rešio problem nebalansiranosti klasa. Ovaj pristup je ujedno korišćen kao jedno od predloženih pristupa za rešavanje ovih zadataka od strane autora. Kao prvo predloženo rešenje, korišćen je *branchLSTM* model koji obrađuje svaku granu konverzacije pomoću LSTM mreže, koje se može prilagoditi za klasifikaciju pojedinačnih poruka ili čitavog niza poruka. Drugi osnovni model je model pod nazivom *NileTMRG* koji koristi linearni SVM model nad BoW reprezentacijama poruka uz uključivanje leksičkih i strukturalnih karakteristika konverzacije [53]. Poslednje predloženo rešenje je zasnovano na prisupu većinskog glasanja u kome se svakom primeru iz skupa podataka dodeljuje najčešća, odnosno većinska klasa.

Najbolje prijavljena takmičarska rešenja na zadatku prepoznavanja tipa delovanja na objavu (**TD**) koristila su BERT-Embd i GPT-Embd za reprezentaciju sadržaja svake poruke [54, 218]. Ova rešenja (*BLCU NLP*, *BUT-FIT*) dostigla su F_1^{Ma} vrednost od 61.87%, odnosno 60.67%, na zadatku klasifikacije tipa delovanja na objavu (pogledati tabelu 9.14). Druga rešenja zasnivala su se na DNN arhitekturama nad granama konverzacije, pri čemu su koristila i tekstualne i strukturne informacije o konverzaciji, uključujući Meta atribute kao što su afektivne informacije [147]. Neka od predloženih rešenja su koristila BERT-Embd za predstavljanje sadržaja poruka, koje su dodatno obogaćene kontekstom prethodnih poruka unutar grane, čime su omogućili eksplicitno modelovanje dijaloga [116].

Tabela 9.14: *Rezultati osnovnih i najuspešnijih rešenja za* **IG** *i* **TD** *zadatke u okviru SemVal2019 takmičenja*

Zadatak	**IG**		**TD**
Takmičarsko rešenje	F_1^{Ma}	RMSE	F_1^{Ma}
eventAI [114]	**0.5765**	**0.6078**	0.5776
BUT-FIT [54]	-	-	**0.6067**
BLCU NLP [54]	0.2525	0.8179	**0.6187**
Osnovno rešenje			
branchLSTM [104]	0.3364	0.7806	0.4920
NileTMRG [53]	0.3089	0.7698	-
Većinsko glasanje [65]	0.2241	0.7115	0.2234

Na zadatku prepoznavanja istinitosti glasina (**IG**), najuspešnije takmičarsko rešenje (*eventAI*) integriše tri vrste podataka: sadržaj poruke, pouzdanost korisnika i obrasce propagiranja informacija kroz konverzaciju [114]. U predloženom rešenju sadržaj poruke se analizira korišćenjem leksičkih i semantičkih atributa, dok se pouzdanost korisnika procenjuje na osnovu njegovih karakteristika i prethodne aktivnosti. Propagacija informacija se modeluje kroz strukture diskusija u obliku stabala koja prikazuju širenje informacija. Kombinovanjem ovih atributa i klasifikacionih modela, sistem je postigao konkurentne rezultate na zadatku klasifikacije tačnosti glasina (istinita, lažna, nepotvrđena), što pokazuje efikasnost višedimenzionalnog pristupa u kontekstu automatske analize dezinformacija na društveni mrežama. Dobijeni rezultati, takođe, ukazuju da je za uspešnu klasifikaciju konverzacionih poruka u pojedinim klasifikacionim zadacima (kao što su **TD** i **IG**) važno uzeti u obzir međusobnu zavisnost koja postoji između poruka, kao i strukturne karakteristike sadržaja poruka i čitave konverzacije.

Rezultati iz razvijene metodologije

Rezultati dobijeni primenom metoda razvijenih u okviru ovog istraživanja pokazuju da Brch BiLSTM arhitektura ostvaruje bolje performanse u odnosu na Msg BiLSTM arhitekturu na oba zadatka, što je u skladu sa očekivanjima, jer model koji obrađuje poruke u kontekstu celih grana diskusija mogu bolje da prepoznaju kontekstualne odnose između poruka i na taj način pomognu u rešavanju postavljenog klasifikacionog zadatka. Vrednosti u Acc, $Prec^{Ma}$, Rec^{Ma} i F_1^{Ma} mera su pokazala konzistentno povećanje između ovih eksperimenata, pri čemu su najveća poboljšanja primećena kod eksperimenata sa pridruženim Meta, EmoAtr i MorAtr atributima. Uključivanje mešovitih pridruženih Meta atributa (leksičkih, konverzacionih, emocionalnih, moralnih) značajno poboljšava performanse klasifikacije, pri čemu EmoAtr i MorAtr atributi uzimaju značajno učešće u ovom skupu atributa. Dodatno, rezultati iz posebnih eksperimenata izvršenih uključivanjem pojedinačno EmoAtr, MorAtr, kao i njihove unije, ukazuju da analiza komunikacije na društvenim mrežama

kroz prizmu emocionalni i moralnih aspekata jezika može pomoći u boljem konverzacije i na taj način poboljšati tačnost prilikom automatske klasifikacije ovih podataka. Najbolji rezultati na zadacima **TD** i **IG** postignuti su korišćenjem BiLSTM arhitektura sa uključenim mehanizmom pažnje i Meta atributima, čime se potvrđuje značaj uključivanja mehanizma pažnje u pronalaženju značajnih informacija iz konteksta poruke.

Tabela 9.15: *Eksperimenti sa Msg i Brch arhitekturama sa i bez uključenih EmoAtr, MorAtr i Meta atributa na zadatku* **TD**

Eksp.	Algoritam	Atributi	Acc	$Prec^{Ma}$	Rec^{Ma}	F_1^{Ma}
Msg	BiLSTM	Embd	0.749	0.510	0.430	0.443
	BiLSTM+Att	Embd	0.710	0.504	0.501	0.491
	BiLSTM	Embd+Meta	0.747	0.727	0.465	0.470
	BiLSTM+Att	Embd+Meta	0.753	0.617	0.505	0.519
		Embd+EmoAtr	0.751	0.548	0.512	0.509
		Embd+MorAtr	0.749	0.572	0.473	0.494
		Embd+EmoAtr+MorAtr	0.735	0.559	0.520	0.524
Brch	BiLSTM	Embd	0.760	0.530	0.450	0.460
	BiLSTM+Att	Embd	0.727	0.525	0.520	0.510
	BiLSTM	Embd+Meta	0.765	0.745	0.480	0.485
	BiLSTM+Att	Embd+Meta	0.770	0.635	0.520	0.535
		Embd+EmoAtr	0.765	0.565	0.530	0.525
		Embd+MorAtr	0.762	0.590	0.490	0.510
		Embd+EmoAtr+MorAtr	0.760	0.575	0.540	0.545

Tabela 9.16: *Eksperimenti nad Brch i Brch-Ext arhitekturama sa i bez uključenih EmoAtr, MorAtr i Meta atributa na zadatku* **IG**

Eksp.	Algoritam	Atributi	Acc	$Prec^{Ma}$	Rec^{Ma}	F_1^{Ma}
Brch	BiLSTM	Embd	0.490	0.400	0.370	0.405
	BiLSTM+Att	Embd	0.505	0.412	0.385	0.417
	BiLSTM	Embd+Meta	0.518	0.431	0.397	0.430
	BiLSTM+Att	Embd+Meta	0.540	0.460	0.420	0.455
		Embd+EmoAtr	0.525	0.439	0.405	0.438
		Embd+MorAtr	0.520	0.432	0.398	0.428
		Embd+EmoAtr+MorAtr	0.532	0.445	0.415	0.445
Brch-Ext	BiLSTM	Embd	0.515	0.425	0.395	0.432
	BiLSTM+Att	Embd	0.528	0.440	0.410	0.448
	BiLSTM	Embd+Meta	0.550	0.460	0.425	0.460
	BiLSTM+Att	Embd+Meta	0.570	0.490	0.455	0.485
		Embd+EmoAtr	0.555	0.470	0.435	0.470
		Embd+MorAtr	0.552	0.465	0.430	0.465
		Embd+EmoAtr+MorAtr	0.560	0.475	0.440	0.475

Dobijeni rezultati potvrđuju da uključivanje dodatnih atributa (Meta, EmoAtr i MorAtr) poboljšava performanse modela u zadatku klasifikacije istinitosti glasina (**IG**). Uočava se da Brch-Ext arhitektura dosledno nadmašuje osnovni Brch pristup, što ukazuje da proširena reprezentacija sadržaja (prepoznati tip delovanja) doprinose boljem razumevanju konteksta glasina. Uvođenje Meta atributa, kao i u slučaju zadatka **TD**, daje primetno poboljšanje u svim merama u odnosu na modele bez ovih atributa. Primetno, uključivanjem isključivo EmoAtr i MorAtr atributa, pojedinačno ili u kombinaciji, unapređuju se performase modela, ali u nešto nižem opsegu u odnosu na uključivanje mešovitog skupa značajnih Meta atributa (F_1^{Ma} od 48.5% za Brch-Ext+Att+Meta). Ipak, ni najbolji eksperiment ne nadmašuje *eventAI* rešenje, koje je postiglo vrednost od 57.65% F_1^{Ma}, što sugeriše

da su potrebni dodatni eksperimenti za postizanje optimalnih performansi. U poređenju sa osnovnim rešenjima (*branchLSTM*, *NileTMRG*, većinsko glasanje), rezultati iz eksperimenata zasnovanih na predloženoj metodologiji u okviru ovog istraživanja ih nadmašuju, što potvrđuje primeljivost razvijene metodologije na rešavanje **TD** i **IG** zadataka klasifikacije konverzacionih poruka sa društvenih mreža.

9.3. Emocionalni i moralni atributi kao zavisne promenljive

9.3.1 Predviđanje emocionalnog afekta

Evaluacija izgrađenih modela za prepoznavanje emocionalnog afekta (pogledati odeljak 8.7) je izvršena korišćenjem standardnih metrika kao što su F_1^{Ma}, F_1^{Mi}, F_1^{w}, Acc, HS i HL kako bismo uporedili performanse modela između različitih eksperimenata. Korišćenje različitih varijanti F_1 metrika za ocenjivanje rezultata klasifikacije je značajno naročito u slučajevima nebalansiranih skupova podataka, kao i u kontekstu složenosti višeznačne klasifikacije emocija [80]. Prema svojoj definiciji F_1^{Mi} metrika je osetljivija na klase sa većim brojem instanci, dok F_1^{Ma} pruža uravnotežen pregled svih klasa, ali može biti podložna uticaju klasa sa malim brojem instanci. Sa druge strane, metrika F_1^{w} kompenzuje neravnotežu tako što klasi dodeljuje težinu prema njenoj učestalosti. U okviru literature, rezultati višeznačne klasifikacije su često neprecizno iskazani preko standardne F_1 mere, bez preciznog naglašavanja o kojoj se tačno varijanti ove mere radi. Vrednosti F_1 mere na istom zadatku u do sada objavljnenim radovima variraju u zavisnosti od broja klasa, vrste klasifikacije (binarna, višeklasna, višeznačna) i balansiranosti kategorija u skupu za obučavanje (pogledati tabelu 9.17). U literaturi, takođe, često nije naglašeno koja tehnika podele podataka na skupove za obuku, proveru i testiranje je korišćena, što je od važnosti kod nebalansiranih skupova podataka i može značajno uticati na ukupne performanse klasifikacije.

Tabela 9.17: *Poređenje performansi klasifikacije emocija u različitim istraživanjima [145]*

Rad	Podaci	#Emocija	Višeznačna	F_1^{Ma}	Acc	Balans.
[1]	Tviter	8	Ne	-	0.95	Ne
[1]	Tviter	24	Ne	0.87	-	Ne
[173]	Tviter	11 + *neutral*	Da	0.64	0.53	Ne
[220]	Tviter	11	Da	0.44	0.46	Ne
[119]	Knjige	8 + *neutral*	Ne	0.60	-	Da (*-disgust;-surprise*)
[49]	Redit	27	Da	0.46	-	Ne
[49]	Redit	6	Ne	0.64	-	Ne
[145]	Titlovi	8 + *neutral*	Da	0.54	0.54	Da

Analiza performansi višeznačne klasifikacije na podkorpusu **Twitter-Emo.SR**, prikazana u tabeli 9.18, pokazuje razlike između osnovnih modela zasnovanih na BERT arhitekturama i različitih pristupa procesiranja podataka: originalnog teksta (*Orig*), maskiranja korisničkih referenci, heš oznaka i internet adresa (*Masked*), kao i kombinacije maskiranja sa korišćenjem strogo emocionalnih obeležja (*Masked+Emo*). Rezultati pokazuju da veći modeli, poput Twitter-XLM-R_{large} i XLM-R_{large}, pokazuju bolje ukupne performanse, pri čemu *Masked+Emo* pristup prikazuje najbolje vrednosti za F_1^{Ma} (54%) i F_1^{w} (62%), kao i relativno visoku vrednost za Acc (41%). Poređenja među modelima pokazuju da primena maskiranih tokena sa korišćenjem strogo emocionalnih obeležja poboljšava F_1 rezultate u odnosu na nemodifikovane tekstualne sekvence ili one kod kojih su samo maskirani tokeni. Ostali modeli, kao što su Twitter-XLM-R_{base}, Jerteh-355 i BERTić, takođe ostvaruju bolje rezultate pri upotrebi modifikovanih sekvenci i izboru kategorija, ali sa nešto nižim

Tabela 9.18: *Rezultati doobučavanja modela zasnovanih na BERT arhitekturama u zavisnosti od različitih pristupa procesiranja ulaznih podataka primenjenih na korpusu* **Twitter-Emo.SR**

Osnovni model	Korpus	F_1^{Ma}	F_1^{Mi}	F_1^{w}	Acc	HS	HL
XLM-R_{base}	*Orig*	0.38	0.59	0.57	0.40	0.54	0.12
	Masked	0.44	0.59	0.57	0.38	0.56	0.12
	Masked+Emo	0.45	0.59	0.56	0.39	0.55	0.13
XLM-R_{large}	*Orig*	0.50	0.59	0.56	0.40	0.56	0.13
	Masked	0.51	0.61	0.60	0.39	0.57	0.14
	Masked+Emo	**0.53**	0.62	0.61	0.40	0.57	0.13
Twitter-XLM-R_{base}	*Orig*	0.40	0.60	0.61	0.39	0.55	0.14
	Masked	0.42	0.61	0.58	0.40	0.57	0.14
	Masked+Emo	0.49	0.61	0.59	0.39	0.57	0.13
Twitter-XLM-R_{large}	*Orig*	0.47	0.61	0.59	0.39	0.57	0.13
	Masked	0.49	0.60	0.58	0.39	0.56	0.14
	Masked+Emo	**0.54**	0.62	0.62	0.41	0.58	0.13
Jerteh-81	*Orig*	0.37	0.57	0.55	0.36	0.52	0.13
	Masked	0.42	0.58	0.56	0.37	0.51	0.14
	Masked+Emo	0.44	0.58	0.56	0.35	0.53	0.13
Jerteh-355	*Orig*	0.46	0.58	0.56	0.38	0.55	0.13
	Masked	0.47	0.59	0.58	0.39	0.56	0.14
	Masked+Emo	0.49	0.60	0.59	0.37	0.55	0.14
BERTić	*Orig*	0.44	0.59	0.58	0.36	0.55	0.14
	Masked	0.45	0.60	0.59	0.39	0.56	0.12
	Masked+Emo	0.49	0.60	0.59	0.37	0.56	0.14

vrednostima u poređenju sa XLM-R_{large}. Iz ovih eksperimenata se može zaključiti da obrada ulaznog teksta, koja uključuje uklanjanje nebitnih informacija uz korišćenje strogo emocionalnih obeležja, doprinosi tačnosti i doslednosti modela u višeznačnoj klasifikaciji emocija na podkorpusu **Twitter-Emo.SR**. Poboljšanje performansi pri korišćenju *Masked* sekvenci u odnosu na *Orig* tekstualne sekvence može se pripisati nekoliko faktora. Maskiranje korisničkih referenci (*@user*), heš oznaka (*#hash*) i url adresa (*http*) pomaže modelu da se fokusira na ključne delove teksta koji nose semantičku i emocionalnu informaciju, umesto na manje relevantne ili ometajuće elemente. Dodatno, razlike se mogu pripisati i tipu poruke, kao što su inicijalne objave naspram komentara koje su u korpusu zastupljenije. Inicijalne objave na društvenim mrežama često imaju jasniji kontekst i direktniji sadržaj, dok komentari mogu biti kraći, neformalniji sa većim brojem korisničkih referenci koje mogu ometati model u pravilnom prepoznavanju emocionalnog tona. Uvođenjem *Emo* skupa, gde su zadržane samo striktno emocionalne kategorije, a uklonjena manje zastupljena kategorija *neutral*, dolazi do povećanja F_1^{Ma} vrednosti jer se model bolje prilagođava na preostale emocionalne kategorije koje su sada ravnomernije zastupljene. Drugi razlog može da leži u manjoj preciznosti na ovoj kategoriji što može biti uslovljeno nedovoljnim brojem odgovarajućih primera. Ukupne performanse modela, takođe mogu biti umanjene zbog nedovoljne zastupljenosti kategorija *digust* i *fear*, što je naročito primetno na podkorpusu **Reddit-Emo.SR**. Za sledeći eksperiment merenja uspešnosti klasifikacije emocionalnog afekta na različitim delovima korpusa **Social-Emo.SR**, koristiće se *Masked+Emo* pristup obrade ulaznih podataka koji je na ovom eksperimentu pokazao najbolje rezultate.

Rezultati poređenja modela prikazani u tabeli 9.19 pokazuju da veći i napredniji modeli, poput XLM-R_{large}, dosledno pružaju bolje performanse u svim metrikama u poređenju sa manjim verzijama kao što je XLM-R_{base}. Ovi modeli postižu visoke vrednosti F_1^{Ma}, F_1^{Mi}, i F_1^{w}, posebno na podkorpusu **Twitter-Emo.SR**, što ukazuje na to da veći modeli mogu efi-

Tabela 9.19: *Rezultati doobučavanja modela zasnovanih na BERT i LLaMA arhitekturama izvršenih na celokupnom korpusu* **Social-Emo.SR** *i njegovim podkorpusima*

Osnovni model	Korpus	F_1^{Ma}	F_1^{Mi}	F_1^{w}	Acc	HS	HL
XLM-R_{base}	**Twitter-Emo.SR**	0.45	0.59	0.56	0.39	0.55	0.13
	Reddit-Emo.SR	0.42	0.57	0.56	0.35	0.51	0.14
	Social-Emo.SR	0.47	0.59	0.57	0.36	0.54	0.14
XLM-R_{large}	**Twitter-Emo.SR**	0.53	0.62	0.61	0.40	0.57	0.13
	Reddit-Emo.SR	0.47	0.60	0.59	0.38	0.54	0.14
	Social-Emo.SR	0.51	0.62	0.61	0.38	0.56	0.13
Twitter-XLM-R_{base}	**Twitter-Emo.SR**	0.49	0.61	0.59	0.39	0.57	0.13
	Reddit-Emo.SR	0.44	0.59	0.58	0.39	0.54	0.13
	Social-Emo.SR	0.48	0.61	0.59	0.38	0.56	0.13
Twitter-XLM-R_{large}	**Twitter-Emo.SR**	0.54	0.62	0.62	0.39	0.58	0.13
	Reddit-Emo.SR	0.48	0.61	0.60	0.38	0.55	0.13
	Social-Emo.SR	0.52	0.61	0.60	0.35	0.55	0.14
Jerteh-81	**Twitter-Emo.SR**	0.44	0.58	0.56	0.35	0.53	0.13
	Reddit-Emo.SR	0.40	0.55	0.54	0.34	0.49	0.14
	Social-Emo.SR	0.45	0.58	0.56	0.35	0.52	0.14
Jerteh-355	**Twitter-Emo.SR**	0.49	0.60	0.59	0.37	0.55	0.14
	Reddit-Emo.SR	0.46	0.60	0.58	0.37	0.54	0.14
	Social-Emo.SR	0.46	0.58	0.57	0.36	0.53	0.14
BERTić	**Twitter-Emo.SR**	0.49	0.60	0.59	0.37	0.56	0.14
	Reddit-Emo.SR	0.43	0.59	0.58	0.37	0.54	0.14
	Social-Emo.SR	0.47	0.60	0.58	0.36	0.55	0.14
LLaMA-3.2-3B-Instruct	**Twitter-Emo.SR**	0.14	0.15	0.14	0.04	0.08	0.21
	Reddit-Emo.SR	0.14	0.17	0.15	0.05	0.11	0.22
	Social-Emo.SR	0.16	0.18	0.16	0.06	0.12	0.22
LLaMA-3.2-3B-Instruct doobučavanje	**Twitter-Emo.SR**	0.43	0.56	0.54	0.16	0.46	0.18
	Reddit-Emo.SR	0.43	0.55	0.55	0.12	0.44	0.19
	Social-Emo.SR	0.44	0.55	0.54	0.13	0.44	0.19

kasnije prepoznati i klasifikovati emocije. Specijalizovani domenski modeli, poput Twitter-XLM-R, pokazuju poboljšane rezultate na domenskim korpusima, što ukazuje na prednost prilagođavanja modela određenoj vrsti podataka. Na primer, Twitter-XLM-R_{large} postiže najvišu F_1^{Ma} vrednost od 54% na podkorpusu **Twitter-Emo.SR**, dok su njegove performanse na drugim korpusima nešto niže, ali i dalje konkurentne sa drugim modelima. S druge strane, modeli kao što su Jerteh-81 i Jerteh-355, iako izrađeni isključivo nad srpskim književnim korpusima podataka, pokazuju stabilne, ali generalno niže performanse u poređenju sa većim XLM-R modelima, što ukazuje na potrebu za dodatnom optimizacijom za postizanje veće preciznosti nad podacima sa društvenih mreža. Model BERTić je pokazao rezultat od 49% F_1^{Ma} na podkorpusu **Twitter-Emo.SR**, što je u rangu sa performansama drugih modela srednjeg nivoa performansi. Generalno, svi modeli su pokazali bolje rezultate na podkorpusu **Twitter-Emo.SR** u poređenju sa drugim korpusima, što ukazuje na bolju pogodnost tog korpusa za klasifikaciju emocija.

Rezultati, takođe, pokazuju da je doobučavanje *LLaMa-3.2-3B-Instruct* modela značajno poboljšalo performanse klasifikacije emocionalnog afekta u srpskom jeziku, pri čemu je F_1^{Mi} porastao sa oko 14–16% na 43–44% na svim korpusima. Ipak, uprkos poboljšanju, rezultati su i dalje relativno niski, što ukazuje na potencijalne izazove, poput nedovoljne količine podataka za doobučavanje, nebalansiranosti između klasa i mogućih ograničenja modela u razumevanju karakteristika srpskog jezika. Doslednost poboljšanja na različitim korpusima sugeriše da model nije previše prilagođen samo jednom domenu, ali da su potrebne dodatne optimizacije modela, kao što su balansiranje klasa, fino po-

dešavanje hiperparametara, obogaćivanje tekstualnih reprezentacija dodatnim semantičkim resursima ili korišćenje osnovnog LLaMA modela sa većim brojem parametara. Nešto niži rezultati u odnosu na osnovne modele BERT arhitekture, ukazuju na njihovu bolju prilagodljivost zadacima klasifikacije teksta i srpskom jeziku.

Tabela 9.20: *Vrednost F_1 mere prikazana po kategoriji emocija na skupu za testiranje korišćenjem Twitter-XLM-R_{large} modela*

Kategorija	`Twitter-Emo.SR`	`Reddit-Emo.SR`	`Social-Emo.SR`
anger	0.79	0.68	0.73
anticipation	0.46	0.40	0.47
disgust	0.32	0.23	0.28
fear	0.32	0.22	0.28
joy	0.69	0.70	0.70
sadness	0.65	0.55	0.60
surprise	0.48	0.49	0.49
trust	0.61	0.66	0.65
F_1^{Ma}	0.54	0.50	0.53
F_1^{Mi}	0.63	0.60	0.61
F_1^{w}	0.62	0.59	0.61

Na osnovu analize performansi modela možemo zaključiti nekoliko važnih aspekata o korpusima, anotacijama i distribuciji kategorija. Najpre, modeli izgrađeni nad podkorpusima **`Twitter-Emo.SR`** i **`Reddit-Emo.SR`** pokazuju različite performanse, što ukazuje na specifične razlike u strukturi i karakteristikama teksta između ovih platformi. Na podkorpusu **`Twitter-Emo.SR`** uočljive su nešto bolje performanse modela, što može ukazivati da kraće, konciznije poruke sa Tviter platforme, na kojoj se emocije jasnije izražavaju, mogu biti lakše za obradu i klasifikaciju. Ove razlike mogu ukazivati i na razlike u kvalitetu i konzistentnosti obeležja, jer je preciznost i objektivnost prilikom kategorizacije emocija od izuzetne važnosti. Stabilne, ali nešto niže performanse na celom korpusu **`Social-Emo.SR`** ukazuju na širu raspodelu emocionalnih obeležja i heterogenost anotacija, što može otežati postizanje konzistentnih performansi. Značajne varijacije u performansama modela u F_1^{Ma} vrednosti mogu ukazivati na to da modeli različito prepoznaju ređe i češće emocionalne kategorije. Ovo može biti rezultat nebalansirane raspodele kategorija, u kojoj dominantne emocije preovlađuju, a ređe kategorije, kao što su *disgust* i *fear*, ostaju manje precizno klasifikovane (pogledati tabelu 9.20). Ove varijacije takođe sugerišu da **`Reddit-Emo.SR`**, sa generalno nižim performansama na svim modelima, može sadržati kontekstualno složene emocije koje predstavljaju izazov za preciznu klasifikaciju na ovom podkorpusu.

9.3.2 Predviđanje moralne vrednosti

Eksperimenti za prepoznavanje iskazane moralnosti u tekstu su izvršeni na zadacima prepoznavanja osnovnih moralnih vrednosti i moralnog sentimenta. U oba slučaja su korišćeni isti osnovni modeli i mere za evaluaciju kao i u eksperimentima za prepoznavanje emocionalnog afekta. Vrednosti F_1 mere na istom zadatku u do sada objavljenim radovima variraju u zavisnosti od podataka (domena), algoritma, broja klasa, vrste klasifikacije i balansiranosti kategorija u skupu za obučavanje (pogledati tabelu 9.21).

Na zadatku prepoznavanja osnovnih moralnih vrednosti i moralnog sentimenta, sa rezultatim prikazanim u tabeli 9.22 i tabeli 9.23, u tom redosledu, uočljive su značajne razlike u performansama među modelima i korpusima koji su korišćeni za njihovo obučavanje. Veći modeli, kao što je XLM-R_{large} pokazuju bolje rezultate u poređenju sa manjim modeli-

Tabela 9.21: *Poređenje performansi klasifikacije moralnih vrednosti u različitim istraživanjima*

Rad	Podaci	#Klasa	Višeznačna	F_1	Balans.
[160]	MFTC, MFRC, Facebook	10 (12)	Ne	0.11-0.32	Ne
[32]	MFTC	11	Ne	0.47-0.77	Da
[31]	MFRC	6	Da	0.65-0.75	Ne

Tabela 9.22: *Rezultati doobučavanja modela zasnovanih na BERT i LLaMA arhitekturama izvršenih na celokupnom korpusu* **`Social-Mor.SR`** *i njegovim podkorpusima na zadatku prepoznavanja osnovnih moralnih vrednosti*

Osnovni model	Korpus	F_1^{Ma}	F_1^{Mi}	F_1^{w}	Acc	HS	HL
XLM-R_{base}	**`Twitter-Mor.SR`**	0.35	0.65	0.58	0.40	0.60	0.19
	`Reddit-Mor.SR`	0.31	0.70	0.61	0.42	0.64	0.17
	`Social-Mor.SR`	0.33	0.67	0.59	0.41	0.61	0.18
XLM-R_{large}	**`Twitter-Mor.SR`**	0.49	0.66	0.64	0.39	0.60	0.19
	`Reddit-Mor.SR`	0.36	0.67	0.63	0.40	0.62	0.19
	`Social-Mor.SR`	0.46	0.68	0.65	0.40	0.62	0.19
Twitter-XLM-R_{base}	**`Twitter-Mor.SR`**	0.42	0.65	0.60	0.40	0.59	0.19
	`Reddit-Mor.SR`	0.33	0.69	0.61	0.44	0.64	0.17
	`Social-Mor.SR`	0.35	0.66	0.59	0.40	0.61	0.18
Twitter-XLM-R_{large}	**`Twitter-Mor.SR`**	0.53	0.67	0.66	0.35	0.60	0.20
	`Reddit-Mor.SR`	0.44	0.68	0.66	0.38	0.62	0.19
	`Social-Mor.SR`	0.46	0.68	0.65	0.40	0.62	0.19
Jerteh-81	**`Twitter-Mor.SR`**	0.38	0.63	0.58	0.38	0.58	0.20
	`Reddit-Mor.SR`	0.33	0.69	0.62	0.43	0.64	0.17
	`Social-Mor.SR`	0.37	0.66	0.60	0.42	0.61	0.18
Jerteh-355	**`Twitter-Mor.SR`**	0.47	0.66	0.63	0.34	0.59	0.20
	`Reddit-Mor.SR`	0.36	0.69	0.63	0.44	0.64	0.17
	`Social-Mor.SR`	0.42	0.68	0.64	0.40	0.62	0.19
BERTić	**`Twitter-Mor.SR`**	0.30	0.62	0.54	0.40	0.57	0.20
	`Reddit-Mor.SR`	0.28	0.68	0.59	0.43	0.64	0.17
	`Social-Mor.SR`	0.27	0.66	0.55	0.43	0.62	0.18
LLaMA-3.2-3B-Instruct	**`Twitter-Emo.SR`**	0.13	0.22	0.18	0.03	0.14	0.44
	`Reddit-Emo.SR`	0.13	0.20	0.20	0.02	0.12	0.45
	`Social-Emo.SR`	0.13	0.21	0.19	0.02	0.13	0.45
LLaMA-3.2-3B-Instruct doobučavanje	**`Twitter-Emo.SR`**	0.28	0.61	0.54	0.11	0.47	0.28
	`Reddit-Emo.SR`	0.22	0.53	0.50	0.08	0.39	0.33
	`Social-Emo.SR`	0.29	0.56	0.55	0.09	0.43	0.30

ma iste arhitekture (XLM-R_{base}), što je naročito uočljivo u vrednostima F_1^{Ma} mere. Rezultati na korpusu **`Twitter-Mor.SR`** variraju, ali ovaj korpus često donosi niže vrednosti tačnosti (Acc) u poređenju sa **`Reddit-Mor.SR`**, koji pokazuje nešto bolje rezultate zbog raznovrsnijeg sadržaja. Model Jerteh-355 i Twitter-XLM-R_{base} izdvajaju se kao najbolji po kombinaciji metrika F_1^{Ma}, F_1^{Mi}, i F_1^{w}, posebno na celokupnom korpusu **`Social-Mor.SR`**. Nasuprot tome, model BERTić postiže konzistentno slabije rezultate, što može ukazivati na ograničenja u prilagođavanju srpskom jeziku i društvenim moralnim kategorijama. Vrednosti metrika HS i HL su relativno stabilne među modelima, što ukazuje na ujednačenu procenu harmonije i konzistencije. Analiza rezultata iz tabele pokazuje promenljive performanse različitih modela u zadatku prepoznavanja moralnog sentimenta na različitim korpusima. Veći modeli poput XLM-R_{large} postižu bolje rezultate, sa vrednostima F_1^{Ma} do 38% na podkorpusu **`Twitter-Mor.SR`** i 35% na celom korpusu **`Social-Mor.SR`**, u poređenju sa manjim modelima, koji dostižu maksimalne vrednosti F_1^{Ma} oko 17-18%. Model Jerteh-355 takođe pokazuje prihvatljive performanse, sa boljim balansom u merama F_1^{Ma} i F_1^{w}.

Tabela 9.23: *Rezultati doobučavanja modela zasnovanih na BERT i LLaMA arhitekturama izvršenih na celokupnom korpusu* **Social-Mor.SR** *i njegovim podkorpusima na zadatku prepoznavanja moralnih kategorija prema moralnom sentimentu*

Osnovni model	Korpus	F_1^{Ma}	F_1^{Mi}	F_1^{w}	Acc	HS	HL
XLM-R_{base}	**Twitter-Mor.SR**	0.16	0.55	0.44	0.35	0.50	0.11
	Reddit-Mor.SR	0.18	0.58	0.49	0.35	0.52	0.11
	Social-Mor.SR	0.17	0.57	0.46	0.35	0.51	0.11
XLM-R_{large}	**Twitter-Mor.SR**	0.32	0.57	0.55	0.32	0.50	0.11
	Reddit-Mor.SR	0.27	0.62	0.57	0.34	0.55	0.11
	Social-Mor.SR	0.35	0.61	0.56	0.37	0.55	0.10
Twitter-XLM-R_{base}	**Twitter-Mor.SR**	0.16	0.55	0.44	0.35	0.49	0.11
	Reddit-Mor.SR	0.17	0.58	0.49	0.34	0.52	0.11
	Social-Mor.SR	0.17	0.56	0.47	0.35	0.50	0.11
Twitter-XLM-R_{large}	**Twitter-Mor.SR**	0.38	0.61	0.57	0.35	0.55	0.10
	Reddit-Mor.SR	0.27	0.61	0.57	0.32	0.53	0.11
	Social-Mor.SR	0.35	0.61	0.56	0.37	0.55	0.10
Jerteh-81	**Twitter-Mor.SR**	0.14	0.53	0.42	0.35	0.47	0.11
	Reddit-Mor.SR	0.14	0.55	0.46	0.33	0.50	0.11
	Social-Mor.SR	0.14	0.54	0.43	0.34	0.48	0.11
Jerteh-355	**Twitter-Mor.SR**	0.21	0.56	0.48	0.37	0.50	0.10
	Reddit-Mor.SR	0.18	0.59	0.50	0.33	0.51	0.11
	Social-Mor.SR	0.22	0.58	0.49	0.35	0.52	0.11
BERTić	**Twitter-Mor.SR**	0.14	0.54	0.43	0.36	0.49	0.11
	Reddit-Mor.SR	0.15	0.58	0.48	0.37	0.53	0.11
	Social-Mor.SR	0.15	0.57	0.45	0.36	0.50	0.11
LLaMA-3.2-3B-Instruct	**Twitter-Emo.SR**	0.03	0.09	0.03	0.01	0.05	0.25
	Reddit-Emo.SR	0.03	0.06	0.02	0.01	0.04	0.27
	Social-Emo.SR	0.03	0.07	0.02	0.01	0.04	0.26
LLaMA-3.2-3B-Instruct doobučavanje	**Twitter-Emo.SR**	0.26	0.47	0.44	0.07	0.36	0.18
	Reddit-Emo.SR	0.22	0.54	0.47	0.10	0.42	0.17
	Social-Emo.SR	0.26	0.52	0.48	0.06	0.39	0.17

Nešto niži rezultati na podkorpusu **Twitter-Mor.SR** u poređenju sa podkorpusom **Reddit-Mor.SR**, ukazuju na potencijalnu složenost u prepoznavanju jezičkih obrazaca kod iskazivanja moralnih stavova u kraćim tekstovima. Ukupne vrednosti F_1^{Mi} i F_1^{w} mera od 54-61% pokazuju umerenu preciznost, dok je mera HS sa vrednostima 49-55% relativno stabilna i niska, što ukazuje na izazove u prepoznavanju moralnih kategorija u oba sistema kategorizacije. Modeli Jerteh-81 i BERTić, iako obučavani isključivo nad korpusima srpskog jezika, su pokazali najniže performanse, dok veći modeli poput Twitter-XLM-R_{large} i XLM-R_{large} su pokazali najbolje rezultate na celokupnom korpusu **Social-Mor.SR**, kao i na njegovim podkorpusima.

Performanse klasifikacije na pojedinačnim kategorijama su proverene korišćenjem doobučenih Twitter-XLM-R_{large} modela, koji su pokazali najbolje ukupne performanse na skupu za testiranje u prethodnim eksperimentima. Vrednosti F_1 mere prikazane u tabeli 9.24 pokazuju značajnu promenljivost između klasifikacionih kategorija i korpusa koji su korišćeni za doobučavanje modela. Kategorije *care* i *harm* imaju najviše vrednosti F_1 mere u klasifikaciji moralnih osnova i moralnog sentimenta u svim korpusima, što ukazuje na sposobnost modela da tačno prepozna ove moralne vrednosti. Nasuprot tome, kategorije poput *subversion*, *cheating* i *degradation* iz kategorizacije moralnog sentimenta pokazuju konzistentno niske rezultate ($1\% \leq F_1 \leq 30\%$) kojima se uzrok može pronaći u težini razlikovanja prilikom izražavanja ovih vrednosti od svojih dihotomnih parova (*authority*, *fairness*, *purity*). Najbolji ukupni rezultati ostvareni su na korpusu **Social-Mor.SR** sa vrednostima

Tabela 9.24: *Performanse doobučenog modela Twitter-XLM-R_{large} na pojedinačnim kategorijama u klasifikaciji moralnih osnova i moralnog sentimenata prikazane korišćenjem F_1 mera*

Kategorija	Twitter-Mor.SR		Reddit-Mor.SR		Social-Mor.SR	
authority	0.28	0.44	0.22	0.07	0.34	0.39
subversion		0.01		0.12		0.13
care	0.85	0.66	0.89	0.75	0.88	0.74
harm		0.75		0.75		0.75
fairness	0.46	0.36	0.42	0.36	0.46	0.34
cheating		0.29		0.28		0.34
loyalty	0.49	0.58	0.31	0.28	0.46	0.36
betrayal		0.39		0.39		0.36
purity	0.39	0.44	0.18	0.01	0.37	0.44
degradation		0.17		0.13		0.30
F_1^{Ma}	0.50	0.41	0.40	0.31	0.50	0.41
F_1^{Mi}	0.66	0.60	0.66	0.61	0.68	0.60
F_1^{w}	0.64	0.58	0.64	0.59	0.67	0.59

F_1^{Mi} od 68% i F_1^{w} od 67%, dok **Reddit-Mor.SR** pokazuje najniže performanse za većinu kategorija (F_1^{Ma} od 31%). Uočene karakteristike sugerišu da je za ispravno prepoznavanje složenih moralnih vrednosti i njihovih sentimenata potreban raznovrsniji skup primera, naročito za kategorije sa nižim rezultatima. Poređenje performansi modela na zadacima klasifikacije moralnih osnova (5 kategorija) i moralnog sentimenata (10 kategorija) ukazuje na značajan uticaj granularnosti zadatka na rezultate klasifikacije, pri čemu sa povećanjem granularnosti (moralni sentiment) performanse izgrađenih modela značajno opadaju, što je delimično i očekivano usled povećane složenosti novog zadatka. Slično kao i u slučaju klasifikacije emocionalnog afekta, LLaMA modeli su pokazali niže performanse i nakon doobučavanja u odnosu na modele BERT arhitektura prilagođene srpskom jeziku. Ovaj uvid sugeriše na dalje potrebe za optimizacijom LLaMA modela u pogledu pronalaženja optimalnih vrednosti hiperparametara, balansiranjem klasa, obogaćivanjem podataka za obučavanje dodatnim resursima i korišćenja većih osnovnih modela.

9.4. Odnos izmedju sentimentalnih, emocionalnih i moralnih atributa

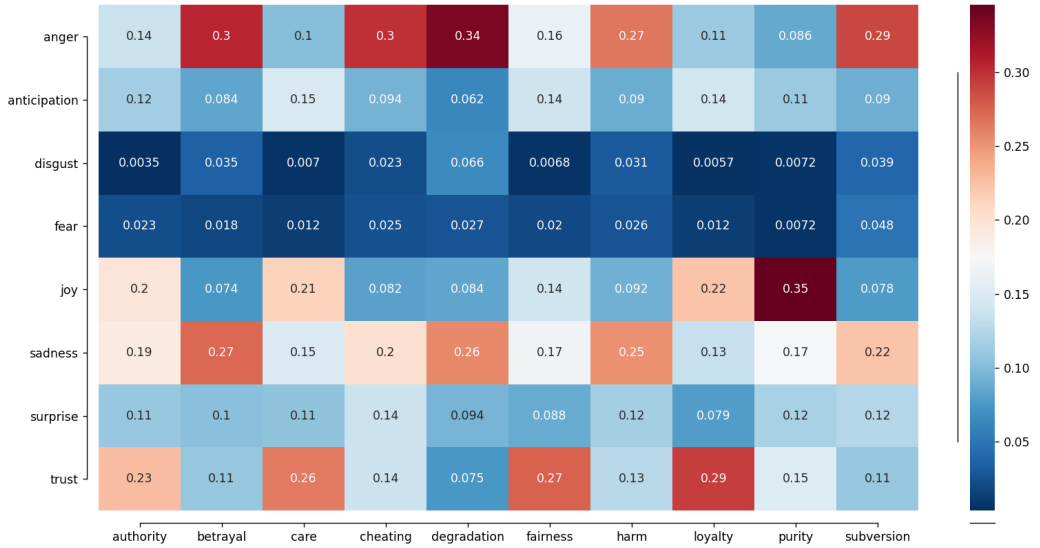
Sentiment predstavlja subjektivni odnos prema nekoj pojavi koji nastaje agregacijom emocija i utvrđenih verovanja. Istraživanja sentimenta i subjektiviteta u tekstu su se naglo proširila sa pojavom većih obeleženih skupova podataka i razumevanja značaja ovakve vrste analize. Za srpski jezik je izvršeno više različitih istraživanja za prepoznavanje sentimenta reči i tekstualnih sekvenci. U jednom istraživanju autori predlažu poluautomatsku metodu za kreiranje leksikona sentimentalnih reči polazeći od ručno kreirane liste reči i prepoznatih sinonima korišćenjem SWN leksikona [133]. Druga istraživanja obuhvataju prepoznavanje sintaktičkih negacija i njihov uticaj na promenu sentimenta tekstualne sekvence [122]. Najnovija istraživanja prepoznaju mogućnosti merenja intenziteta sentimenta reči i celokupnih tekstualnih sekvenci korišćenjem sintaktičkih pokretača intenziteta sentimenta u tekstovima (pogledati odeljak 8.2.1)

Iako se sentiment i emocije često koriste kao sinonimi, u suštini predstavljaju sasvim drugačije koncepte. Sentiment jeste određen višestrukim i kompleksnim kombinacijama emocionalnih reakcija, ali i misaonim delovanjima kroz izražene moralne stavove na određeni stimulans. Emocije daju granularniji i detaljniji prikaz u analizi sadržaja i odgova-

raju na pitanja šta se nalazi iza određenih sentimentalnih odrednica, kao što je prikazano u primerima u tabeli 9.25. U razvijenim metodologijama klasifikacije osnovnih ljudskih emocija, autori često naglašavaju opšti sentiment za svaku od kategorija. Tako su u Plutčikovom modelu kategorizacije emocija, kategorije *anger*, *disgust*, *fear*, *sadness* prepoznate kao kategorije sa negativnim sentimentom, kategorija *joy* kao nosilac pozitivnog sentimenta, dok su kategorije *anticipation* i *surprise* mešovitog karaktera, odnosno mogu se pojaviti kako u pozitivnim, tako i u negativnim sentimentalnim kontekstima. Sa druge strane, kategorizacija morala prema MFT prepoznaje vrlinu (opšte pozitivno) i manu (opšte negativno) za svaku osnovnu moralnu vrednost, odnosno definiše kategorije kao dihotomne parove podeljene prema moralnom sentimentu date sledećom kategorizacijom: *authority/subversion*, *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *purity/subversion*. S obzirom da se u razvijenoj metodologiji za klasifikaciju konverzacionih tekstova među zavisnim i nezavisnim atributima nalaze atributi iz sva tri aspekta jezika (sentimentalni, emocionalni, moralni), ovim delom istraživanja ćemo pokušati da utvrdimo njihove potencijalne korelacije i zavisnosti.

Tabela 9.25: *Primeri odnosa intenziteta sentimenta i prepoznatih emocionalnih reči u rečenicama na srpskom jeziku sa mešovitim pozitivnim i negativnim kontekstom*

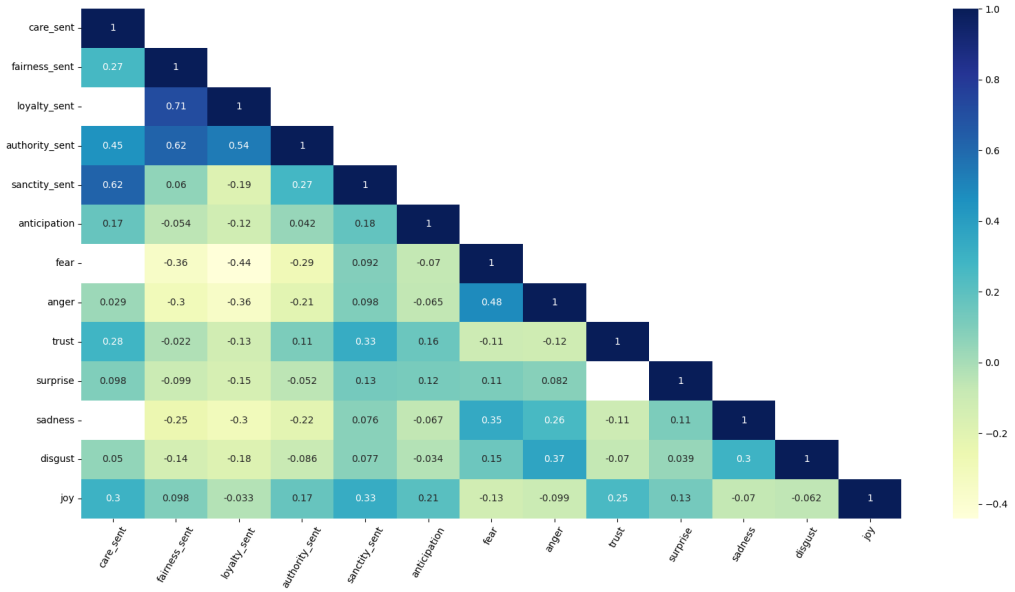
Primer	SRPOL	*EmoLex.SR*
Ovaj film je na mene ostavio jako **loš utisak**, iako sam dobio preporuke da ga **pogledam**	-1(-0.45)	(„*loš*“, ADJ) - *anger*, *disgust*, *fear* („*utisak*“, NOUN) - *surprise* („*pogledati*“, VERB) - *anticipation*
I pored **dobrih** vizuelnih efekata, ovaj film ne **zaslužuje** visoku **ocenu**	1 (0.28)	(„*dobar*“, ADJ) - *anticipation*, *joy*, *surprise*, *trust* („*zaslužiti*“, VERB) - *anticipation*, *trust* („*ocena*“, NOUN) - *anger*, *fear*, *sadness*
Dopao mi se film, iako me je u pojedinim trenucima **rastužio**	1 (0.17)	(„*dopadati*“, VERB) - *joy*, *trust* („*rastužiti*“, VERB) - *fear*, *sadness*



Slika 9.9: *Korelacija pojavljivanja emocionalnih i moralnih kategorija u konačnim obeležjima korpusa* **`Twitter-Mor.SR`**

Odnos između atributa sentimenta, emocionalnosti i moralnosti, koji su nastali kao rezultat ručnog obeležavanja i harmonizacije obeležja, je proveren u korpusu **`Social-Mor.SR`**, odnosno njegovom podkorpusu **`Twitter-Mor.SR`**. Višeznačna obeležja su najpre pretvorena u jednoznačna na taj način što je svakoj poruci iz korpusa dodeljena svaka pojedinačna

kategorija iz višeznačnog obeležja. Odnos između atributa koji predstavlja intenzitet sentimenta (SRPOL) i atributa moralnosti i emocionalnosti u podkorpusu **Twitter-Mor.SR** je predstavljen prilikom statističke analize ovog korpusa kada je utvrđena pravilnost u njihovim međusobnim interakcijama (pogledati odeljke 8.4.1 i 8.5.1). Iz dobijenih rezultata o odnosu emocionalnih i moralnih atributa (pogledati sliku 9.9), koji predstavljaju normalizovanu učestalost pojavljivanja emocionalnih kategorija po svakoj moralnoj kategoriji, moguće je utvrditi visok stepen zavisnosti koje između njih postoje. Uočljivo je da emocionalne kategorije *anger* i *sadness* pokazuju visok stepen i sličnu raspodelu učešća u moralnim kategorijama *degradation*, *cheating*, *betrayal*, *subversion* i *harm*. Sa druge strane, emocionalna kategorija *joy* je najzastupljenija u moralnoj kategoriji *purity*, a zatim u kategorijama *loyalty* i *care*, dok je kategorija *trust* pokazala najdominantnije prisustvo u kategorijama *loyalty* i *fairness*. Ovi, donekle očekivani rezultati u kojima je primetno zajedničko grupisanje kategorija sa pozitivnim i negativnim sentimentom između dva aspekta jezika, još jednom potvrđuju ispravnost uspostavljenih šema obeležavanja i postupka harmonizacije obeležja u korpusima **Social-Emo.SR** i **Social-Mor.SR**. Pored toga, uočljivo je da su emocionalne kategorije *sadness* i *anticipation* sledeće po stepenu zastupljenosti u moralnoj kategoriji *care*, a koje se inače upotrebljavaju kao emocionalna stanja kod iskazivanja brige za drugim.



Slika 9.10: *Pirsonovi koeficijenti korelacije (r, p≤0.05) između nezavisnih atributa intenziteta emocionalnog afekta i moralnog sentimenta u konverzacionim porukama na zadatku klasifikacije tipa delovanja na glasinu (**TD**)*

Pored toga, utvrđeno je postojanje statistički značajnog (p≤0.05) stepena Pirsonovih korelacija (*r*) između nezavisnih emocionalnih i moralnih atributa (EmoAtr, MorAtr) koji su korišćeni na zadatku klasifikacije tipa delovanja na glasinu (**TD**) (pogledati sliku 9.10). Na ovom zadatku su korišćene poruke sa Tviter i Redit društvenih mreža na engleskom jeziku, a emocionalni i moralni atributi su izračunati korišćenjem razvijenih *NRC.EmoInt* i *eMFD* leksikona za engleski jezik. Ovi atributi prema svojoj konstrukciji mere emocionalni intenzitet i intenzitet moralnog sentimenta za svaku kategoriju. Na slici su predstavljeni samo statistički značajni koeficijenti korelacije (p≤0.05), dok su koeficijenti bez statističkog značaja izostavljeni. Najveći intenzitet je prepoznat između atributa u svakoj od grupa (0.25-0.71), dok korelacije između atributa iz različitih grupa postoje, ali sa nešto manjim koeficijentima (0.17-0.3), što je donekle i očekivano zbog različitih pristupa u konstrukciji i pokrivenosti korišćenih leksikona.

10. Diskusija

U okviru ovog istraživanja dokazane su sledeće postavljene hipoteze (pogledati odeljak 1.1):

H1 ***Uključivanje moralnih i emocionalnih aspekata jezika značajno doprinosi tačnosti modela klasifikacije konverzacionih tekstova, što ukazuje na njihov značaj u interpretaciji i razumevanju konverzacionih tekstova.***

Izgrađeni modeli za klasifikaciju konverzacionih poruka elektronske pošte na zadatku **PL** i poruka sa društvenih mreža na zadatku **TD**, potvrđuju tačnost navedene hipoteze. U skupovima dodatnih atributa teksta uključeni su emocionalni i moralni atributi izgrađeni pomoću odgovarajućih semantičkih leksikona za engleski jezik. Pokrenuti su nezavisni eksperimenti u kojima su izgrađeni klasifikacion modeli zasnovani na BiLSTM+Att arhitekturama za analizu pojedinačnih poruka, ali i sekvence poruka u nizu, i to:

- Bez uključivanja pridruženih atributa;
- Uključivanjem raznovrsnih skupova pridruženih atributa teksta (Meta);
- Uključivanjem isključivo emocionalnih i moralnih pridruženih atributa (EmoAtr, MorAtr).

Svi modeli koriste vektorsku reprezentaciju poruka kao standardne atribute (BoW, Tf-Idf, Embd), dok su dodatni semantički atributi opciono uključivani u određene slojeve arhitekture. Evaluacija na nezavisnom testnom skupu je pokazala značajnu razliku u tačnosti između ovih pristupa. Pored toga što su arhitekture koje koriste sekvencijalne zavisnosti između poruka doprinele povećanju tačnosti klasifikacije, izgrađeni modeli nad ovim arhitekturama koji su uključili moralne i emocionalne aspekte su postigli povećanje tačnosti za +4.2% na **IG** i +3.8% na **TD** zadatku, u odnosu na modele bez uključivanja ovih atributa. Dobijeni rezultati potvrđuju da moralni i emocionalni aspekti jezika doprinose tačnosti modela klasifikacije, što ukazuje na njihov značaj u interpretaciji i klasifikaciji konverzacionih tekstova na različitim zadacima.

H2 ***Razvijeni semantički leksikoni moralnih i emocionalnih afekata reči za srpski jezik mogu doprineti prepoznavanju ovih aspekata u tekstu.***

Razvijeni semantički leksikon *SentiWords.SR* intenziteta sentimenta i prateći alat SR-POL za izračunavanje intenziteta sentimenta u tekstualnim sekvencama korišćenjem leksikona i skupova pravila kojima se utvrđuju pokretači sentimenta na osnovu sintaksnih i semantičkih karakteristika tekstualne sekvence predstavlja važan resurs i prvi takve vrste razvijen za srpski jezik (pogledati odeljak 8.2.1). Ovaj resurs je korišćen kao pomoćni alat za razvoj i proveru razvijenih resursa za emocionalnost – *EmoLex.SR* sa ∼9.8k lema_{Sr}-PoS parova (pogledati odeljak 8.2.3) i moralnost – *MFD.SR* sa ∼4.3k lema_{Sr}-PoS parova (pogledati odeljak 8.6).

Intenzivnom i nezavisnom proverom nad različitim domenskim kolekcijama tekstualnih podataka označenih na emocionalni afekt, utvrđeno je da je jezički i kulturološki prilagođen emocionalni leksikon *EmoLex.SR* u mogućnosti da prepozna emocionalne signale u tekstualnim sekvencama (pogledati odeljak 9.1.2). Dodatno, razvijeni leksikon u prepoznavanju emocionalnih signala nadmašuje leksikone koji su nastali

automatskim prevođenjem postojećih leksikona emocionalnog afekta na engleskom jeziku.

Analogno, prvi leksikon moralnih vrednosti na srpskom jeziku *MFD.SR*, potvrdio je svoju uspešnost u prepoznavanju dominantnih moralnih vrednosti u opisima njihovog razumevanja ispitanika sa srpskog govornog područja (pogledati odeljak 9.1.3).

H3 ***Kvantitativnom analizom atributa moguće je utvrditi korelaciju između iskazanih moralnih stavova i emocionalnih reakcija u konverzacionim porukama.***

Kvantitativna analiza sprovedena je nad korpusom **`Social-Mor.SR`** koji sadrži ∼13.6k konverzacionih poruka, sa višeznačnim emocionalnim i moralnim obeležjima. Rezultati korelacione analize pokazali su značajnu povezanost između moralnih stavova i emocionalnih reakcija u obeleženom korpusu (pogledati odeljak 9.4). Na primer, emocionalne kategorije *anger* i *sadness* su identifikovane kao najdominantnije u svim kategorijama negativnog, a nasuprot njima kategorije *trust* i *joy* pozitivnog moralnog sentimenta.

Međuzavisnosti su potvrđene postojanjem prihvatljivog stepena korelacija (r=0.17-0.3, p$\leq$0.05) između nezavisnih moralnih i emocionalnih atributa korišćenih na zadacima klasifikacije teksta u okviru ovog istraživanja. Ovi nalazi ukazuju na postojanje povezanosti između moralnih i emocionalnih aspekata jezika i potvrđuju mogućnost kvantitativnog merenja ove povezanosti.

H4 ***Moguće je razviti klasifikacione modele prihvatljive tačnosti koji na osnovu karakteristika tekstualnih sadržaja iz razvijenih obeleženih konverzacionih korpusa na srpskom jeziku, mogu da prepoznaju moralne i emocionalne aspekte jezika.***

Obeleženi korpusi konverzacionih tekstova na srpskom jeziku **`Social-Emo.SR`** sa ∼34.6k i **`Social-Mor.SR`** ∼13.6k konverzacionih poruka, korišćeni su za obučavanje klasifikacionih modela za prepoznavanje emocionalnog afekta i moralnih vrednosti u tekstualnim sekvencama na srpskom jeziku. Izgrađeni klasifikacioni modeli su postigli performanse koje su u rangu sa performansama modela na istim zadacima višeznačne klasifikacije koja su prijavljena u drugim istraživanjima. Inicijalni rezultati prepoznavanja emocionalnog afekta u tekstovima na srpskom jeziku su dostigli vrednosti od na celom korpusu **`Social-Emo.SR`**, i nešto više vrednosti na podkorpusu **`Twitter-Emo.SR`** (pogledati odeljak 9.3.1).

Dodatno, doobučavanje LLaMA modela, verzije LLaMA-3.2-3B-Instruct za prepoznavanje emocionalnih i moralnih signala u tekstovima na srpskom jeziku, nad obeleženim korpusima u okviru ovog istraživanja, značajno poboljšava performanse ovog LLM modela, čime se otvaraju nove mogućnosti za precizno prepoznavanje ovih aspekata na srpskom jeziku. Ovi nalazi ukazuju na važnost razvoja odgovarajućih semantičkih jezičkih resursa, ali istovremeno potvrđuju potrebu za prilagođavanjem već izgrađenih AI modela specifičnostima srpskog jezika i zadatku koji se rešava.

Projektovanje optimalne arhitekture neuronske mreže zahteva balansiranje između kompleksnosti i preciznosti modela. Eksperimentisanjem sa različitim brojem slojeva i jedinica po sloju može se pronaći konfiguracija koja najbolje odgovara zadatku klasifikacije. Uvođenje slojeva za normalizaciju težina i izbacivanje čvorova mreže može pomoći u regularizaciji modela i sprečavanju preprilagođavanja. U primeni predložene metodologije na zadacima **`PL`** i **`TD`**, dobijeni rezultati za različite tehnike predstavljanja teksta, specijalno BERT-Embd, zavise od brojnih faktora i prvenstveno su određeni ograničenjima u dostupnim računarskim resursima neophodnim za obučavanje modela. Najpre, za kreiranje BERT-

Embd reči korišćeni su opšti, već obučeni, BERT modeli sa manjim brojem parametara što može uticati na ukupne performanse klasifikacionog modela u ovim eksperimentima. Nadalje, podaci korišćeni u zadacima **PL** i **TD** pripadaju specifičnim domenima i stilovima komunikacije (poslovna i neformalna), te bi u cilju poboljšanja performansi bilo potrebno doobučiti BERT model za razumevanje ovakvih tipova podataka, čime bi rečnik uključio i reči karakteristične za ove domene, odnosno stilove izražavanja. Najzad, ograničenje koje postoji u veličini rečnika za BERT modele (30,522), kao i za maksimalnu dužinu ulazne sekvence (MSL) koje ovi modeli podržavaju (128, 256, 512, 768), u mnogome može uticati na krajnje performanse. Uočena ograničenja će poslužiti kao smernice za dalja unapređenja predložene metodologije.

Pored L2 normalizacije i kombinacije metoda za izbor Meta atributa, kao što su F_s, permutacija vrednosti i isključivanje jednog atributa, moguća unapređenja uključuju implementaciju SHAP algoritma [125] kako bi se dobio podrobniji uvid u doprinos svakog atributa u klasifikacionom modelu. Takođe, konačan izbor atributa se može unaprediti težinskim uprosečavanjem vrednosti dobijenih iz pojedinačnih metoda, što može doprineti stabilnijem i pouzdanijem izboru relevantnih atributa. Pored toga, primena tehnika za smanjivanje dimenzionalnosti, poput autoenkodera ili analize glavnih komponenti (eng. *Principal Component Analysis*, *PCA*), može otkriti potencijalne latentne faktore i smanjiti redundantnost među izabranim atributima. Na kraju, primena drugih metoda, kao što je rekurzivna eliminacija atributa (eng. *Recursive Feature Elimination*, *RFE*) koja omogućava iterativno uklanjanje manje značajnih atributa, potencijalno može poboljšati predloženi pristup za kreiranje konačne liste najznačajnijih Meta atributa.

Prilikom formiranja semantičkih leksikona, jedan od ključnih uočenih nedostataka ogleda se u ograničenoj sposobnosti razlikovanja značenja polisemičnih reči, kao što je, na primer, reč „*kocka*“ (NOUN), koja u zavisnosti od konteksta može označavati geometrijsko telo (*neutral*; *non-moral*) ili igru na sreću (*joy*, *sadness*, *trust*; *fairness*, *betrayal*). U statičkim leksikonima, gde su vrednosti dodeljene na nivou leme i obeležja vrste reči bez obzira na kontekst upotrebe, ovakve reči mogu biti netačno interpretirane ili kategorizovane, što dovodi do potencijalnih grešaka u identifikovanju emocionalnog ili moralnog sadržaja. Nemogućnost razlikovanja konteksta prilikom korišćenja leksikona znači da vrednost može biti pogrešno dodeljena, ili obrnuto, da relevantna upotreba može ostati neprepoznata, čime je znatno smanjena preciznost analize datog aspekta jezika. Ovaj problem može se prevazići primenom modela transformer arhitekture (BERT, GPT), koji omogućavaju dinamičku reprezentaciju značenja reči u zavisnosti od njihovog konteksta u rečenici, čime se može postići preciznije identifikovanje određenih aspekata jezika u zavisnosti od konteksta.

Upotreba LLM uvodi metodološke rizike (halucinacije, pristrasnosti) koji se u izlazima manifestuju kao pogrešna lematizacija, netačno PoS obeležavanje (PoS-Tagging), konstrukcija višerečnih izraza (eng. *Multi-Word Expression*, *MWE*) pri prevođenju i pronalaženju sinonima, nenormirana upotreba ijekavske varijante, redundantne sekvence kod generisanja ili interlingvalna interferencija iz srodnih jezika. Predupređivanje ovih izazova je u okviru istraživanja rešeno primenom automatske i ručne provere u svakom slučaju korišćenja LLM. Prilikom konstrukcije *EmoLex.SR* leksikona, LLM, konkretno Čet-GPT *gpt-3.5-turbo* model, je korišćen za rešavanje zadataka T1–T4, na kojima je LLM, nakon primenjene provere, pokazao značajna unapređenja u odnosu na postojeće metode, i to:

- Prevođenja pojedinačne reči (T1) za +22.5% u odnosu na GT alat.
- Obeležavanja pojedinačnih reči u kategorije emocionalnog afekta (T2) pri čemu je tačnost poklapanja sa konačnim obeležjem povećana za +27% u odnosu na automatski preveden leksikon sa engleskog jezika.

- Formiranja liste sinonima za datu reč (T3) pri čemu je tačnost preklapanja sa konačnim skupom sinonima povećana za +32.5% u odnosu na SWN leksikon i ručne metode.
- Formiranja paralelnog korpusa obeleženog na emocionalni afekt (T4) pri čemu je efikasnost povećana za približno ∼+49.5%[84] u odnosu na GT alat (prevođenje rečenica + obeležavanje).

U okviru predobeležavanja korpusa na moralnu vrednost, korišćeni model *Falcon-7B-Instruct* je pokazao pristrasnost prema kategorijama *care* i *harm*, sa značajno većim brojem pojavljivanja ovih kategorija u obeleženim podacima dobijenih pomoću inžinjeringa instrukcija nad ovim modelom. Ovo ponašanje je delimično očekivano usled različitih kulturoloških i jezičkih karakteristika koje su na modele prenete iz podataka korišćenih za njihovo obučavanje. Uočena pristrasnost je u značajnoj meri korigovana tokom ručne provere obeležja. Dodatno, na ovom slučaju korišćenja LLM, kao i u slučaju doobučavanja LLaMA modela, instrukcije su konstruisane na engleskom jeziku sa ciljem da ih model bolje razme, a zbog nedovoljne podrške za srpski jezik ovih modela. Na ovaj način je delimično prevaziđen problem nedovoljne podrške za srpski jezik, jer se od modela očekivalo tačno razumevanje instrukcija (engleski → bolja podrška), razumevanje pridruženog teksta poruke (srpski → slabija podrška) i što tačniji generisani sadržaj, odnosno obeležja (engleski → bolja podrška).

Ručna anotacija emocionalnog afekta i moralnih vrednosti u tekstu je predstavljala zahtevan zadatak zbog složenosti prepoznavanja ovih aspekata, subjektivnosti tumačenja i razlike u percepciji među anotatorima iz različitih socio-demografskih grupa. Na obeleženim korpusima, Kohenov kapa koeficijent (k), koji je korišćen za izračunavanje slaganja među anotatorima, je pokazao umerene vrednosti usklađenosti obeležavanja u proseku od ∼0.26 na korpusu **`Social-Emo.SR`** i ∼0.17 na korpusu **`Social-Mor.SR`**. Dalja istraživanja će biti usmerena na povećanje stepena usaglašenosti kroz postupak provere tačnosti obeležja, uspostavljanjem preciznijih instrukcija za obeležavanje i uvođenjem dodatnih ljudskih anotatora ($n > 2$). Provera se može izvršiti korišćenjem nekih od naprednih tehnika nadgledanog ili polunadgledanog učenja nad primerima na kojima su postignuta potpuna ili delimična slaganja, čime bi se postupak ubrzao i učinio efikasnijim.

Izgradnja modela za prepoznavanje emocionalnih i moralnih dimenzija na srpskom jeziku, doobučavanjem postojećih modela BERT i LLaMA arhitektura, je predstavljala poseban izazov zbog uočene umerene konzistentnosti obeležja u korpusima korišćenim za obučavanje, specifičnih jezičkih osobina poruka sa društvenih mreža i dostupnih računarskih resursa. Poređenje rezultata izgrađenih klasifikacionih modela ukazuje da prilagođavanje modela specifičnim domenima i korišćenje većih arhitektura pruža bolje performanse u preciznoj klasifikaciji emocionalnog afekta i moralnih vrednosti. Lošiji rezultati doobučavanja modela *LLaMa-3.2-3B-Instruct* u odnosu na BERT modele mogu se objasniti razlikama u arhitekturi, stepenom prilagođenosti modela srpskom jeziku i mogućnostima finog podešavanja modela. BERT arhitekture koriste MLM, koji je bolje prilagođen za zadatke klasifikacije teksta, dok je LLaMA optimizovan za generativne zadatke, što može smanjiti njegovu preciznost na klasifikacionim zadacima. Uzrok se može potražiti i u algoritmima za tokenizaciju koje ovi modeli koriste (WP/SP naspram BPE), kao i u količini podataka neophodnih za efektivno doobučavanje modela na specifičnim zadacima. Neke od tehnika koje će se u narednim koracima koristiti za poboljšanje performansi

[84] Kvantitativna procena je izračunata približno jednostavnim sabiranjem unapređenja po pojedinačnim koracima u odnosu na postojeće metode koje obuhvataju automatsko prevođenje i ručnu proveru tačnosti prevoda i dodeljenih obeležja.

modela jesu optimizacija hiperparametara modela, dodatna provera obeležja, povećanje obima i raznovrsnosti podataka za obuku, korišćenje semantičkih baza znanja, kao i korišćenje osnovnih modela sa većim brojem parametara (LLaMA-3-7B/40B/405B-Instruct). Za očekivati je da prethodno doobučavanje LLaMA modela nad velikim i opštim korpusima srpskog jezika, naročito onim specifičnim za određene domene (pravo, medicina, književna literatura, društvene mreže), značajno poboljša performanse klasifikacionih modela na specifičnim zadacima, koji obuhvataju i zadatke prepoznavanja emocionalnog afekta i moralne vrednosti u tekstu.

11. Zaključak

Modelovanje moralnih i emocionalnih aspekata jezika u klasifikaciji konverzacionih tekstova predstavlja važan zadatak za razumevanje konverzacionih tekstova i ujedno predstavlja aktuelan istraživački problem u računarskoj lingvistici. Prepoznate moralne i emocionalne dimenzije teksta omogućavaju dublje razumevanje interakcija između učesnika u konverzaciji. Moralne vrednosti, kao što su pravičnost, lojalnost, i autoritet, često prožimaju konverzaciju i utiču na način na koji ljudi odgovaraju i reaguju. Slično tome, emocionalni ton poruke, prepoznat kroz emocionalna stanja radosti, ljutnje ili straha, može značajno oblikovati tok razgovora i uticati na njegov ishod. Pridruživanje pridruženih moralnih i emocionalnih atributa, kao što su intenzitet emocija ili verovatnoća da poruka sadrži moralne vrednosti, kao što su briga ili autoritet, omogućava klasifikacionim algoritmima dublje razumevanje konverzacionog konteksta. Ovi atributi omogućavaju složeniji uvid u konverzaciju i pomažu u prepoznavanju suptilnih varijacija u značenju, što je od velike važnosti u pojedinim klasifikacionim zadacima kao što su prepoznavanje privatnih poruka elektronske pošte ili lažnih objava na društvenim mrežama.

Istraživanje ukazuje na složenost i značaj razvoja naprednih metoda za analizu konverzacionih tekstualnih sadržaja. Upotrebom naprednih arhitektura dubokog učenja i neuronskih mreža, kao što su višeslojne BiLSTM nad povezanim porukama, omogućeno je postizanje visoke tačnosti u klasifikaciji konverzacionih tekstova. Specijalizovane BiLSTM arhitekture su pokazale naročitu sposobnost u dekodiranju sekvencijalnih zavisnosti između neposrednih poruka koje karakterišu konverzacije, što ih čini pogodnim za analize koje nadilaze jednostavne klasifikacije tekstualnog sadržaja i prelaze u sferu razumevanja dimenzija dinamike konverzacije. Pored arhitekturalnog aspekta, istraživanje je obuhvatilo integraciju pridruženih atributa sadržaja, u kojima su moralni i emocionalni atributi dobili poseban značaj. Ovi atributi omogućavaju precizniju identifikaciju i klasifikaciju skrivenih vrednosnih sudova i afektivnih elemenata koji su potvrdili važnu ulogu u konverzacionim interakcijama. Osim toga, dodavanje konverzacionih atributa je ML modelima omogućilo da identifikuju strukturu interakcija, prepoznajući implicitne promene u kontekstu dijaloga. Kombinacija ovih atributa sa leksičkim i sintaktičkim karakteristikama je pružila holistički pristup u klasifikaciji, koji je neophodan za dublju analizu kompleksnih konverzacionih tekstova i poboljšanje performansi klasifikacionih modela.

Razvoj specijalizovanih leksičkih resursa, poput semantičkih leksikona, obeleženih korpusa i ML modela, predstavlja značajan iskorak ka uspostavljanju efikasnih alata za prepoznavanje moralnih i emocionalnih aspekata u srpskom jeziku. U okviru ovog istraživanja konstruisani su emocionalni (*WNA.SR*, *EmoLex.SR*) i moralni (*MFD.SR*) leksikoni, specifično prilagođeni lingvističkim osobenostima srpskog jezika i kulturološkim karakteristikama srpskog govornog područja. Posebno, SWN leksikon je proširen obeležjima emocionalnog afekta za $\sim$1000 postojećih sinseta. Pored toga, obeleženi su korpusi poruka na srpskom jeziku preuzetih sa društvenih mreža Tviter i Redit, u pogledu emocionalnog afekta i moralnih vrednosti, pod oznakama **`Social-Emo.SR`** i **`Social-Mor.SR`**, tim redosledom. Obeležavanje je sprovedeno kombinacijom inovativnih tehnika mašinskog učenja, kao što su transferno učenje i inženjering instrukcija, i ručne provere koja je osigurala tačnost automatski obeleženih podataka. Razvoj ovih resursa predstavljao je složen proces zbog nekoliko ključnih faktora. Najpre, precizno definisanje granica između različitih moralnih i emocionalnih kategorija se pokazalo izuzetno zahtevnim zbog njihove često preklapajuće prirode. Drugo, prikupljanje dovoljno obimnog i raznovrsnog korpusa koji sadrži relevant-

ne moralne i emocionalne signale u tekstu je zahtevalo značajne napore u pronalaženju odgovarajućih tekstualnih izvora. Najzad, automatske metode su iziskivale provere kvaliteta oznaka dobijenih na ovaj način, što je zahtevalo dodatne resurse (ljudske, vremenske, materijalne) kako bi se postigla preciznost i konzistentnost razvijenih obeleženih korpusa. Na osnovu obeleženih korpusa, razvijeni su ML modeli za automatsko prepoznavanje ovih aspekata jezika u tekstovima na srpskom jeziku. Modeli su nastali doobučavanjem postojećih modela BERT i LLaMA arhitektura koji imaju sposobnost razumevanja tekstova na srpskom jeziku. Postignuti rezultati ukazuju na konkurentne performanse modela, koje su u skladu sa rezultatima drugih savremenih rešenja razvijenih za isti zadatak, što potvrđuje njegovu primenljivost i pouzdanost.

Dalja istraživanja biće usmerena na unapređenje razvijenih resursa kroz nekoliko koraka. Prvenstveno, planira se kontinuirana provera i proširivanje postojećih leksičkih resursa i korpusa za srpski jezik kako bi se postigla sveobuhvatnija pokrivenost različitih kontekstualnih varijacija u jeziku. Takođe, planirano je poboljšanje performansi izgrađenih klasifikacionih ML modela, korišćenjem boljih varijanti i arhitektura osnovnih modela, kao i prethodnim doobučavanjem osnovnih modela nad većim korpusima srpskog jezika. Pored proširenja i adaptacije resursa, dalja istraživanja biće usmerena na primenu naprednih metoda mašinskog učenja u cilju automatizacije postupka obeležavanja i provere leksičkih resursa, čime bi se smanjila zavisnost od ručnog rada i ubrzao razvoj kvalitetnih resursa. Naročita pažnja biće stavljena na pronalaženje efikasnih matematičkih pristupa za automatsko izračunavanje i dodelu klasifikacionih obeležja. U okviru ovog plana, važan korak će biti i istraživanje mogućnosti za primenu razvijenih resursa u različitim NLP sistemima na srpskom jeziku, kao što su sistemi za analizu konverzacija na društvenim mrežama ili sistemi za prepoznavanje emocionalnih stanja i moralnih vrednosti u konverzacijama.

Iako rezultati ovog istraživanja otvaraju mogućnosti za naprednu analizu moralnih i emocionalnih aspekata srpskog jezika, njihova primena nosi etičke rizike, koja uglavnom podrazumeva mogućnost diskriminacije i podržavanja potencijalnih pristrasnosti u društvenim normama. Modeli trenirani nad specifičnim kulturnim obrascima mogu reflektovati postojeće stereotipe, što može rezultovati nepravilnim tumačenjem komunikacija i diskriminacijom korisnika. Stoga će naredni istraživački koraci uključiti otkrivanje i korekciju potencijalnih pristrasnosti, kao i uspostavljanje etičkih regulativa kako bi se osigurala odgovorna i pravična upotreba modela razvijenih za prepoznavanje emocionalnih i moralnih aspekata u tekstovima napisanim na srpskom jeziku.

12. Bibliografija

[1] M. Abdul-Mageed and L. Ungar. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728, 2017.

[2] M. Abdulhai, G. Serapio-García, C. Crepy, D. Valter, J. Canny, and N. Jaques. Moral Foundations of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, 2024.

[3] G. Agrawal, S. Gummuluri, and C. Spera. Beyond-RAG: Question Identification and Answer Generation in Real-Time Conversations. *arXiv preprint arXiv:2410.10136*, 2024.

[4] V. Ahire and S. Borse. Emotion Detection from Social Media Using Machine Learning Techniques: A Survey. In *Applied Information Processing Systems: Proceedings of ICCET 2021*, pages 83–92. Springer, 2022.

[5] R. Akkiraju, A. Xu, D. Bora, et al. FACTS About Building Retrieval Augmented Generation-based Chatbots. *arXiv preprint arXiv:2407.07858*, 2024.

[6] A. Alhogail and A. Alsabih. Applying Machine Learning and Natural Language Processing to Detect Phishing Email. *Computers & Security*, 110:102414, 2021.

[7] R. S. H. Ali and N. El Gayar. Sentiment Analysis using Unlabeled Email data. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pages 328–333. IEEE, 2019.

[8] S. Alkhereyf and O. Rambow. Work Hard, Play Hard: Email Classification on the Avocado and Enron Corpora. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 57–65, 2017.

[9] S. Alkhereyf and O. Rambow. Email Classification Incorporating Social Networks and Thread Structure. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1336–1345, 2020.

[10] S. Angel Deborah, T. Mirnalinee, and S. M. Rajendram. Emotion Analysis on Text Using Multiple Kernel Gaussian... *Neural Processing Letters*, 53:1187–1203, 2021.

[11] D. Antonakaki, P. Fragopoulou, and S. Ioannidis. A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, 164:114006, 2021.

[12] O. Araque, L. Gatti, and K. Kalimeri. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184, 2020.

[13] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[14] I. Asimov. Three laws of robotics. *Asimov, I. Runaround*, 2, 1941.

[15] M. Atari, J. Haidt, J. Graham, S. Koleva, S. T. Stevens, and M. Dehghani. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157–1188, 2023.

[16] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38:100311, 2020.

[17] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[18] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, 2022.

[19] V. Batanović, M. Cvetanović, and B. Nikolić. A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLoS One*, 15(11):e0242050, 2020.

[20] V. Batanović, N. Ljubešić, T. Samardžić, and T. Erjavec. Serbian linguistic training corpus SETimes.SR 2.0, 2023. Slovenian language resource repository – CLARIN.SI.

[21] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM journal of Research and Development*, 63(4/5):4–1, 2019.

[22] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[23] V. Bobicev, V. Maxim, T. Prodan, N. Burciu, and V. Angheluş. Emotions in words: Developing a multilingual wordnet-affect. In *Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010, Iaşi, Romania, March 21-27, 2010. Proceedings 11*, pages 375–384. Springer, 2010.

[24] P. Boddington. *Towards a code of ethics for artificial intelligence*. Springer, 2017.

[25] M. Bogdanović, J. Kocić, and L. Stoimenov. SRBerta – A Transformer Language Model for Serbian Cyrillic Legal Texts. *Information*, 15(2):74, 2024.

[26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[27] C. Bothe, C. Weber, S. Magg, and S. Wermter. EDA: Enriching Emotional Dialogue Acts using an Ensemble of Neural Annotators. *arXiv preprint arXiv:1912.00819*, 2019.

[28] C. Bothe and S. Wermter. Conversational Analysis of Daily Dialog Data using Polite Emotional Dialogue Acts. *arXiv preprint arXiv:2205.02921*, 2022.

[29] M. M. Bradley and P. J. Lang. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Technical Report C-1 1, The Center for Research in Psychophysiology, University of Florida, 1999.

[30] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017.

[31] L. Bulla, A. Gangemi, and M. Mongiovì. Do Language Models Understand Morality? Towards a Robust Detection of Moral Content. In *International Workshop on Value Engineering in AI*, pages 98–113. Springer, 2023.

[32] L. Bulla, S. D. Giorgis, A. Gangemi, L. Marinucci, and M. Mongiovì. Detection of morality in tweets based on the moral foundation theory. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 1–13. Springer, 2022.

[33] J. Camacho-Collados and M. T. Pilehvar. On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 40–46, 2018.

[34] E. Cambria, R. Speer, C. Havasi, and A. Hussain. SenticNet: A Publicly Available Semantic Resource for Opinion Mining. In *2010 AAAI Fall Symposium (FS-10-02)*, pages 14–18, 2010.

[35] E. Cambria and B. White. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine*, 9(2):48–57, 2014.

[36] C. D. Cameron, K. A. Lindquist, and K. Gray. A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and social psychology review*, 19(4):371–394, 2015.

[37] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & electrical engineering*, 40(1):16–28, 2014.

[38] M. X. Chen, B. N. Lee, G. Bansal, Y. Cao, S. Zhang, J. Lu, J. Tsay, Y. Wang, A. M. Dai, Z. Chen, et al. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2287–2295, 2019.

[39] Z. Chen, J. M. Zhang, M. Hort, M. Harman, and F. Sarro. Fairness Testing: A Comprehensive Survey and Analysis of Trends. *ACM Transactions on Software Engineering and Methodology*, 33(5):1–59, 2024.

[40] N. Chhaya, K. Chawla, T. Goyal, P. Chanda, and J. Singh. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86, 2018.

[41] S. Choudhari, N. Choudhary, S. Kaware, and A. Shaikh. Email Prioritization Using Machine Learning. *Available at SSRN 3568518*, 2020.

[42] A. I. Cislowska, B. P. Acuna, et al. Integration of Chatbots in Additional Language Education: A Systematic Review. *European journal of Educational Research*, 13(4):1607–1625, 2024.

[43] D. J. Ciuk and J. Rottman. Moral conviction, emotion, and the influence of episodic versus thematic frames. *Political Communication*, 38(5):519–538, 2021.

[44] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[45] A. Conneau and G. Lample. Cross-lingual Language Model Pretraining. In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, volume 32, pages 7059 – 7069, 2019.

[46] K. Crockett, D. Mclean, A. Latham, and N. Alnajran. Cluster analysis of twitter data: A review of algorithms. In *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, volume 2, pages 239–249, 2017.

[47] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[48] T. M. De los Santos and R. L. Nabi. Emotionally charged: Exploring the role of emotion in online news information seeking and processing. *journal of Broadcasting & Electronic Media*, 63(1):39–58, 2019.

[49] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, 2020.

[50] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and A. Zubiaga. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, 2017.

[51] P. E. Ekman and R. J. Davidson. *The nature of emotion: Fundamental questions.* Oxford University Press, 1994.

[52] M. El-Assady, R. Sevastjanova, D. Keim, and C. Collins. ThreadReconstructor: Modeling reply-chains to untangle conversational text through visual analytics. In *Computer Graphics Forum*, volume 37, pages 351–365. Wiley Online Library, 2018.

[53] O. Enayet and S. R. El-Beltagy. NileTMRG at SemEval-2017 task 8: Determining rumour and veracity support for rumours on Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 470–474, 2017.

[54] M. Fajcik, P. Smrz, and L. Burget. BUT-FIT at SemEval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104, 2019.

[55] M. Farooq, V. De Silva, H. Tibebu, and X. Shi. Conversational Emotion Detection and Elicitation: A Preliminary Study. In *2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET)*, pages 1–5. IEEE, 2023.

[56] C. Fellbaum. WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

[57] A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *journal of Machine Learning Research*, 20(177):1–81, 2019.

[58] O. Flanagan. *How to Do Things with Emotions: The Morality of Anger and Shame Across Cultures*. Princeton University Press, 2021.

[59] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

[60] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32:221, 1948.

[61] P. Foot. *The problem of abortion and the doctrine of double effect*, volume 5. Oxford, 1967.

[62] J. A. Frimer, R. Boghrati, J. Haidt, J. Graham, and M. Dehgani. Moral foundations dictionary for linguistic analyses 2.0. *Unpublished manuscript*, 2019.

[63] L. Gatti, M. Guerini, and M. Turchi. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421, 2015.

[64] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[65] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski. SemEval-2019 Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*, pages 845–854, 2019.

[66] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

[67] J. Graham, J. Haidt, and B. A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *journal of personality and social psychology*, 96(5):1029, 2009.

[68] A. Grattafiori, A. Dubey, A. Jauhri, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[69] A. Graves. *Long Short-Term Memory*, pages 37–45. Springer, 2012.

[70] J. Greene and J. Haidt. How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12):517–523, 2002.

[71] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[72] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[73] J. Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001.

[74] Z. Halim, M. Waqar, and M. Tahir. A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. *Knowledge-Based Systems*, 208:106443, 2020.

[75] K. Hämmerl, B. Deiseroth, P. Schramowski, J. Libovický, C. Rothkopf, A. Fraser, and K. Kersting. Speaking Multiple Languages Affects the Moral Bias of Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156, 2023.

[76] A. Haque and M. Abulaish. An emotion-enriched and psycholinguistics features-based approach for rumor detection on online social media. In *Proceedings of the 11th International Workshop on Natural Language Processing for Social Media*, pages 28–37, 2023.

[77] M. Hasan, E. Rundensteiner, and E. Agu. Automatic emotion detection in text streams by analyzing Twitter data. *International Journal of Data Science and Analytics*, 7:35–51, 2019.

[78] T. Hastie, R. Tibshirani, J. Friedman, et al. *The elements of statistical learning*. Springer, 2009.

[79] Z. He, S. Guo, A. Rao, and K. Lerman. Whose Emotions and Moral Sentiments do Language Models Reflect? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6611–6631, 2024.

[80] M. C. Hinojosa Lee, J. Braet, and J. Springael. Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores. *Applied Sciences*, 14(21):9863, 2024.

[81] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.

[82] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246, 2021.

[83] E. Hoque and G. Carenini. Multiconvis: A visual text analytics system for exploring a collection of online conversations. In *Proceedings of the 21st international conference on intelligent user interfaces*, pages 96–107, 2016.

[84] D. Hovy, A. Johannsen, and A. Søgaard. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, pages 452–461, 2015.

[85] J. Howard and S. Ruder. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

[86] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018.

[87] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[88] T. Ivanović, R. Stanković, B. Š. Todorović, and C. Krstev. Corpus-based bilingual terminology extraction in the power engineering domain. *Terminology*, 28(2):228–263, 2022.

[89] S. Jabbari, B. Allison, D. Guthrie, and L. Guthrie. Towards the Orwellian Nightmare: Separation of Business and Personal Emails. In *Proceedings of the COLING/ACL 2006 Main conference poster sessions*, pages 407–411, 2006.

[90] P. Janičić and M. Nikolić. *Veštačka inteligencija*. Univerzitet u Beogradu, Matematički fakultet, 2021.

[91] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[92] S. Kaplar, M. Radojičić, I. Obradović, B. Lazić, and R. Stanković. Solution for quantitative analysis of texts in Serbian based on syllables. In *ICIST 2018 Proceedings*, volume 2, pages 315–20, 2018.

[93] J. Kaur and J. R. Saini. Emotion detection and sentiment analysis in text corpus: a differential study with informal and formal writing styles. *International journal of Computer Application, ISSN*, pages 0975–8887, 2014.

[94] J. D. Kenton, M.-W. Chang, and L. K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, volume 1, pages 4171–4186, 2019.

[95] B. Kerr. Thread arcs: An Email Thread Visualization. In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*, pages 211–218. IEEE, 2003.

[96] V. Kešelj and D. Šipka. Pristup izgradnji stemera i lematizora za jezike s bogatom fleksijom i oskudnim resursima zasnovan na obuhvatanju sufiksa. *Infoteka – časopis za digitalnu humanistiku*, 9(1-2):21–31, 2008.

[97] J. S. Kessler. Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 85–90, 2017.

[98] J. Y. Kim, R. A. Calvo, K. Yacef, and N. J. Enfield. A Review on Dyadic Conversation Visualizations - Purposes, Data, Lens of Analysis. *arXiv preprint arXiv:1905.00653*, 2019.

[99] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and Flesch Reading Ease formula) for Navy enlisted personnel. *Institute for Simulation and Training*, 56, 1975.

[100] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2017.

[101] S. Kiritchenko and S. Mohammad. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers)*, pages 465–470, 2017.

[102] B. Klimt and Y. Yang. The Enron Corpus: A New Dataset for Email Classification Research. In *European Conference on Machine Learning*, pages 217–226. Springer, 2004.

[103] M. B. Kmainasi, R. Khan, A. E. Shahroor, B. Bendou, M. Hasanain, and F. Alam. Native vs Non-native Language Prompting: A Comparative Analysis. In *International Conference on Web Information Systems Engineering*, pages 406–420. Springer, 2024.

[104] E. Kochkina, M. Liakata, and I. Augenstein. Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, 2017.

[105] E. Kochkina, M. Liakata, and A. Zubiaga. All-in-one: Multi-task Learning for Rumour Verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, 2018.

[106] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, et al. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861, 2023.

[107] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1143, 1995.

[108] M. Kosanović. *Adjectives Denoting Emotions in English and Serbian: A Cognitive Linguistic Analysis [Pridevi koji označavaju emocije u engleskom i srpskom jeziku: kognitivnolingvistička analiza]*. PhD thesis, University of Belgrade, Faculty of Philology, 2016.

[109] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotic: Emotions in context dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 61–69, 2017.

[110] N. C. Krämer and S. Winter. Impression management 2.0: The relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. *journal of media psychology*, 20(3):106–116, 2008.

[111] D. L. Krebs. The evolution of moral behaviors. In *Handbook of Evolutionary Psychology: Ideas, Issues, and Applications*, pages 337–368. Lawrence Erlbaum Associates Mahwah, NJ, 1998.

[112] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, 2016.

[113] M. S. A. Lee and J. Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

[114] Q. Li, Q. Zhang, and L. Si. eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 855–859, 2019.

[115] W. Li and H. Xu. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749, 2014.

[116] Y. Li and C. Scarton. Revisiting Rumour Stance Classification: Dealing with Imbalanced Data. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 38–44, 2020.

[117] T. Lischetzke and M. Eid. Why Extraverts Are Happier Than Introverts: The Role of Mood Regulation. *Journal of Personality*, 74(4):1127–1162, 2006.

[118] P. Lison and J. Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[119] C. Liu, M. Osama, and A. De Andrade. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6293–6298, 2019.

[120] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[121] Z. Liu, T. Zhang, K. Yang, P. Thompson, Z. Yu, and S. Ananiadou. Emotion detection for misinformation: A review. *Information Fusion*, page 102300, 2024.

[122] A. Ljajić and U. Marovac. Improving sentiment analysis for twitter data by handling negation rules in the Serbian language. *Computer Science and Information Systems*, 16(1):289–311, 2019.

[123] N. Ljubešić, T. Erjavec, and D. Fišer. Corpus-based diacritic restoration for south slavic languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3612–3616, 2016.

[124] N. Ljubešić and D. Lauc. Bertić – The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, 2021.

[125] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[126] R. Mansoor, N. D. Jayasinghe, and M. M. A. Muslam. A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms. In *2021 International Conference on Information Networking (ICOIN)*, pages 327–332. IEEE, 2021.

[127] M. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330, 1993.

[128] A. M. McNutt and G. L. Kindlmann. Improving the Scalability of Interactive Visualization Systems for Exploring Threaded Conversations. In *EuroVis (Posters)*, pages 53–55, 2019.

[129] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, 2013.

[130] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[131] A. Milenković. *Semantics and creative analysis of verbs denoting feelings in the modern Serbian language [Semantička i tvorbena analiza glagola kojima se označavaju osećanja u savremenom srpskom jeziku]*. PhD thesis, University of Belgrade, Faculty of Philology, 2017.

[132] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

[133] M. Mladenović, J. Mitrović, C. Krstev, and D. Vitas. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3):599–620, 2016.

[134] S. Moderc, R. Stanković, A. Tomašević, and M. Škorić. An italian-serbian sentence aligned parallel literary corpus. *Review of the NCD*, 43:78–91, 2023.

[135] S. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, 2018.

[136] S. Mohammad and P. Turney. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34, 2010.

[137] S. M. Mohammad. Word Affect Intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, pages 174–183, 2018.

[138] S. M. Mohammad and S. Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

[139] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.

[140] S. Mukherjee, S. Mukherjee, M. Hasegawa, A. Hassan Awadallah, and R. White. Smart To-Do: Automatic Generation of To-Do Items from Emails. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8680–8689, 2020.

[141] N. Nisar, N. Rakesh, and M. Chhabra. Review on Email Spam Filtering Techniques. *International Journal of Performability Engineering*, 17(2):178–190, 2021.

[142] J. Nivre, M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, et al. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666. European Language Resources Association (ELRA), 2016.

[143] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[144] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE international conference on mobile data management (MDM)*, pages 371–375. IEEE, 2017.

[145] E. Öhman, M. Pàmies, K. Kajava, and J. Tiedemann. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, 2020.

[146] I. Pak and P. L. Teh. Text Segmentation Techniques: A Critical Review. *Innovative Computing, Optimization and Its Applications*, 741:167–181, 2018.

[147] E. W. Pamungkas, V. Basile, and V. Patti. Stance Classification for Rumour Analysis in Twitter: Exploiting Affective Information and Conversation Structure. *arXiv preprint arXiv:1901.01911*, 2019.

[148] S.-H. Park, B.-C. Bae, and Y.-G. Cheong. Emotion recognition from text stories using an emotion embedding model. In *2020 IEEE international conference on big data and smart computing (BigComp)*, pages 579–583. IEEE, 2020.

[149] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[150] J. W. Pennebaker, R. J. Booth, R. L. Boyd, and M. E. Francis. *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates, 2015.

[151] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[152] F. M. Plaza-del Arco, S. M. Jiménez Zafra, A. Montejo Ráez, M. D. Molina González, L. A. Ureña López, and M. T. Martín Valdivia. Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 2021.

[153] R. Plutchik. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[154] L. Pokhun and M. Y. Chuttur. Emotions in texts. *Bulletin of Social Informatics Theory and Application*, 4(2):59–69, 2020.

[155] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38, 2013.

[156] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019.

[157] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332, 2021.

[158] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*, 7:100943–100953, 2019.

[159] L. Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[160] V. Preniqi, I. Ghinassi, J. Ive, C. Saitis, and K. Kalimeri. MoralBERT: a fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 433–442, 2024.

[161] N. Proferes, N. Jones, S. Gilbert, C. Fiesler, and M. Zimmer. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7(2):20563051211019004, 2021.

[162] I. Pujiono, I. M. Agtyaputra, Y. Ruldeviyani, et al. Implementing retrieval-augmented generation and vector databases for chatbots in public services agencies context. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, 10(1):216–223, 2024.

[163] C. Puryear, J. A. Vandello, and K. Gray. Moral panics on social media are fueled by signals of virality. *Journal of Personality and Social Psychology*, 127(1):84–103, 2024.

[164] Y. Qi and Z. Shabrina. Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach. *Social network analysis and mining*, 13(1):31, 2023.

[165] Y. Rao, Q. Li, X. Mao, and L. Wenyin. Sentiment topic models for social emotion mining. *Information Sciences*, 266:90–100, 2014.

[166] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[167] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062):1518–1524, 2011.

[168] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?"Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[169] B. Rujević, M. Kaplar, S. Kaplar, R. Stanković, I. Obradović, and J. Mačutek. Quantitative analysis of syllable properties in Croatian, Serbian, Russian, and Ukrainian. *Language and text: Data, models, information and applications*, pages 55–67, 2021.

[170] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[171] T. Saha, A. P. Patra, S. Saha, and P. Bhattacharyya. A transformer based approach for identification of tweet acts. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[172] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[173] A. E. Samy, S. R. El-Beltagy, and E. Hassanien. A Context Integrated Model for Multi-label Emotion Detection. *Procedia Computer Science*, 142:61–71, 2018.

[174] P. Schober, C. Boer, and L. A. Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, 2018.

[175] H. Schuff, J. Barnes, J. Mohme, S. Padó, and R. Klinger. Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus. In *Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 13–23, 2017.

[176] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[177] K. Sechidis, G. Tsoumakas, and I. Vlahavas. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer, 2011.

[178] Z. Shao, R. Chandramouli, K. Subbalakshmi, and C. T. Boyadjiev. An analytical system for user emotion extraction, mental state modeling, and rating. *Expert Systems with Applications*, 124:82–96, 2019.

[179] A. Sharaff and N. K. Nagwani. Identifying categorical terms based on latent Dirichlet allocation for email categorization. In *Emerging Technologies in Data Mining and Information Security*, pages 431–437. Springer, 2019.

[180] C. Silva and B. Ribeiro. The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 1661–1666. IEEE, 2003.

[181] M. Škorić. Novi jezički modeli za srpski jezik. *Infoteka*, 24, 2024.

[182] E. A. Smith and R. Senter. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, 1967.

[183] M. S. Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18(1):25, 2010.

[184] M. Šošić. SRPOL – A Lexicon Based Framework for Sentiment Strength of Serbian Texts. *Review of the NCD*, 41:58–73, 2022.

[185] M. Šošić and J. Graovac. Effective Methods for Email Classification: Is it a Business or Personal Email? *Computer Science and Information Systems*, 19(3):1155–1175, 2022.

[186] M. Šošić, J. Graovac, and R. Stanković. Building an Emotion Lexicon for Serbian Using Curated Language Resources. *Language Resources and Evaluation*, 2025. Forthcoming.

[187] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[188] B. C. Stahl and D. Wright. Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, 16(3):26–33, 2018.

[189] R. Stanković, M. Košprdić, M. I. Nešić, and T. Radović. Sentiment Analysis of Serbian Old Novels. In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, pages 31–38, 2022.

[190] R. Stanković, C. Krstev, B. Š. Todorović, and M. Škorić. Annotation of the Serbian ELTeC Collection. *Infotheca - Journal for Digital Humanities*, 21(2):43–59, 2022.

[191] R. Stanković, B. Šandrih, C. Krstev, M. Utvić, and M. Škorić. Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3954–3962, 2020.

[192] R. Stanković, M. Škorić, K. Cvetana, and D. Vitas. SrpMD4Tagging – Serbian Morphological Dictionaries for Tagging. version 1.0.0, 2021.

[193] M. Stewart and U. Schultze. Producing solidarity in social media activism: The case of My Stealthy Freedom. *Information and organization*, 29(3):100251, 2019.

[194] C. Strapparava, A. Valitutti, et al. WordNet-Affect: an Affective Extension of WordNet. In *LREC*, volume 4, pages 1083–1086. Lisbon, Portugal, 2004.

[195] H. Suresh and J. Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, 2021.

[196] A. H. Syrjämäki, M. Ilves, P. Isokoski, J. Kiskola, A. Rantasila, T. Olsson, G. Bente, and V. Surakka. Emotionally toned online discussions evoke subjectively experienced emotional responses. *journal of Media Psychology: Theories, Methods, and Applications*, 35(1):55, 2023.

[197] D. Tang, B. Qin, and T. Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.

[198] R. Teper, C.-B. Zhong, and M. Inzlicht. How emotions shape moral behavior: Some answers (and questions) for the field of moral psychology. *Social and Personality Psychology Compass*, 9(1):1–14, 2015.

[199] J. Thiergart, S. Huber, and T. Übellacker. Understanding Emails and Drafting Responses – An Approach Using GPT-3. *arXiv preprint arXiv:2102.03062*, 2021.

[200] M. Thway, J. Recatala-Gomez, F. S. Lim, K. Hippalgaonkar, and L. W. T. Ng. Battling Botpoop using GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbots Impact on Learning. *arXiv preprint arXiv:2406.07796*, 2024.

[201] Y. Torii, D. Das, S. Bandyopadhyay, and M. Okumura. Developing Japanese WordNet affect for analyzing emotions. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 80–86, 2011.

[202] J. Trager, A. S. Ziabari, A. M. Davani, et al. The Moral Foundations Reddit Corpus. *arXiv preprint arXiv:2208.05545*, 2022.

[203] Z. Tufekci. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *Proceedings of the Eighth International AAAI conference on Weblogs and Social Media*, volume 8, pages 505–514, 2014.

[204] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. *arXiv preprint arXiv:1908.08962*, 2019.

[205] P. Valdesolo. Getting emotions right in moral psychology. *Atlas of moral psychology*, pages 88–93, 2018.

[206] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[207] D. Vera, O. Araque, and C. A. Iglesias. GSI-UPM at IberLEF2021: Emotion Analysis of Spanish Tweets by Fine-tuning the XLM-RoBERTa Language Model. In *IberLEF@ SEPLN*, pages 16–26, 2021.

[208] D. Vitas, R. Stanković, and C. Krstev. The Many Faces of SrpKor. In *Proceedings of the International Conference South Slavic Languages in the Digital Environment – JuDig*. University of Belgrade, Faculty of Philology, 2024. Abstract.

[209] P. Vossen. *EuroWordNet: A multilingual database with lexical semantic networks*. Springer, 1998.

[210] Y. Wang, Y. Hou, W. Che, and T. Liu. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, 11:1611–1630, 2020.

[211] A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

[212] J. Weizenbaum. ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[213] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2019.

[214] L. Williams, M. Arribas-Ayllon, A. Artemiou, and I. Spasić. Comparing the utility of different classification schemes for emotive language analysis. *Journal of Classification*, 36:619–648, 2019.

[215] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, 1994.

[216] R. Xi and M. P. Singh. Moral Sparks in Social Media Narratives. *arXiv preprint arXiv:2310.19268*, 2024.

[217] E. Yang, J. Amar, J. H. Lee, B. Kumar, and Y. Jia. The Geometry of Queries: Query-Based Innovations in Retrieval-Augmented Generation for Healthcare QA. *arXiv preprint arXiv:2407.18044*, 2025.

[218] R. Yang, W. Xie, C. Liu, and D. Yu. BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 1090–1096, 2019.

[219] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, and R. Kurzweil. Multilingual Universal Sentence Encoder for Semantic Retrieval. In A. Celikyilmaz and T.-H. Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, 2020.

[220] J. Yu, L. Marujo, J. Jiang, P. Karuturi, and W. Brendel. Improving Multi-label Emotion Classification via Sentiment Classification with Dual Attention Transfer Network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[221] R. Zhang and J. Tetreault. This Email Could Save Your Life: Introducing the Task of Email Subject Line Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 446–456, 2019.

[222] S. Zhang, A. Celikyilmaz, J. Gao, and M. Bansal. EmailSum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6895–6909, 2021.

[223] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

[224] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, and K. Shu. Mining Dual Emotion for Fake News Detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476, 2021.

[225] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, 2021.

[226] C. Ziems, J. Yu, Y.-C. Wang, A. Halevy, and D. Yang. The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, 2022.

[227] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen. The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):1–29, 2018.

Prilozi

A. Pridruženi atributi klasifikacije

Tabela A.1: *Detaljni opis atributa po grupama - Leksički, Konverzacioni, Atributi Izražajnosti, Moralnosti i Emocionalnosti*

	Atribut	Opis	Tip	Zadatak
LexAtr	content_words_count	Broj reči u sadržaju	I	PL/TD/IG
	subject_words_count	Broj reči u naslovu	I	PL
	content_length	Broj karaktera u sadržaju	I	PL/TD/IG
	subject_length	Broj karaktera u naslovu	I	PL
	sentences	Broj rečenica u sadržaju	I	PL/TD/IG
	syllable_count	Broj slogova u sadržaju	I	PL/TD/IG
	newlines	Broj novih redova u sadržaju	I	PL/TD/IG
	avg_word_length	Prosečna dužina reči	D	PL/TD/IG
	avg_sentence_length	Prosečna dužina rečenice	D	PL/TD/IG
	noun_phrases_ratio	Prosečan broj imeničkih izraza	D	PL/TD/IG
	difficult_words_ratio	Prosečan broj „teških“ reči (reči koje imaju više od 2 sloga)	D	PL/TD/IG
	words_density	Gustina reči u tekstualnoj sekvenci	D	PL/TD/IG
	sentences_density	Gustina rečenica u tekstualnoj sekvenci	D	PL/TD/IG
	ASPW	Prosečan broj slogova po reči	D	PL/TD/IG
	ASPS	Prosečan broj slogova po rečenici	D	PL/TD/IG
	nouns_ratio	Prosečan broj imenica u tekstualnoj sekvenci	D	TD/IG
	adj_ratio	Prosečan broj prideva u tekstualnoj sekvenci	D	TD/IG
	adv_ratio	Prosečan broj priloga u tekstualnoj sekvenci	D	TD/IG
	verb_ratio	Prosečan broj glagola u tekstualnoj sekvenci	D	TD
IndAtr	is_plain	Identifikator da li je poruka inicijalna objava	I	TD/IG
	is_retweet	Identifikator da li je poruka podeljena	I	TD/IG
	is_mention	Identifikator da li je poruka komentar ili referencira korisnika	I	TD/IG
	business_indicator	Prosečan broj poslovnih reči u tekstualnoj sekvenci	D	PL
	acronyms_indicator	Prosečan broj akronima u tekstualnoj sekvenci	D	PL
PncAtr	dots_ratio	Prosečan broj tačaka u odnosu na ukupan broj interpunkcijskih znakova	D	PL/TD/IG
	question_marks_ratio	Prosečan broj znakova pitanja u odnosu na ukupan broj interpunkcijskih znakova	D	PL/TD/IG
	exclamation_marks_ratio	Prosečan broj uzvičnika u odnosu na ukupan broj interpunkcijskih znakova	D	PL/TD/IG
	hash_tags_ratio	Prosečan broj znakova „#“ u odnosu na ukupan broj interpunkcijskih znakova	D	PL/TD/IG
	reference_tags_ratio	Prosečan broj znakova „@“ u odnosu na ukupan broj interpunkcijskih znakova	D	PL/TD/IG
NERAtr	names_ratio	Prosečan broj ličnih imena u tekstualnoj sekvenci	D	PL/TD/IG
	org_ratio	Prosečan broj imena organizacija u tekstualnoj sekvenci	D	PL/TD/IG
	num_ratio	Prosečan broj reči koje sadrže brojeve u tekstualnoj sekvenci	D	PL/TD/IG
	conn_ratio	Prosečan broj konektora (predloga, veznika) u tekstualnoj sekvenci	D	PL/TD/IG
	months_ratio	Prosečan broj naziva meseci u godini u tekstualnoj sekvenci	D	PL/TD/IG
	days_ratio	Prosečan broj naziva dana u nedelji u tekstualnoj sekvenci	D	PL/TD/IG

Nastavak na sledećoj strani

Tabela A.1 – *Nastavak sa prethodne strane*

	Atribut	Opis	Tip	Zadatak
	emails_ratio	Prosečan broj adresa elektronske pošte u tekstualnoj sekvenci	D	PL/TD/IG
	url_ratio	Prosečan broj internet adresa u tekstualnoj sekvenci	D	PL/TD/IG
ConAtr	free_domains_ratio	Prosečan broj besplatnih domena elektronske pošte u odnosu na ukupan broj domena	D	PL
	number_of_recipients	Broj primalaca poruke	I	PL/TD/IG
	recipient_dom_coherency	Koherentnost domena primalaca poruke (u okviru istog domena, izvan, mešovito)	I	PL
	depth	Dubina poruke u konverzacionom nizu	I	PL/TD
MorAtr	care_p	Verovatnoća da poruka pripada kategoriji *care*	D	PL/TD/IG
	sanctity_p	Verovatnoća da poruka pripada kategoriji *sanctity (purity)*	D	PL/TD/IG
	authority_p	Verovatnoća da poruka pripada kategoriji *authority*	D	PL/TD/IG
	loyalty_p	Verovatnoća da poruka pripada kategoriji *loyalty*	D	PL/TD/IG
	fairness_p	Verovatnoća da poruka pripada kategoriji *fairness*	D	PL/TD/IG
	care_sent	Sentiment poruke za kategoriju *care*	D	PL/TD/IG
	sanctity_sent	Sentiment poruke za kategoriju *sanctity (purity)*	D	PL/TD/IG
	authority_sent	Sentiment poruke za kategoriju *authority*	D	PL/TD/IG
	loyalty_sent	Sentiment poruke za kategoriju *loyalty*	D	PL/TD/IG
	fairness_sent	Sentiment poruke za kategoriju *fairness*	D	PL/TD/IG
	moral_nonmoral_ratio	Odnos broja moralnih i preostalih reči u poruci	D	PL/TD/IG
EmoAtr	trust	Intenzitet poruke za kategoriju *trust*	D	PL/TD/IG
	joy	Intenzitet poruke za kategoriju *joy*	D	PL/TD/IG
	anger	Intenzitet poruke za kategoriju *anger*	D	PL/TD/IG
	disgust	Intenzitet poruke za kategoriju *disgust*	D	PL/TD/IG
	sadness	Intenzitet poruke za kategoriju *sadness*	D	PL/TD/IG
	fear	Intenzitet poruke za kategoriju *fear*	D	PL/TD/IG
	surprise	Intenzitet poruke za kategoriju *surprise*	D	PL/TD/IG
ExpAtr	ARI	Automatski indeks čitljivosti	D	PL/TD/IG
	FRES	Flešova ocena lakoće čitanja	D	PL/TD/IG
	LWM	Linsearova mera čitljivosti	D	PL/TD/IG
	polarity	Intenzitet sentimenta poruke	D	PL/TD/IG
	subjectivity	Intenzitet subjektivnosti poruke	D	PL/TD/IG

Tabela A.2: *Pristup jedne promenljive primenjen u analizi značaja atributa na zadacima PL/TD/IG. Prikazani su statistički značajni atributi (p-vrednost < 0.005)*

PL		**TD**		**IG**	
Atribut	F_s	Atribut	F_s	Atrubut	F_s
recipient_dom_coherency	510.08	question_marks_ratio	468.88	avg_sentence_length	240.39
free_domains_ratio	459.74	sentences	387.91	ASPS	219.51
joy	369.24	trust	338.62	fear	213.54
difficult_words_ratio	255.37	adv_ratio	322.64	fairness_sent	207.13
ASPS	219.92	spaces_ratio	275.13	LWM	206.49
LWM	203.53	noun_ratio	247.13	anger	188.59
ARI	192.83	content_lex_count	245.26	authority_p	173.77
business_indicator	178.55	syllable_count	232.67	names_lex_ratio	169.14
exclamation_marks_ratio	172.30	content_lex_length	202.90	anticipation	160.90
avg_sentence_length	144.63	noun_phrases_ratio	187.58	care_p	150.28
avg_word_length	134.04	authority_sent	156.30	disgust	148.90
subject_lex_length	127.22	care_sent	150.55	polarity	142.73

Nastavak na sledećoj strani

Tabela A.2 – *Nastavak sa prethodne strane*

PL		TD		IG	
Atribut	F_s	Atribut	F_s	Atrubut	F_s
surprise	121.76	words_density	144.93	authority_sent	123.72
FRES	111.07	hash_tags_ratio	135.17	loyalty_sent	114.44
adv_ratio	97.38	conns_lex_ratio	128.53	fairness_p	113.41
disgust	92.41	ARI	122.83	care_sent	108.70
loyalty_sent	89.01	loyalty_sent	116.96	moral_nonmoral_ratio	104.74
noun_ratio	85.79	fear	114.19	ARI	98.22
ASPW	85.77	fairness_sent	113.61	spaces_ratio	87.74
subjectivity	62.40	joy	108.45	sadness	86.72
orgs_lex_ratio	60.12	sanctity_sent	99.14	hash_tags_ratio	86.06
noun_phrases_ratio	58.15	polarity	74.14	subjectivity	77.56
adj_ratio	56.06	anger	72.74	noun_ratio	69.97
subject_lex_count	53.26	newlines	63.73	avg_word_length	68.66
sanctity_sent	52.82	difficult_words_ratio	63.08	sanctity_sent	65.91
polarity	47.70	subjectivity	57.17	is_mention	62.13
trust	37.61	disgust	51.95	loyalty_p	61.65
fairness_sent	37.26	FRES	48.17	is_plain	60.97
moral_nonmoral_ratio	37.09	avg_word_length	39.75	noun_phrases_ratio	55.39
anticipation	33.56	ASPW	32.79	difficult_words_ratio	52.88
sanctity_p	27.85	exclamation_marks_ratio	27.86	verb_ratio	51.97
reference_tags_ratio	23.15	reference_tags_ratio	19.65	sentences_density	50.65
sentences_density	21.97	names_lex_ratio	18.89	conns_lex_ratio	47.89
anger	21.73	orgs_lex_ratio	18.83	FRES	41.03
recipients_count	21.20	care_p	17.87	newlines	40.60
authority_sent	17.74	surprise	17.73	reference_tags_ratio	36.08
syllable_count	15.35	verb_ratio	16.55	content_lex_count	32.29
spaces_ratio	15.33	moral_nonmoral_ratio	15.67	content_lex_length	31.74
numbers_lex_ratio	15.17	is_plain	14.72	syllable_count	31.16
question_marks_ratio	14.74	numbers_lex_ratio	14.29	ASPW	25.77
authority_p	14.36	authority_p	14.08	trust	25.77
words_density	13.21	is_mention	13.96	numbers_lex_ratio	24.02
sadness	12.85	sadness	13.25	adj_ratio	23.57
content_lex_length	12.83	ASPS	12.70	dots_ratio	20.47
care_p	11.92	adj_ratio	10.50	words_density	20.24
names_lex_ratio	11.65	dots_ratio	8.99	question_marks_ratio	10.14
acronyms_indicator	11.11	sentences_density	7.29	surprise	6.62
newlines	10.85	anticipation	7.18	is_retweet	6.13
dots_ratio	8.69	sanctity_p	6.71		
conns_lex_ratio	8.14	fairness_p	5.23		
care_sent	8.13				
content_lex_count	7.97				

B. Anotacione šeme za prepoznavanje emocionalnosti i moralnosti

Sa Tviter i Redit socijalnih platformi prikupljene su tektualne poruke napisane na srpskom jeziku. U skupu podataka se nalaze inicijalne objave, kao i komentari na objave. Cilj je da ove poruke na što precizniji način obeležimo na prisustvo emocionalnog afekta i moralne vrednosti, ukoliko se one u tekstualnom sadržaju mogu prepoznati.

1. **Obeležavanje na prisustvo emocionalnog afekta** – Poruke se prema prisustvu emocionalnog afekta klasifikuju u 8 + 1 kategorija sa nazivima na engleskom jeziku: anger, anticipation, disgust, fear, joy, sadness, surprise, trust + neutral, prema modelu koji je postavio američki psiholog Robert Plutčik. Plutčikov model prepoznaje 8 emocionalnih kategorija, sa tri nivoa emocionalnih intenziteta u svakoj kategoriji, pri čemu su osnovne emocije postavljene u sredini svake kategorije, sa okvirnim definicijama zadatim na sledeći način:
   - **`anger`** (ljutnja) – emocije povezane sa osećajem ljutnje, nezadovoljstva i frustracije.
   - **`anticipation`** (iščekivanje) – emocije koje se fokusiraju na iščekivanje određenog događaja.
   - **`disgust`** (odvratnost) – emocije povezane sa izbegavanjem i odbacivanjem.
   - **`fear`** (strah) – negativna osećanja povezana za uplašenost i anksioznost.
   - **`joy`** (radost) – pozitivna emocija koja obuhvata sreću i satisfakciju.
   - **`surprise`** (iznenađenje) – emocije povezane sa izenenađenjem i čuđenjem.
   - **`sadness`** (tuga) – negativne emocije povezane sa osećajem gubitka i razočarenja.
   - **`trust`** (poverenje) – osećanja koja su povezana sa mirom i povezivanjem.
2. **Obeležavanje na prisustvo moralne vrednosti** – Poruke se prema prisustvu moralne vrednosti klasifikuju u 10 + 1 kategorija sa nazivima na engleskom jeziku: *care*, *harm*, *loyalty*, *betrayal*, *authority*, *subversion*, *sanctity*, *degradation* + *non-moral*, prema modelu koji definiše MFT. Model MFT prepoznaje 5 osnovnih moralnih vrednosti, predstavljenih dihotomnim parovima suprotstavljenih polariteta, sa okvirnim opisima njihovog značenja datim na sledeći način:
   - **`care/harm`** (briga/povreda) – senzitivnost prema patnji i dobrobiti drugih, zasnovana na empatiji i emocionalnoj povezanosti. Motiviše ponašanja koja uključuju zaštitu, negu i pomoć.
   - **`fairness/cheating`** (pravednost/prevara) – težnja ka pravičnim i ravnopravnim odnosima među pojedincima, koja je utemeljena na principima jednakosti, reciprociteta i poštenja.
   - **`loyalty/betrayal`** (lojalnost/izdaja) – identifikacija i posvećenost sopstvenoj grupi. Podrazumeva spremnost na žrtvu u cilju očuvanja kohezije i stabilnosti grupe.
   - **`authority/subversion`** (autoritet/subverzija) – poštovanje legitimnih autoriteta, društvenih hijerarhija i tradicije, koje doprinosi očuvanju reda, discipline i institucionalne stabilnosti.
   - **`sanctity/degradation`** (svetost/degradacija) – ideal čistote i uzdržavanja od telesnih i moralno nepoželjnih uticaja, često povezan sa religijskim normama i konceptima duhovnog uzdizanja.

Da bi se olakšao i ubrzao proces anotacije, poruke su predobeležene nekim od postojećih tehnika i modela računarske lingvistike. S obzirom da ove metode nisu najtačnije iz potpuno legitimnih razloga (znanja preuzeta iz drugih jezika, preciznosti alata, modela i drugih), dodeljena obeležja je potrebno proveriti. Svakoj poruci u skupu podataka su pridružene tri obeležja sa nazivima label_1, label_2 i label_3. Obeležja se mogu poklapati, preklapati, biti potpuno različite ili u pojedinim slučajevima potpuno nedostajati. Dodatno, svakoj poruci je dodeljeno obeležje annotator_label koja je predviđena za opcioni unos odgovarajuće labele, koja se konstruiše od strane anotatora i sadrži jednu ili više emocionalnih kategorija odvojenih zarezom.

Precizne instrukcije za korake u toku obeležavanja:

- Pažljivo pročitati sadržaj poruke, uzimajući u obzir tekstualni sadržaj i pridružene emocionalne simbole (emotikone, smajlije) u cilju što tačnijeg prepoznavanja emocionalniog/moralnog tona poruke koji je njen kreator želeo da pošalje.
- Uporediti prepoznati emocionalni/moralni ton poruke sa kategorijama iz anotacione šeme i odabrati jednu ili više emocionalnih/moralnih kategorija koje joj najviše odgovaraju.

- Proveriti da li neka od pridruženih obeležja sadrži odgovarajuću kategorizaciju i ukoliko to jeste slučaj, obarati njen naziv iz liste mogućih vrednosti u polju za validaciju obeležja.
- Ukoliko nijedno od pridruženih obeležja ne sadrži odgovarajuće kategorije ili su one samo delimično tačne, napraviti novo obeležje u polju ručno kreiranje obeležja i odabrati njen naziv u polju za validaciju obeležja.
- Ukoliko u sadržaju poruke nije prepoznat emocionalni afekt/moralna vrednost, dodeliti joj obeležje sa kategorijom *neutral*, odnosno *non-moral* (što je takođe podržano od nekih automatski dodeljenih obeležja).
- Ukoliko sadržaj poruke nije napisan na srpskom jeziku, odabrati odgovarajući jezik u polju za izbor jezika.

Prilikom obeležavanja poruka, anotatorima je preporučeno da provere jezik na kome je poruka napisana (unapred postavljeno na srpski), čime je obezbeđena dodatna provera tačnosti u pogledu jezika koji se analizira. U cilju što preciznije i konzistentnije kategorizacije poruka, anotatorima je preporučeno da svako dodeljeno obeležje sadrži najviše tri kategorije. Ovo ograničenje je omogućilo da svaka poruka bude klasifikovana prema najrelevantnijim dimenzijama emocionalnog afekta, odnosno moralne vrednosti. Takođe, anotatorima je naglašeno da obeležavanje vrše isključivo na osnovu eksplicitnog značenja sadržaja poruke, bez uključivanja šireg konteksta ukoliko je prepoznat.

Dodatno, u cilju daljeg unapređenja polu-automatizovanog procesa anotacije, kreirane su liste osnovnih ključnih reči za emocionalne i moralne kategorije iz anotacionih šema. Prvi korak je uključio identifikaciju glavnih emocija ili moralnih pojmova na osnovu psiholoških modela, kao što su Plutčikov krug emocija ili MFT. Prema Plutčikovom modelu identifikovani su svi intenziteti, odnosno podkategorije, u emocionalnim stubovima, kao i kombinacije emocija koje formiraju složenije emocionalne reakcije, obeležene kao međukategorije. Na sličan način, iz šeme za anotaciju moralnih vrednosti identifikovane su sve moralne kategorije zajedno sa sentimentom i značenjem koje imaju. Za svaku kategoriju (podkategoriju, međukategoriju) su prema njenom značenju prikupljeni sinonimi, srodne reči i fraze koje često prate ili opisuju te emocije ili moralne pojmove. Nakon početne pripreme, lista je unapređena na osnovu povratnih informacija od strane anotatora, čime se osigurala uključenost termina koji pokrivaju različite manifestacije kategorija. Rezultujuće liste ključnih reči zasnovane na imenicama (koje podrazumevaju uključivanje njihovih varijacija i derivacija) za prepoznavanje emocionalnih (pogledati tabelu B.1) i moralnih kategorija (pogledati tabelu B.2) su služile kao smernica anotatorima prilikom kategorizacije i identifikovanja emocija i moralnih termina u tekstovima, čime je obezbeđena doslednost i tačnost tokom procesa anotacije.

Tabela B.1: *Izdvojene ključne reči u srpskom jeziku za emocionalne podkategorije i međkategorije prema Plutčikovom modelu*

Podkategorija Ključne Reči	Intenzitet - Podkategorija	Kategorija	Međukategorija	Međukategorija Ključne Reči
interesovati se, zainteresovati se	1 - zainteresovanost (Interest)		AGRESIVNOST (AGGRESSIVENESS)	agresivnost, ratobornost, grubost, napadnost
očekivanje, iščekivanje, predviđanje, nestrpljenje, naslućivanje	2 - IŠČEKIVANJE (ANTICIPATION)	IŠČEKIVANJE (ANTICIPATION)		
opreznost, budnost, predostrožnost	3 - opreznost (Vigilance)		OPTIMIZAM (OPTIMISM)	optimizam, optimističan, vedrina, entuzijazam
spokojstvo, mirnoća, ležernost smirenost	1 - spokoj (Serenity)			
radost, sreća, slavlje, satisfakcija, razdraganost	2 - RADOST (JOY)	RADOST (JOY)		
ushićenost, oduševljenje, ekstaza osećanje beskrajne radosti, trans	3 - oduševljenje (Ecstasy)		LJUBAV (LOVE)	ljubav, voleti, privženost, zaljubljenost, prijateljstvo
prihvatanje, primanje, izbor	1 - prihvatanje (Acceptance)			
verovati, vera, nadati se, čuvati, imati poverenja	2 - POVERENJE (TRUST)	POVERENJE (TRUST)		
divljenje	3 - divljenje (Admiration)		POKORNOST (SUBMISSION)	poniznost, pokornost, potčinjenost
bojazan, uznemirenost	1 - uznemirenost (Apprehension)			
strah, uplašenost, strepnja, anksioznost	2 - STRAH (FEAR)	STRAH (FEAR)		
užas	3 - užas (Terror)		STRAHOPOŠTOVANJE (AWE)	strahopoštovanje, ustreptalost, respekt
odvraćanje pažnje, ometanje	1 - ometanje (Distraction)			
iznenađenje, iznenaditi	2 - IZNENAĐENJE (SURPRISE)	IZNENAĐENJE (SURPRISE)		
čuđenje, začuđenost, zadivljenost, zapanjenost	3 - čuđenje (Amazement)		NEODOBRAVANJE (DISAPPROVAL)	neodobravanje, neslaganje, osuda
zamišljenost, briga, zapitanost	1 - zamišljenost (Pensiveness)			
tuga, razočarenje, gubitak	2 - TUGA (SADNESS)	TUGA (SADNESS)		
bol, žalost, beda, jad	3 - patnja (Grief)		POKAJANJE (REMORSE)	kajanje, žaljenje, griža savesti
dosada, gnjavaža, čamotinja	1 - dosađivanje (Boredom)			
izbegavanje, odbacivanje	2 - GAĐENJE (DISGUST)	GAĐENJE (DISGUST)		
gnušanje, odvratnost	3 - gnušanje (Loathing)		PREZIR (CONTEMPT)	prezir, nepoštovanje, nadmenost, oholost
muka, sekiracija, iritacija	1 - iritacija (Annoyance)			
ljutnja, nezadovoljstvo, frustracija, gnev, srdžba	2 - LJUTNJA (ANGER)	LJUTNJA (ANGER)		
bes, žestina, jarost	3 - bes (Rage)		AGRESIVNOST (AGGRESSIVENESS)	agresivnost, ratobornost, grubost, napadnost

Tabela B.2: *Izdvojene ključne reči u srpskom jeziku za svaku od moralnih kategorija prema sentimentu kategorije*

Sentiment	Moralna Kategorija	Ključne Reči
VRLINA (VIRTUE)	BRIGA (CARE)	briga, empatija, razmatranje, osetljivost, rastuženost, ohrabrenje, pažnja, solidarnost, brižljivost, saosećanje, nesebičnost, razumevanje, podrška, ljubaznost, tolerancija, razumevanje, saosećajnost, nesebičnost, razmatranje
MANA (VICE)	POVREDA (HARM)	povreda, bol, bes, nasilje, nepravda, žrtva, agresija, sukob, tuga, ljutnja, neprijateljstvo, ogorčenje, konflikt, šok, oštećenje, povređenost, oštećenje, gnev, muka, očaj
VRLINA	PRAVEDNOST (FAIRNESS)	pravda, fer, ispravnost, etika, vrednost, zakon, moral, dostojanstvo, ravnopravnost, poštenje, integritet, odgovornost, transparentnost, solidarnost, princip, pravilnost, nepristrasnost, moralnost
MANA	VARANJE (CHEATING)	varanje, obmana, prevara, lukavstvo, laž, neiskrenost, manipulacija, trik, nevera, licemerje, prećutkivanje, manipulacija, prikrivanje, prevarant, muljanje, falsifikovanje, podvala, zavaravanje
VRLINA	LOJALNOST (LOYALTY)	lojalnost, vernost, poverenje, posvećenost, odanost, iskrenost, pouzdanost, diskrecija, predanost, sigurnost, privrženost, čvrstina, nepokolebljivost, stabilnost
MANA	IZDAJA (BETRAYAL)	izdaja, prevara, nepoštovanje, neverstvo, neiskrenost, laž, izdajnik, varalica, licemerje, dvoličnost, prevrtljivost, prevrtljivost, nečasnost
VRLINA	AUTORITET (AUTHORITY)	autoritet, poštovanje, poslušnost, vođa, lider, nadređeni, ugled, moć, uvaženost, respekt, autoritarnost, dominacija, vođstvo, hijerarhija, ovlašćenje, upravljanje, komanda
MANA	SUBVERZIJA (SUBVERSION)	subverzija, revolt, pobuna, protest, anarhija, opozicija, otpor, kontra, rezistencija, ustanak, neposlušnost, delikventnost, obaranje, rušenje, bunt
VRLINA	SVETOST (PURITY)	svetost, božanstvo, bogoljublje, religioznost, religija, posvećenje, obožavanje, molitva, hram, oltar, obred, liturgija, sakralnost, svetost, svetinja, poštovanje, čestitost, nevinost, čistota, počast, vrednost, dostojanstvo, poštovanje, integritet, posvećenost, duhovnost, svečanost, poštovanje, autentičnost
MANA	DEGRADACIJA (DEGRADATION)	degradacija, poniženje, propadanje, unazađivanje, osmišljavanje, osvajanje, razaranje, srozavanje, sluđivanje, ponižavanje, propadanje, dekadentnost

C. Inženjering instrukcija

Inženjering instrukcija je proces dizajniranja i optimizacije ulaznih instrukcija koje se koriste za interakciju sa LLM. Cilj inženjeringa instrukcija jeste da se modelu pruže jasne i precizne instrukcije kako bi generisao tačne, relevantne i korisne odgovore. Ovaj proces postaje ključan u radu sa LLM, jer dobro osmišljena instrukcija može značajno poboljšati performanse modela bez dodatnog treniranja ili prilagođavanja modela. Instrukcije mogu biti jednostavne strukture, kao što su zahtevi za kreiranje sadržaja, sažimanje informacija ili rešavanje specifičnih problema, ali mogu biti i složenije strukture, kada se kombinuju instrukcije za različite zadatke. Na primer, inženjering instrukcija može uključivati formulisanje zahteva sa više detalja, korišćenje specifičnih ključnih reči ili pružanje konteksta koji pomaže modelu da generiše preciznije odgovore. U okviru ovog istraživanja tehnike inženjeringa instrukcija su korišćene za:

1. Izgradnju emocionalnog rečnika za srpski jezik - LLM - Čet-GPT, u okviru koga su rešavani sledeći zadaci:
   - Prevođenje i obeležavanje vrsta reči.
   - Obeležavanje reči na prisustvo emocionalnog afekta.
   - Generisanje sinonima za datu reč.
   - Generisanje korpusa paralelnih rečenica na srpskom i engleskom jeziku.
2. Evaluaciju i doobučavanje LLM za prepoznavanje emocionalne kategorije u tekstualnoj sekvenci na srpskom jeziku - LLM: LLaMA Instruct.
3. Obeležavanje tekstualnih sekvenci na srpskom jeziku na prisustvo moralnih vrednosti - LLM: Falcon-7B-Instruct.

U okviru izgranje emocionalnog rečnika za srpski jezik, korišćene su tehnike Čet-GPT inženjeringa instrukcija za rešavanje četiri različita lingvistička problema, koji uključuju prevod pojedinačne reči, obeležavanje (reč i rečenice) i generisanje teksta (sinonimi i paralelni segmenti teksta). Za svaki od zadataka, upiti su pažljivo dizajnirani da generišu rezultate u željenim formatima, kao što su pojedinačne reči ili rečenice odvojene zarezom. Dodatno, svaki od zadataka je imao različit broj iteracija i vrednosti parametara da bi se zadovoljio kvalitet generisanih izlaznih rezultata (pogledati tabele C.1 i C.4). Naročito:

- Prevod reči i PoS obeležavanje izvedeni su kroz tri nezavisne iteracije. U jednoj sesiji, engleska reč sa dodeljenim PoS obeležjem je prevedena na srpski jezik i dodeljeno joj je PoS obeležje. Koristeći većinsko glasanje kroz ponovljene iteracije, odabran je tačan prevod i PoS obeležje prevoda.
- Višeznačna kategorizacija afekta reči izvedena je kroz tri nezavisne iteracije koristeći Plutčikov skup kategorija emocija. Konačno obeležje LLM za svaku reč je određena kao spoj kategorija koje se najčešće pojavljuju u generisanim obeležjima u tri različite iteracije.
- Za sve zadatke odabran je model *gpt-3.5-turbo*.
- Nivo temperature je postavljen na 0,1 za zadatke koji zahtevaju deterministički izlaz (prevod, kategoričko obeležje), a na 0,7 za kreativnije rezultate kao što je generisanje sinonima ili rečenica.

Tehnika inženjeringa instrukcija u okviru ovog rada je korišćena i za doobučavanje LLM korišćenjem specijalno dizajniranih i optimizovanih instrukcija sa kontekstualnim informacijama koje usmeravaju model da pravilno interpretira postavljeni zadatak. Zadatak koji se rešavao na ovaj način jeste kategorizacija tekstualnih sekvenci na srpskom jeziku u predefinisane kategorije emocionalnog afekta. Za ovaj zadatak je korišćen LLaMA 3 Instruct model koji je prethodno doobučen za prepoznavanje korisničkih instrukcija (pogledati tabelu C.2).

Inženjering instrukcija je, takođe, korišćen na zadatku obeležavanja tekstualnih sekvenci na srpskom jeziku za detektovanje moralnog konteksta, odnosno prepoznavanja etičkih stavova izraženih u tekstualnim sadržajima. Zadatak je rešavan upotrebom Falcon-7B-Instruct LLM i pažljivo dizajniranim instrukcijama postavljenim na engleskom jeziku (pogledati tabelu C.3). Korišćenje engleskog jezika kao posrednika za anotaciju tekstova na srpskom omogućava modelima treniranim na engleskom da razumeju instrukcije i generišu rezultate u engleskim kategorijama, dok analiziraju srpski tekst [103]. Ovaj pristup olakšava primenu naprednih jezičkih modela, iako može naići na poteškoće pri prepoznavanju lokalnih izraza karakterističnih za

Tabela C.1: *Čet-GPT inženjering instrukcija na zadacima za kreiranje emocionalnog rečnika - prevođenje i anotacija teksta*

Parametar	Vrednost
Zadatak:	$lema_{Sr}$-PoS prevođenje
Model:	gpt-3.5-turbo
Iteracija:	3
Temperatura:	0.1
Sistemska instrukcija:	„You are a language translator assistant. Generate a translation of the English word provided with the assigned part of speech tag in Serbian. Provide an exact answer.“
Korisnička instrukcija:	„Please translate the word with the assigned part of speech into the Serbian language.“
Zadatak:	$lema_{Sr}$-PoS obeležavanje emocionalnog afekta
Model:	gpt-3.5-turbo
Iteracija:	3
Temperatura:	0.1
Sistemska instrukcija:	„You are an emotional affect annotation assistant. Annotate provided Serbian words with assigned part of speech tags with one or multiple discrete emotional categories from the following list: [anger, trust, fear, sadness, joy, anticipation, surprise, disgust, neutral]. Separate the outputs with a comma.“
Korisnička instrukcija:	„Please annotate the Serbian word with the assigned part of the speech tag with one or multiple categories of emotional affects.“

Tabela C.2: *Primena inženjeringa instrukcija na zadatku doobučavanja modela za prepoznavanje emocionalnog afekta u tekstovima srpskom jeziku*

Parametar	Vrednost
Zadatak:	Obeležavanje teksta
Model:	LLaMA 3 Instruct
Iteracija:	1
Temperatura:	0.1
Instrukcija:	„You are a language annotation assistant. Your task is to: Classify the text according to their emotional tone into one or multiple emotional categories (maximum 3) from the list: [anger, anticipation, disgust, fear, joy, neutral, sadness, surprise, trust], and return the answer as the emotional label containing assigned categories separated by a comma. text: data["text"] label: data["label"] “

srpski jezik. Obeležja dobijena na ovakav način korišćena su u cilju poboljšanja efikasnosti obeležavanja većih tekstualnih korpusa i pažljivo su proverena u procesu ručnog obeležavanja i evaluacije .

Korišćenje LLM za zadatke na srpskom jeziku suočava se s nekoliko izazova, s obzirom na to da srpski jezik još uvek nije zvanično podržan u mnogim popularnim modelima, iako je moguća interakcija sa ovim modelima na srpskom jeziku. Potencijalni izazov na zadacima koji su rešavani u okviru ovog istraživanja je prevaziđen korišćenjem instrukcija na engleskom jeziku i ponovljenim iteracijama na zadacima koji zahtevaju veću tačnost u generisanim odgovorima. Dodatno, najveći broj rešavanih zadataka od modela zahteva kategorizaciju podataka (PoS obeležja, emocionalni afekt, moralna kategorija) sa konačnim obeležjima datim na engleskom jeziku. Na zadacima generisanja teksta na srpskom jeziku, poput generisanja sinonima za zadatu reč, dobijeni odgovori se koriste za evaluaciju sposobnosti modela da izvrši postavljeni zadatak, kao i za proveru postojećih lista sinonima iz srpskih rečnika. Na kraju, svi rezultati generisani uz pomoć LLM pažljivo su provereni i ispitana je njihova tačnost.

Tabela C.3: *Primena inženjeringa instrukcija na zadatku obeležavanja moralne vrednosti*

Parametar	Vrednost
Zadatak:	Anotacija teksta
Model:	Falcon-7B-Instruct
Iteracija:	1
Temperatura:	0.1
Instrukcija:	„Given this text, decide what its moral category is. Categories are the following: [care, harm, fairness, cheating, loyalty, betrayal, authority, subversion, purity, degradation, non-moral] Text: message Category: “

Tabela C.4: *Čet-GPT inženjeringa instrukcija na zadacima za kreiranje emocionalnog rečnika - generisanje teksta*

Parametar	Vrednost
Zadatak:	lema_{Sr}-PoS generisanje sinonima
Model:	gpt-3.5-turbo
Iteracija:	1
Temperatura:	0.7
Sistemska instrukcija:	„You are a language generator assistant. Your task is to generate synonyms for a given Serbian word with assigned PoS tags. Synonyms are words with the same PoS tag that have the same or very similar meaning. Separate the outputs with a comma.“
Korisnička instrukcija:	„Please generate synonyms for the Serbian word with the assigned part of the speech tag.“
Zadatak:	Parallel Sentences Generation
Model:	gpt-3.5-turbo
Iteracija:	1
Temperatura:	0.7
Sistemska instrukcija:	„You are language generator and annotation assistant. Your task is to: 1. Generate sentence in English and translate generated sentence to Serbian language. 2. Categorise sentences with one or multiple discrete emotional and sentiment categories from the following list: [anger, trust, fear, sadness, joy, anticipation, surprise, disgust, neutral, positive, negative]. There should be 3 items in the output separated by a comma: English Sentence, Serbian Sentence, Emotional Categories.“
Korisnička instrukcija:	„Please generate pair of parallel sentences in Serbian and English with the assigned emotional categories.“

D. Anketa o razumevanju moralnih vrednosti

U cilju izgradnje i evaluacije jezičkih resursa za prepoznavanje moralnih vrednosti u srpskom jeziku, razvijena je potpuno anonimna anketa za bolje razumevanje uspostavljenih moralnih vrednosti među populacijom koja govori srpski jezik. Anketa je napravljena pomoću Gugl Formulara (eng. *Google Forms*) alata koji omogućava efikasan način za prikupljanje podataka od ispitanika putem interneta, pomoću različitih formata očekivanih odgovora, kao što su jedinstven izbor od nekoliko ponuđenih, duži tekstualni paragrafi ili rangiranje odgovora na numeričkoj skali. Ispitanicima su objašnjena značenja osnovnih moralnih vrednosti prema definicijama datim u MFT[85]. Sve kategorije, definicije i pridruženi upitnici su prevedeni na srpski jezik. Od ispitanika je zahtevano da tekstualne odgovore popunjavaju na srpskom jeziku i ekavskom izgovoru, na ćiriličnom ili latiničnom pismu. Anketa je prosleđena najširoj grupi ispitanika preko društvenih mreža i digitalno uređenih studentskih grupa. Studenti su ispitivani sa ciljem da se istraži kako pripadnici ove društvene grupe razumeju i primenjuju moralne norme u različitim aspektima svakodnevnog života. Pitanja u anketi su podeljena u tri grupe:

- Prvu grupu sačinjavaju 9 pitanja o socio-demografskim karakteristikama ispitanika: *pol, starost, stepen obrazovanja, afilijacija, veličina mesta stanovanja, regionalna pripadnost, religioznost, omiljena vrsta hrane i muzike*.
- U drugoj grupi se nalaze 10 pitanja o razumevanju 5 osnovnih moralnih vrednosti podeljenih u dihotomne parove: *briga/povreda, pravednost/varanje, lojalnost/izdaja, autoritet/subverzija i svetost/degradacija*.
- Treća grupa sadrži 32 pitanja iz MFQ koju su napravili MFT autori.

Tabela D.1: *Poređenje prosečnih MFQ skorova ispitanika iz različitih društvenih grupa za svaku od moralnih vrednosti iz MFT*

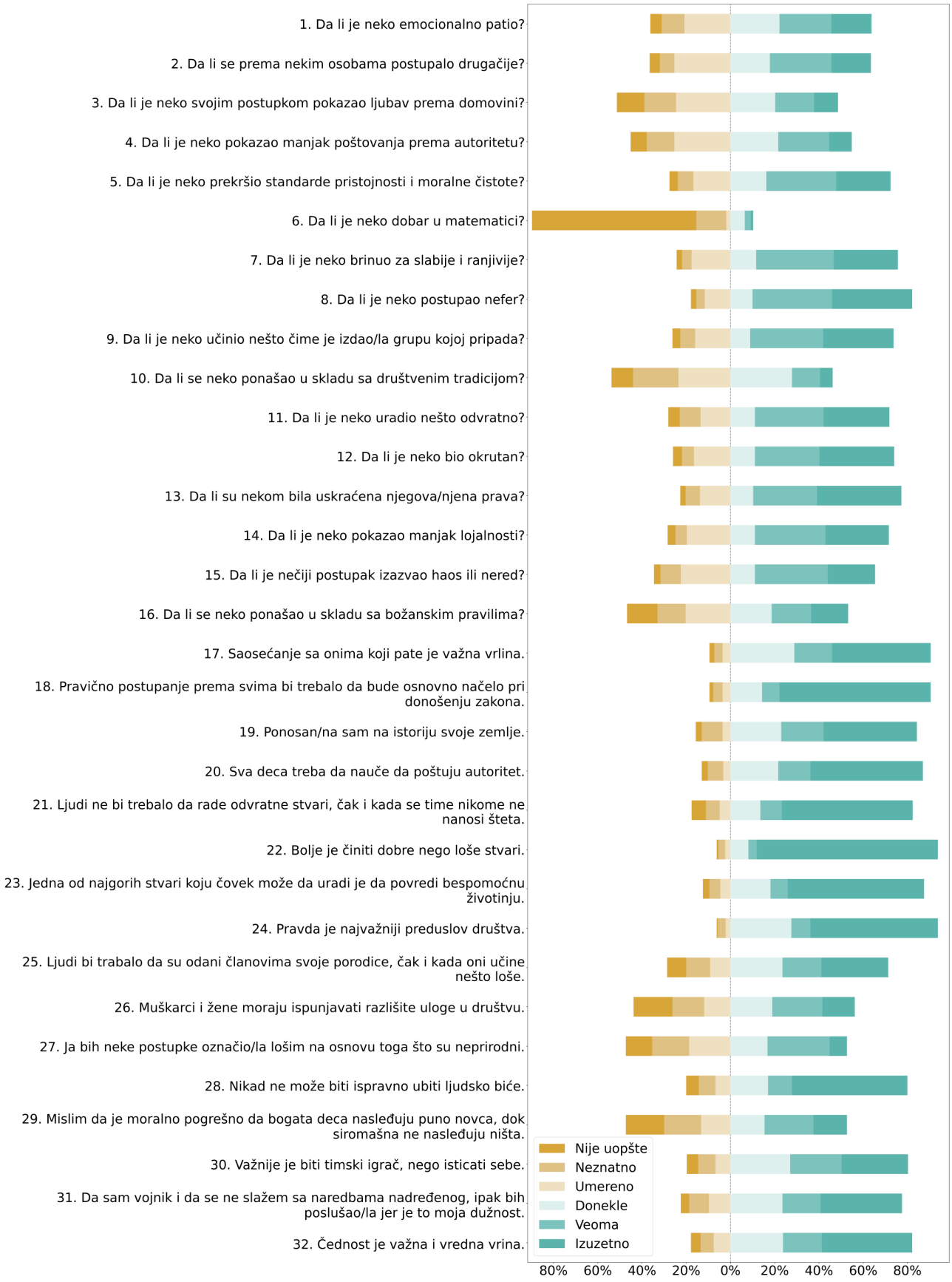
Grupa	`authority`	`care`	`fairness`	`loyalty`	`purity`
američki ispitanik	16.5	20.2	20.5	16.0	12.6
srpski ispitanik	15.6	23.5	22.3	19.6	19.1
srpski student	20.4	10.9	22.2	21.9	18.7
srpski (prosek)	18.5	22.3	21.9	20.3	19.6

Prva grupa pitanja je sastavljena tako da pokrije širok spektar informacija o demografskim karakteristikama ispitanika, koji uključuju pol, starost, stepen obrazovanja, radno angažovanje i regionalne razlike. Pored toga, anketa istražuje uticaj religioznih uverenja i savremenih društvenih uticaja na moralne vrednosti ispitanika, kao što su vegetarijanski način ishrane ili slušanje internacionalne muzike. U drugoj grupi se nalaze pitanja koja od ispitanika očekuju da opišu kako doživljavaju svaku MFT moralnu vrednost ili da navedu primere i situacije koje ih podržavaju ili krše. Pet osnovnih moralnih vrednosti je dato dihotomnim parovima: *briga/povreda, pravednost/varanje, lojalnost/izdaja, autoritet/subverzija i svetost/degradacija*. Odgovori na ova pitanja su slobodnog tekstualnog formata, sa očekivanom dužinom svakog odgovora u granici od 0 do 100 reči. Treću grupu pitanja predstavlja srpski prevod MFQ[86] koji su razvili autori MFT. Upitnik je dizajniran tako da kvantifikuje sklonost pojedinca, ali i čitavih grupa, ka svakoj od 5 moralnih vrednosti na numeričkoj skali od 0 do 30. Svako od pitanja je dodeljeno jednoj moralnoj vrednosti, pri čemu dva pitanja (redni broj 6 i 22) predstavljaju „lažni“ uzorak pitanja i ne ulaze u konačne skorove. Odgovori na pitanja su vrednovani na sledeći način: [0: „*Nije uopšte*“, 1: „*Neznatno*“, 2: „*Umereno*“, 3: „*Donekle*“, 4: „*Veoma*“, 5: „*Izuzetno*“][87]. Dobijeni rezultati po svakom pitanju iz treće grupe pitanja (MFQ) u okviru pokrenute ankete su predstavljeni na slici D.1, na kojoj je sa nagomilavanjima ka ekstremnim vrednostima („*Nije uopšte*“, „*Izuzetno*“) uočljivo da su ispitanici ankete prepoznali umetnuta, odnosno „lažna“ pitanja.

[85] https://moralfoundations.org/

[86] https://moralfoundations.org/questionnaires/

[87] Prevodi odgovora na srpski jezik su delimično prilagođeni radi jednostavnijeg izračunavanja rezultata. U odgovorima za prvih 16 pitanja, oznake se odnose na relevantnost koje dato tvrđenje ima za ispitanika po pitanju moralnosti. U drugom delu upitnika, od ukupno 16 pitanja, oznake se odnose na stepen slaganja ispitanika sa datim tvrđenjem.



Slika D.1: *Statistika odgovora učesnika ankete na treću grupu pitanja (MFQ)*

U anketi je učestvovalo 370 ispitanika, od kojih je 260 studenata i 110 korisnika društvenih mreža. Demografsku strukturu ispitanika sačinjava 62% žena i 38% muškaraca. Najveća starosna grupa su mlađi učesnici između 18 i 30 godina (65.2%), zatim slede ispitanici od 30 do 50 godina (25.7%), dok je oko 10% ispitanika starije od 50 godina. Prema obrazovnom nivou, 52.4% ispitanika ima završeno srednje obrazovanje, 42.1% visokoškolsko obrazovanje (osnovne studije), a 5.5% poslediplomske kvalifikacije (master, specijalističke ili doktorske studije). U pogledu prebivališta, 45.4% ispitanika je iz velikih gradova, 29.7% iz manjih gradova, a 15.9% sa sela. Ispitanici su pokazali veće vrednosti u svim moralnim kategorijama u poređenju sa prosečnim američkim ispitanikom izračunatim prema MFQ sistemu vrednovanja (pogledati tabelu D.1), koje su naročito izražene u kategorijama *loyalty* (+25%) i *purity* (+55%), kao i kategorijama *authority* (+25%) i *care* (-50%) na podgrupi studenata. Struktura ispitanika ukazuje na dominaciju mlađe i obrazovane populacije iz urbanih sredina, što može prouzrokovati specifične obrasce u prikupljenim odgovorima koji su karakteristični za ove društvene grupe, ali ne i za celu populaciju.

Tabela D.2: *Poređenje prosečnih vrednosti moralnih vrednosti ispitanika iz različitih društvenih grupa prema MFQ*

Pitanje	Grupa	authority	care	fairness	loyalty	purity
Pol	žene	18.6	23.1	22.5	20.5	19.9
	muškarci	18.5	20.3	20.5	19.9	18.8
Starost	do 20	18.7	22.3	22.0	20.6	19.7
	20 - 30	18.5	21.7	21.5	18.9	19.1
	30 - 40	18.3	22.7	24.3	22.7	22.5
	40 - 50	16.4	22.2	19.9	18.0	18.0
	50 - 60	16.2	26.3	23.5	19.7	17.5
	preko 60	11.7	24.0	24.3	18.7	22.0
Stepen obrazovanja	srednja škola	18.4	21.8	21.7	20.1	19.3
	viša škola	16.0	15.0	15.5	15.5	16.0
	fakultet	19.0	22.9	22.2	20.8	20.0
	magistratura	18.7	25.3	25.3	21.0	19.0
	doktorat	15.8	24.3	22.7	19.9	19.8
Radno angažovanje	državna organizacija	16.6	23.8	22.8	19.8	18.6
	privatna organizacija	20.8	23.7	23.4	23.5	24.2
	nezaposlen	18.0	21.7	21.0	20.3	19.9
	penzioner	7.0	28.0	29.0	22.0	23.0
	student	18.7	22.2	21.9	20.3	19.5
Veličina mesta stanovanja	selo	19.4	23.2	22.6	21.6	20.9
	varošica (do 15000)	18.8	23.1	21.9	20.2	19.8
	manji grad (15000 - 100,000)	18.5	21.9	21.6	19.9	19.4
	veći grad (preko 100,000)	18.2	22.1	22.0	20.3	19.3
Geografska pripadnost	Grad Beograd	18.7	22.1	22.4	20.5	19.7
	Južna Srbija	18.5	22.0	21.1	19.6	18.8
	Zapadna Srbija	17.2	21.6	21.4	19.3	18.8
	Istočna Srbija	17.6	20.9	20.2	17.7	18.3
	Severna Srbija	19.2	24.2	22.9	21.9	20.1
	Centralna Srbija	19.1	22.8	22.1	21.1	20.3
	Region (Crna Gora, Bosna, Hrvatska)	17.4	20.9	19.3	19.4	19.8
	Inostranstvo	17.0	21.0	25.0	21.0	18.0
Religiozna uverenja	Da	19.2	22.3	21.8	21.0	20.3
	Ne	14.5	22.2	23.2	16.6	15.4
Omiljena hrana	jelo sa roštilja	19.1	22.4	22.5	20.5	20.2
	brza hrana	19.0	23.3	23.6	21.6	19.7
	vegetarijaska kuhinja	20.0	27.3	25.7	22.3	23.0
	domaća srpska kuhinja	19.5	22.5	21.9	21.4	20.9
	internacionalna kuhinja	15.9	21.7	21.3	17.9	16.9
	nemam omiljenu vrstu hrane	17.1	21.5	21.3	18.4	17.4
Omiljena muzika	domaći pop-rok	18.1	23.1	22.5	20.1	20.1
	domaći folk	19.5	21.9	21.3	21.0	19.9
	internacionalni pop-rok	17.4	22.3	22.7	18.7	17.7
	klasična muzika	18.1	24.3	22.5	21.6	21.3
	nemam omiljenu vrstu muzike	18.8	21.9	21.6	20.8	20.0

Dodatna analiza je pokazala razlike između različitih socio-demografskih grupa u prosečnim moralnim

vrednostima merenim prema MFQ (pogledati tabelu D.2). Žene pokazuju nešto veće vrednosti za kategorije *care*, *fairness* i *purity*, dok su izjednačene sa muškarcima u preostalim kategorijama. Starosna grupa od 30 do 40 godina beleži visoke vrednosti u *fairness* i *loyalty*, dok stariji od 50 godina pokazuju opadanje u kategorijama *authority* i *loyalty*, ali porast u *care* i *fairness*. Visoko obrazovani i radno angažovani u privatnom sektoru pokazuju veće vrednosti u svim kategorijama, sa tendencijom opadanja vrednosti u kategorijama *authority* i *loyalty* sa porastom stepena obrazovanja. Primetan je, takođe, opšti trend smanjenja vrednosti moralnih kategorija sa porastom veličine mesta stanovanja. Geografska pripadnost otkriva varijacije u kojima su oblasti severne i centralne Srbije najviše okrenute ka kategorijama *care* i *loyalty*, dok su ispitanici koji žive u inostranstvu više okrenuti kategoriji *fairness*, a manje ka kategorijama *authority* i *purity*. Religiozna uverenja potvrđuju veće vrednosti kod religioznih ispitanika, posebno u dimenzijama *loyalty* i *purity*. Na kraju, preferencije u hrani i muzici sugerišu da vegetarijanci i ljubitelji klasične muzike beleže najviše vrednosti u većini moralnih kategorija, dok ispitanici sa sklonostima ka internacionalnim kulturama (hrana, muzika) pokazuju nešto niže vrednosti u kategorijama *authority*, *loyalty* i *purity*. Ova analiza ukazuje na višedimenzionalne uticaje društvenih, demografskih i kulturnih faktora na usvojene moralne vrednosti.

Prvenstveni cilj ankete je da se identifikuju moralni aspekti srpskog jezika i načini u iskazivanju moralnih stavova među različitim socio-demografskim grupama na srpskom govornom području. Prikupljeni podaci iz ankete su korišćeni za proveru izgrađenih jezičkih resursa za aspekt moralnosti, konkretno MFD.SR leksikona moralnih reči, ali mogu poslužiti istraživačima i za bolje razumevanje dinamike moralnog rasuđivanja u srpskoj jezičkoj i kulturnoj zajednici među mlađom i urbanom populacijom. Rezultati ankete se mogu iskoristiti i za praktične primene kao što su obrazovni programi i inicijative za unapređenje etičkog ponašanja u društvu. U daljim iteracijama pokretanja ankete, pokušaće se sa proširivanjem populacionog uzorka, kao i obuhvatanjem starijih, manje obrazovanih demografskih grupa i stanovnika ruralnih sredina kako bi se osigurala reprezentativnost uzorka i izbegla potencijalna pristrasnost u donošenju zaključaka na osnovu dobijenih rezultata.

E. Emocionalnost, moralnost i etička pitanja u veštačkoj inteligenciji

Sa ubrzanim razvojem AI pojavila su se brojna etička pitanja o načinu njenog korišćenja i potencijalnim opasnostima koje iz njega mogu proisteći. Uočeno je da AI, odnosno ML, modeli mogu pokazati određeni nivo pristrasnosti (eng. *bias*) koje se u modele prenose iz podataka koji su korišćeni za njihovo obučavanje [143]. Podaci iz realnog života mogu sadržati neregularnosti kao što su rasizam, seksizam ili razne druge vrste diskriminacija prema određenim društvenim grupama. Dodatno, postavljaju se i pitanja o mogućoj zloupotrebi privatnosti podataka pojedinaca ukoliko su modeli obučavani nad privatnim podacima kao što su zaštićeni kompanijski podaci, poruke elektronske pošte ili objave na društvenim mrežama [188]. Obučeni modeli takođe mogu dati pogrešne procene i kreirati nepotpun ili nedovoljno precizan sadržaj koji zatim može prouzrokovati dalekosežne negativne posledice ukoliko se takvi rezultati koriste bez dodatne provere, što je naročito primetno kod LLM najnovije generacije koji su dostupni za široku upotrebu. Pristrasnost može biti posledica neodgovarajuće obrade podataka [195] ili procesa donošenja odluka zasnovanih na veštačkoj inteligenciji i sa etičkim dilemama koje takve odluke podstiču [72]. Kako bi se ublažio negativan uticaj AI algoritama, razvijeni su alati za reviziju razvijenih AI modela u cilju provere njihovog rada [39]. Neki od tih alata su otvorenog koda [113], kao što su *IBM Fairness-Aif 360*[88] [21], Guglov *What-if* alat [213] i *Fairlearn*[89]. Drugi primeri alata uključuju *Ethics and Algorithm Toolkit*[90] ili *AI Explorable*[91]. Svi navedeni alati poseduju funkcionalnosti za ublažavanje pristrasnosti i analizu različitih uspostavljenih metrika pravičnosti.

Iz etičkih razloga otvorila su se i pitanja vlasništva nad podacima i izgrađenim modelima, iz kojih proizilaze i odgovornosti za njihovo korišćenje. Na primer, prilikom korišćenja popularnog Čet-GPT alata za generisanje odgovora, koji je nedvosmisleno napravio iskorak u efektivnoj primeni računarske lingvistike za brzo i efikasno pronalaženje informacija, uočeno je da alat generiše pogrešne ili nedovoljno precizne odgovore. U međunarodnoj javnosti su se otvorila pitanja o etičkom korišćenju alata koji su zasnovani na LLM po pitanjima autorstva i integriteta generisanih sadržaja, zaštiti od korišćenja privatnih podataka i podsticanju društvenih nejednakosti. Naučnici su takođe pokušali da utvrde kako se moralne vrednosti iz podataka preslikavaju u ove modele, odnosno da li u njima postoje pristrasnosti prema određenim moralnim vrednostima koje su preuzete iz podataka za obučavanje [75, 2].

Vodeći svetski istraživači, grupe i institucije poslednjih godina intenzivno rade na uspostavljanju osnovnih postulata na kojima bi se zasnivao dalji AI razvoj u cilju obezbeđivanja daljeg pozitivnog razvoja ljudske civilizacije [24, 91]. Tako je, na primer, nakon pojave velikih jezičkih modela GPT-4 generacije, grupa naučnika i lidera u ovoj oblasti uputila poziv široj tehnološkoj javnosti da se dalji razvoj velikih jezičkih modela, moćnijih od GPT-4, privremeno zaustavi dok se ne utvrde sve moguće posledice njihovog korišćenja i pronađu načini da se potencijalne negativne posledice preduprede. U tom cilju inicijatori su pozvali na poštovanje postojećih etičkih postulata, dalji razvoj bezbednosnih protokola, centralizovano upravljanje, uspostavljanje regulatornih metoda za proveru tačnosti, transparentnosti i načina korišćenja izgrađenih modela. Takođe su ukazali na brz i nepredvidiv rast tehnologije u ovoj oblasti koji donosi potencijalne rizike koje ovako izgrađeni sistemi mogu predstavljati za društvo i čovečanstvo u celini. Kao preteča ovih postulata mogu se uzeti „Tri zakona robotike“, koje je pisac romana iz oblasti naučne fantastike, Isak Asimov sredinom prošlog veka na neki način dalekosežno predvideo u svojim delima [14]:

1. Prvi zakon — robot ne sme da povredi ljudsko biće ili, nečinjenjem, dozvoli da se ljudskom biću nanese šteta.
2. Drugi zakon — robot mora da se povinuje naredbama koje mu daju ljudska bića, osim ako bi takva naređenja bila u suprotnosti sa Prvim zakonom.
3. Treći zakon — robot mora da štiti sopstvenu egzistenciju sve dok takva zaštita nije u suprotnosti sa Prvim ili Drugim zakonom.

[88] https://aif360.res.ibm.com/
[89] https://fairlearn.org/
[90] https://ethicstoolkit.ai/
[91] https://pair.withgoogle.com/explorables/

Uspostavljene regulative koje se odnose na prepoznavanje emocija i moralnih vrednosti pomoću AI sistema usmerene su na osiguravanje etičke upotrebe ovih tehnologija, uz poseban fokus na rizike povezane sa pristrasnošću, diskriminacijom i povredom privatnosti. Preporuka o etici veštačke inteligencije koju je kreirao UNESCO[92], naglašava potrebu da AI sistemi, koji prepoznaju moralne vrednosti i emocije, budu transparentni i u skladu sa ljudskim pravima i uspostavljenim moralnim vrednostima. Dokument o regulisanju AI sistema u Evropskoj Uniji, EU AI Akt[93], klasifikuje sisteme za prepoznavanje emocija kao visokorizične tehnologije, i zahteva strogu regulaciju kako bi se sprečila zloupotreba u osetljivim kontekstima, kao što su zapošljavanje ili obrazovanje. AI sistemi koji analiziraju emocije i moralne vrednosti mogu da izazovu neetične posledice, poput nepravednih odluka zasnovanih na pogrešnim tumačenjima, što potencijalno može da ugrozi ljudski integritet i pravo na privatnost. Sve ove regulative podstiču razvoj AI tehnologija koje obezbeđuju pravednost, transparentnost u tumačenjima i zaštitu od negativnih društvenih posledica.

[92] https://unesdoc.unesco.org/ark:/48223/pf0000381137

[93] https://artificialintelligenceact.eu/the-act/

Skraćenice

Oznaka	Opis	Stranica
F_s	F-statistika (eng. *F-statistic*)	24
$ReLU$	Funkcija ispravljene linearne jedinice (eng. *Rectified Linear Unit*)	30
ρ	Spirmanov koeficijent korelacije (eng. *Spearman correlation coefficient*)	26
σ	Sigmoidna funkcija (eng. *Sigmoid*)	30
k	Kohenov kapa koeficijent (eng. *Cohen's Kappa coefficient*)	99
r	Pirsonov koeficijent korelacije (eng. *Pearson correlation coefficient*)	26
$softmax$	Funkcija mekog maksimuma (eng. *Softmax*)	30
$tanh$	Funkcija hiperboličkog tangensa (eng. *Hyperbolic Tangent*)	30
Acc	Tačnost (eng. *Accuracy*)	21
Acc_{Bal}	Balansirana tačnost (eng. *Balanced Accuracy*)	22
AI	Veštačka inteligencija (eng. *Artificial Intelligence*)	1
API	Programski interfejs aplikacije (eng. *Application Programming Interface*)	91
ARI	Automatski indeks čitljivosti (eng. *Automatic Readability Index*)	62
ASPS	Prosečan broj slogova u rečenici (eng. *average syllable per sentence*)	61
ASPW	Prosečan broj slogova u reči (eng. *average syllable per word*)	61
Att	Mehanizam pažnje (eng. *Attention mechanism*)	34
BCE	Binarna unakrsna entropija (eng. *Binary Cross-Entropy*)	65
BERT	Dvosmerne enkoderske reprezentacije iz transformera (eng. *Bidirectional Encoder Representations from Transformers*)	36
BiLSTM	Dvosmerna LSTM mreža (eng. *Bidirectional LSTM*)	33
BoW	Vreća reči (eng. *Bag-of-Words*)	43
Brch	Konverzaciona grana (eng. *A conversation Branch*)	59
CCE	Kategorička unakrsna entropija (eng. *Categorical Cross-Entropy*)	65
Chr	Karakter (eng. *Character*)	45
ConAtr	Konverzacioni atributi (eng. *Conversational Attributes*)	62
CV	Unakrsna validacija (eng. *Cross-Validation*)	20
DL	Duboko mašinsko učenje (eng. *Deep Learning*)	30
DNN	Duboke neuronske mreže (eng. *Deep Neural Networks*)	1
Embd	Ugnježdeni vektori reči (eng. *word embeddings*)	43
EmoAtr	Atributi emocionalnosti (eng. *Emotional Attributes*)	63
EmoInt	Leksikon emocionalnih intenziteta (eng. *The Emotion Intensity Lexicon*)	52
EmoLex	Leksikon asocijacija između reči i emocija (eng. *The Word-Emotion Association Lexicon*)	52
EMR	Egzaktna tačnost (eng. *Exact Match Ratio*)	23
ERT	Ekstremno nasumična stabla (eng. *Extremely Randomized Trees*)	29
ExpAtr	Atributi izražajnosti (eng. *Expressional Attributes*)	62
F_1	F_1-mera (eng. *F_1-measure*)	21
F_1^{Ma}	Makro F_1-mera (eng. *Macro F_1-measure*)	21
F_1^{Mi}	Mikro F_1-mera (eng. *Micro F_1-measure*)	22
F_1^{w}	Težinska F_1-mera (eng. *Weighted F_1-measure*)	22
F_k	Fleisova kapa (eng. *Fleiss's Kappa*)	93
FCNN	Potpuno povezane neuronske mreže (eng. *Fully Connected Neural Networks*)	30
FFN	Mreže sa propagacijom unapred (eng. *Feed Forward Neural Networks*)	30
FN	Pogrešna negativna (eng. *False Negative*)	20
FP	Pogrešna pozitivna (eng. *False Positive*)	20
FRES	Flešova ocena lakoće čitanja (eng. *Flesch Reading Ease Score*)	62
GD	Gradijentni spust (eng. *gradient descent*)	28
GPT	Generativni prethodno trenirani transformer (eng. *Generative Pre-Trained Transformer*)	38
GT	Gugl prevodilac (eng. *Google Translate*)	83
HL	Hamingov gubitak (eng. *Hamming Loss*)	23
HS	Hamingov skor (eng. *Hamming Score*)	23
Inc	Nekorekntno (eng. *Incorrect*)	93
IndAtr	Atributi indikacije (eng. *Indication Attributes*)	61
KDE	Procena gustine kernelom (eng. *Kernel Density Estimation*)	135
L2	Euklidska norma (eng. *Euclidean norm*)	74

Oznaka	Opis	Stranica
LC	Lingvistička i kulturološka prilagođavanja (eng. *Linguistic and Cultural adjustments*)	92
lema	Lema reči (eng. *lemma*)	45
lema_{En}	Lema reči na engleskom jeziku	83
lema_{Sr}	Lema reči na srpskom jeziku	83
LexAtr	Leksički atributi (eng. *Lexical Attributes*)	61
LexCvg	Pokrivanje leksikona (eng. *Lexicon Coverage*)	130
LLaMA	Veliki jezički Meta AI modeli (eng. *Large Language Models Meta AI*)	39
LLM	Veliki jezički model (eng. *Large Language Model*)	2
LR	Logistička regresija (eng. *Logistic Regression*)	26
LSTM	Duga kratkoročna memorija (eng. *Long Short-Term Memory*)	32
LWM	Linsearova mera čitljivosti (eng. *Linsear Write Metric*)	62
Man	Ručno (eng. *Manual*)	93
mBERT	Višejezični BERT (eng. *Multilingual BERT*)	38
Meta	Pridruženi atributi (eng. *Associated Meta Attributes*)	59
MFQ	Upitinik o moralnim osnovama (eng. *Moral Foundations Questionnaire*)	133
MFT	Teorija o moralnim osnovama (eng. *Moral Foundations Theory*)	5
ML	Mašinsko učenje (eng. *Machine Learning*)	1
MLM	Predviđanje skrivenog tokena (eng. *Masked Language Model*)	36
MorAtr	Atributi moralnosti (eng. *Morality Attributes*)	63
Msg	Pojedinačna poruka (eng. *An individual Message*)	59
MSL	Maksimalna dužina sekvence (eng. *Maximum Sequence Length*)	73
NERAtr	Atributi imenovanih entiteta (eng. *NER-based Attributes*)	62
Ngram	Sekvenca od N uzastopnih tokena (eng. *Sequence of N adjacent tokens*)	46
NLP	Obrada prirodnih jezika (eng. *Natural Language Processing*)	2
NRC	Nacionalni istraživački cantar Kanada (eng. *National Research Council Canada*)	52
NSP	Predviđanje nastupajuće rečenice (eng. *Next Sentence Prediction*)	36
PncAtr	Atributi interpunkcije (eng. *Punctuation Attributes*)	62
PoS	Vrsta reči (eng. *Part of Speech*)	44
PoS-Tagging	Određivanje vrste reči (eng. *Part of Speech Tagging*)	44
Prec	Preciznost (eng. *Precision*)	21
Prec^{Ma}	Makro preciznost (eng. *Macro Precision*)	21
Prec^{Mi}	Mikro preciznost (eng. *Micro Precision*)	22
PWN	Princeton verzija WordNet leksikona (eng. *Princeton WordNet Lexicon*)	52
R	Robusno optimizovani BERT (eng. *Robustly Optimized BERT*)	38
RAG	Generisanje potpomognuto pretragom (eng. *Retrieval-Augmented Generation*)	16
Rec	Odziv (eng. *Recall*)	21
Rec^{Ma}	Makro odziv (eng. *Macro Recall*)	21
Rec^{Mi}	Mikro odziv (eng. *Micro Recall*)	22
RF	Slučajne šume (eng. *Random Forests*)	26
RNN	Rekurentne neuronske mreže (eng. *Reccurent Neural Networks*)	30
SGD	Stohastički gradijentni spust (eng. *Stochastic Gradient Descent*)	27
SntAtr	Sintaktički atributi (eng. *Syntactic Attributes*)	61
SRPOL	Alat za izračunavanje intenziteta sentimenta za srpski jezik (eng. *Serbian Polarity Framework*)	85
SVM	Metoda potpornih vektora (eng. *Support Vector Machine*)	26
SWN	srpska verzija WordNet leksikona (eng. *Serbian WordNet Lexicon*)	87
Syn	Lista sinonima (eng. *Synonyms List*)	93
Tf-Idf	Frekvencija termina - inverzna frekvencija dokumenata (eng. *Term frequency-Inverse document frequency*)	43
TLM	Predviđanje skrivenog tokena prevođenjem (eng. *Translation Language Model*)	38
TN	Tačna negativna (eng. *True Negative*)	20
TP	Tačna pozitivna (eng. *True Positive*)	20
Unigram	Sekvenca od jednog tokena (eng. *Sequence of a single token*)	46
WNA	WordNet leksikon afekata (eng. *WordNet-Affect Lexicon*)	52
XLM	Unakrsno-jezički model (eng. *Cross-lingual Language Model*)	38
Čet-GPT	Čet generativni prethodno trenirani transformer (eng. *Chat Generative Pre-Trained Transformer*)	38

Spisak slika

Spisak tabela

Spisak fragmenata koda

Spisak algoritama

Biografija autora

Milena Šošić rođena je 7. juna 1978. godine u Požarevcu. Osnovnu školu „Njegoš“ i prirodno-matematički smer Gimnazije „Jovan Šerbanović“ u Požarevcu završila je kao odličan učenik, učesnik brojnih takmičenja iz prirodnih nauka i nosilac diploma „Vuk Karadžić“ koja se dodeljuje za postignut izuzetan uspeh u toku obrazovanja. Paralelno je pohađala i završila osnovnu i srednju muzičku školu „Stevan Mokranjac“ u Požarevcu, nakon čega je stekla zvanje muzički izvođač – flautista. Školske 1997/1998. godine upisala je osnovne akademske studije na smeru Računarstvo i informatika, studijskog programa Matematika, na Matematičkom fakultetu Univerziteta u Beogradu. Diplomirala je u aprilu 2004. godine sa prosečnom ocenom 8.86.

Magistarske studije na smeru Računarstvo i informatika na Matematičkom fakultetu Univerziteta u Beogradu upisala je školske 2004/2005. godine. Položila je sve predmete predviđene programom studija sa prosečnom ocenom 10.00. Magistarski rad, na temu „**Primena klasifikacije na N-gramsku analizu genoma**“ iz naučne oblasti Bioinformatika, pod mentorstvom profesora dr Nenada Mitića, odbranila je u decembru 2010. godine. Nakon kraće pauze u studiranju, doktorske akademske studije studijskog programa Informatika na Matematičkom fakultetu Univerziteta u Beogradu upisala je školske 2020/2021. godine sa primarnim usmerenjem na usavršavanje svog obrazovanja u oblasti obrade prirodnih jezika.

Od avgusta 2004. godine angažovana je u brojnim kompanijama na različitim pozicijama u računarskoj oblasti, a od decembra 2013. njena poslovna angažovanja prvenstveno uključuju istraživanje podataka i primenu algoritama mašinskog učenja na rešavanje poslovnih zadataka u različitim domenima. Učesnik je nekoliko nacionalnih i međunarodnih konferencija u oblasti mašinskog učenja. Član je akademske zajednice JeRTeH, koja se bavi izgradnjom jezičkih resursa za srpski jezik. Uža oblast njenog naučnog interesovanja obuhvata mašinsko učenje, sa posebnim fokusom na obradu prirodnih jezika. U užem smislu, bavi se razvojem metoda za pretraživanje informacija, automatsku klasifikaciju teksta i izgradnju semantičkih baza znanja.

Прилог 1.

Изјава о ауторству

Потписани-а Милена Шошић

број уписа 2030/2020

Изјављујем

да је докторска дисертација под насловом

Моделовање моралних и емоционалних аспеката језика у класификацији конверзационих текстова

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 01.09.2025.

Милена Шошић

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Милена Шошић

Број уписа 2030/2020

Студијски програм Информатика

Наслов рада Моделовање моралних и емоционалних аспеката језика у класификацији конверзационих текстова

Ментор проф. др Јелена Граовац

Потписани Милена Шошић

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 01.09.2025.

Милена Шошић

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Моделовање моралних и емоционалних аспеката језика у класификацији конверзационих текстова

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 01.09.2025.

Милена Шошић

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.