

Наставно-научном већу
Математичког факултета
Универзитета у Београду

Одлуком Наставно-научног већа Математичког факултета у Београду донетом на **425.** седници одржаној **26.9.2025.** године именовани смо за чланове комисије за оцену докторске дисертације **„Моделовање моралних и емоционалних аспеката језика у класификацији конверзационих текстова”** кандидаткиње **Милене М. Шошић**, магистра рачунарства. Након прегледа предатог рукописа, комисија подноси Наставно-научном већу следећи

Извештај

Биографски подаци

Кандидаткиња Милена М. Шошић је рођена 7. јуна 1978. године у Пожаревцу. Основну школу „Његош“ и природно-математички смер Гимназије „Јован Шербановић“ у Пожаревцу завршила је као одличан ученик, учесник је бројних такмичења из природних наука и носилац дипломе „Вук Караџић“ која се додељује за постигнут изузетан успех током школовања. Паралелно је похађала и завршила основну и средњу музичку школу „Стеван Мокрањац“ у Пожаревцу, након чега је стекла звање музички извођач, флаутиста. Школске 1997/1998. године уписала је основне академске студије на смеру Рачунарство и информатика, студијског програма Математика, на Математичком факултету Универзитета у Београду. Дипломирала је у априлу 2004. године са просечном оценом 8,86.

Магистарске студије на смеру Рачунарство и информатика на Математичком факултету Универзитета у Београду уписала је школске 2004/2005. године. Положила је све испите предвиђене програмом студија са просечном оценом 10.00. Магистарски рад, на тему „Примена класификације на N-грамску анализу генома“ из научне области Биоинформатика, под менторством проф. др Ненада Митића, одбранила је у децембру 2010. године. Након краће паузе у студирању, докторске академске студије студијског програма Информатика на Математичком факултету Универзитета у Београду уписала је школске 2020/2021. године са примарним усмерењем на усавршавање свог образовања у области обраде природних језика.

Од августа 2004. године ангажована је у бројним компанијама на различитим позицијама у рачунарској области, а од децембра 2013. године њена пословна ангажовања првенствено укључују истраживање података и примену алгоритама машинског учења на решавање пословних задатака у различитим доменима. Учесник је неколико националних и међународних конференција у области машинског учења.

Члан је академске заједнице JeRTeH, која се бави изградњом језичких ресурса за српски језик. Ужа област њеног научног интересовања обухвата машинско учење, са посебним

фокусом на обраду природних језика. У ужем смислу, бави се развојем метода за претраживање информација, аутоматску класификацију текста и изградњу семантичких база знања.

Радови у међународним часописима са SCI листе:

1. M. Šošić, J. Graovac, R. Stanković. Building an Emotion Lexicon for Serbian Using Curated Language Resources. *Language Resources and Evaluation*. Springer. Accepted for publication, 2025 (M23, IF=2.0)
2. M. Šošić and J. Graovac. Effective Methods for Email Classification: Is it a Business or Personal Email? *Computer Science and Information Systems*, ComSIS Consortium, 19(3):1155–1175, 2022 (M23, IF=1.5)

Радови у домаћим часописима:

3. M. Šošić. SRPOL – A Lexicon-Based Framework for Sentiment Strength of Serbian Texts. *Review of the National Center for Digitization*, 41:58–73, 2022.

Саопштења на међународним конференцијама штампана у целини или изводу:

4. M. Šošić, J. Graovac, R. Stanković. Social-Emo.Sr: Emotional Multi-Label Categorization of Conversational Messages from Social Networks X and Reddit. *JUDIG - Book of Abstracts*, Belgrade, 2024.
5. M. Šošić, J. Graovac. Categorization of Serbian Words into Emotional and Moral Categories. *AI conference - Book of Abstracts*, Belgrade, 2023.

Предмет и циљ дисертације

Предмет ове докторске дисертације је анализа улоге моралних и емоционалних аспеката језика у задацима класификације конверзационих текстова. У савременом дигиталном окружењу, све већи део комуникације одвија се путем конверзационих порука, преко друштвених мрежа, апликација за размену порука, електронске поште и различитих система за дигиталну комуникацију. Ове поруке не служе само за пренос информација, већ често одражавају дубља уверења, вредности и емоционална стања учесника у комуникацији, што их чини изузетно релевантним за лингвистичку и семантичку анализу.

Циљ овог истраживања је формулисање општег хибридног приступа у класификацији конверзационих текстова уз употребу моралних и емоционалних атрибута садржаја порука. Посебан допринос рада огледа се у развоју нових ресурса за српски језик, лексикона и корпуса обележених моралним и емоционалним категоријама. Ови ресурси представљају основу за изградњу класификационих модела заснованих на алгоритмима машинског учења, са циљем унапређења аутоматског препознавања моралних и

емоционалних аспеката у текстовима на српском језику. Један од кључних циљева дисертације је и испитивање утицаја ових аспеката на класификацију појединачних порука и низова конверзационих порука, као и анализа степена њихове међусобне повезаности.

Очекује се да резултати овог истраживања допринесу развоју савремених метода анализе конверзационих текстова и прошире разумевање улоге моралних и емоционалних елемената у људској комуникацији, са посебним освртом на специфичности дигиталних облика изражавања.

Приказ дисертације

Рукопис има 209 страна и написан је на српском језику. Структура рада обухвата следећа поглавља: 1. Увод 2. Моралност и емоције 3. Конверзациони текстови 4. Класификација као проблем машинског учења 5. Представљање текста 6. Моралност и емоције у класификацији текста 7. Предложена методологија класификације 8. Моделовање емоционалних и моралних аспеката у српском језику 9. Евалуација резултата 10. Дискусија 11. Закључак 12. Библиографија (садржи 227 библиографске јединице). Дисертација садржи и 5 прилога: Придружени атрибути класификације, Анотационе шеме за препознавање емоционалности и моралности, Инжењеринг инструкција, Анкета о разумевању моралних вредности, Емоционалност, моралност и етичка питања у вештачкој интелигенцији.

У уводном поглављу приказује се значај анализе моралних и емоционалних аспеката језика у класификацији конверзационих текстова. Истиче се улога ових аспеката у бољем разумевању дигиталне комуникације и потреба за развојем ресурса за српски језик. Представљени су циљеви истраживања и формулисане хипотезе о утицају моралних и емоционалних аспеката језика на тачност класификације.

У другом поглављу разматрају се појмови моралности и емоција из филозофске, психолошке и друштвене перспективе. Приказане су кључне теорије о моралним основама, укључујући Теорију моралних основа и мисаоне експерименте попут проблема трамваја. У делу о емоцијама представљене су најутицајније теорије о основним људским емоцијама, укључујући Екманову и Плутчикову, уз нагласак на класификацији емоција и њихов значај за разумевање људског понашања. Ово поглавље поставља теоријски оквир за примену моралних и емоционалних категоризација у даљем истраживању.

Поглавље 3 анализира структуру, врсте конверзационих текстова, укључујући поруке електронске поште, поруке на друштвеним мрежама и остале врсте конверзационих порука. Посебна пажња посвећена је аутоматској обради и класификацији порука применом техника машинског учења. Разматра се и улога алата за аутоматско генерисање одговора у различитим доменима. На крају се приказују различите технике визуелизације конверзационих текстова које се користе у циљу бољег разумевања структуре и динамике комуникације.

У четвртом поглављу детаљно је обрађена класификација као проблем машинског учења.

Објашњени су основни појмови као што су фазе обучавања и тестирања, врсте класификације, као и мере за евалуацију модела. Посебна пажња посвећена је избору значајних атрибута класификације. Описани су традиционални предиктивни алгоритми машинског учења, укључујући оптимизоване линеарне методе и ансамбл методе. Приказани су и алгоритми и архитектуре дубоког учења, с нагласком на њиховој примени у обради текстуалних података. Обрађене су рекурентне неуронске мреже, с посебним освртом на механизам пажње који омогућава моделима да динамички идентификују и обраде најрелевантније делове текста. Посебна пажња посвећена је трансформер архитектури, као и моделима као што су BERT (енг. *Bidirectional Encoder Representations from Transformers*) и GPT (енг. *Generative Pre-Trained Transformer*) који представљају темељ савремених система за разумевање и генерисање природног језика.

У петом поглављу обрађују се технике представљања текста у облику који је погодан за примену алгоритама машинског учења. Описани су различити приступи векторизацији текста као што су врећа речи (енг. *Bag-of-Words*), фреквенција термина - инверзна фреквенција документа (енг. *Term frequency-Inverse document frequency*) и угњеждени вектори речи (енг. *Word embeddings*). Посебан део посвећен је техникама нормализације текста које укључују стандардизацију садржаја, правописну проверу, одређивање врсте речи, морфолошку нормализацију, уклањање стоп речи и шума, као и различите технике токенизације текста. Истакнута је и важност прилагођавања ових техника конкретном језику, задатку и моделу, чиме се омогућава ефикаснија обрада и тачније моделирање текстуалних података.

Поглавље 6 се бави улогом моралности и емоција у класификацији текста. Представљен је приступ заснован на прављењу семантичке базе знања моралних и емоционалних речи, приступ заснован на машинском учењу, као и хибридни приступ који комбинује претходна два приступа ради постизања веће прецизности. Описане су ручне методе обележавања, обележавање засновано на лексичким ресурсима и обележавање коришћењем математичких метода и техника машинског учења. Приказани су и кључни лексикони и корпуси обележени емоционалним и моралним категоријама, који омогућавају дубље разумевање афективних и моралних димензија текста.

Поглавље 7 описује методологију класификације конверзационих текстова, с посебним освртом на придружене атрибуте који могу имати значајну улогу у моделирању конверзационих текстова коришћењем алгоритама машинског учења. Придружени атрибути су организовани у пет група: лексички, конверзациони, атрибути изражајности, емоционалности и моралности. Показује се да ови атрибути значајно доприносе унапређењу перформанси модела машинског учења, посебно у задацима као што су анализа порука електронске поште и откривање истинитости информација на друштвеним мрежама. Предложени су модели дубоког учења засновани на архитектури двосмерних мрежа са дугом краткорочном меморијом (енг. *Bidirectional Long Short-Term Memory, BiLSTM*). Моделовање је извршено над појединачним порукама, као и конверзационим гранама непосредних порука. Приказани су резултати класификације порука електронске поште на пословне и личне, као и класификација истинитости гласина и типа деловања на објављену гласину.

У поглављу 8 анализирано је моделовање емоционалних и моралних аспеката у српском језику. Представљени су кључни изазови који се јављају при примени техника обраде природних језика на српски, као што су морфолошка сложеност, изговорне и правописне варијације, недостатак обележених језичких ресурса, као и недовољан број обучених модела и специјализованих алата. Детаљно је описан развој емоционалног лексикона EmoLex.SR и лексикона моралности MFD.SR, који омогућавају прецизније препознавање емоција и моралних категорија у тексту на српском језику. Такође, представљен је процес прикупљања корпуса конверзационих података са друштвених мрежа као што су Twitter и Reddit, и њиховог обележавања у емоционалне и моралне категорије користећи стандардизоване шеме категоризације. У раду је представљен развој модела машинског учења заснованих на савременим трансформер архитектурама, који су прилагођени српском језику кроз дообучавање на задатку класификације текстова. Показан је значај комбиновања аутоматизованих метода и ручне провере у циљу постизања високог квалитета података и резултата.

У деветом поглављу представљен је процес евалуације изграђених семантичких ресурса за српски језик, као и резултати класификације конверзационих текстова на енглеском језику. Резултати тестирања развијеног лексикона SentiWords.SR и алата SRPOL за мерење интензитета сентимента анализирани су на више различитих текстуалних корпуса. Оцењена је и ефикасност лексикона EmoLex.SR у класификацији текстова, применом на различитим скуповима текстова на српском језику, уз коришћење паралелних српско-енглеских корпуса и великих језичких модела. Поред тога, у поглављу је обрађена провера квалитета развијеног лексикона MFD.SR, коришћењем анкете о разумевању моралних вредности, спроведене међу испитаницима са српског говорног подручја. Посебна пажња посвећена је значају различитих атрибута у класификацији конверзационих текстова. Изложени су и резултати класификације електронске поште на пословну и личну, као и резултати идентификације истинитости гласина и типова деловања у коментарима на гласине.

У поглављу 10 разматрају се резултати добијени кроз тестирање хипотеза везаних за примену моралних и емоционалних атрибута у моделима класификације. Потврђено је да укључивање ових атрибута значајно побољшава тачност модела, посебно у комбинацији са двосмерним мрежама са дугом краткорочном меморијом и механизмом пажње. Такође је показано да развијени српски семантички лексикони прецизније препознају моралне и емоционалне особине у тексту од лексикона који су добијени аутоматским преводњем са других језика. У раду су истакнути изазови у коришћењу великих језичких модела, као и потреба за унапређењем процеса ручног обележавања.

У закључку дисертације су објашњени главни резултати и значај истраживања, са посебним освртом на улогу моралних и емоционалних аспеката у класификацији конверзационих текстова. Истакнут је значајан допринос развоју лингвистичких ресурса за српски језик, као и примена савремених модела машинског учења у решавању класификационих задатака на конверзационим текстовима. Такође су предложени правци за будућа истраживања и указано је на важност етичке употребе развијених система.

Научни допринос дисертације

Дисертација **„Моделовање моралних и емоционалних аспеката језика у класификацији конверзационих текстова“** доноси значајан допринос у области обраде природних језика, са посебним значајем за српски језик. Први научни допринос огледа се у развоју и имплементацији хибридног модела који интегрише моралне и емоционалне атрибуте као семантичке карактеристике за побољшање тачности класификације конверзационих текстова. Веома значајан допринос дисертације представља израда нових, прилагођених језичких ресурса за српски језик, укључујући емоционалне (WNA.SR, EmoLex.SR) и моралне (MFD.SR) лексиконе, као и обележене корпусе Social-Emo.SR и Social-Mor.SR. Поред тога, развијени су и модели машинског учења обучени на овим ресурсима, што омогућава боље разумевање и класификацију текста. Значајан допринос је и лексикон SentiWords.SR, који садржи вредности интензитета сентимента, као и пратећи програмски алат SRPOL, који се користи за мерење интензитета сентимента у конверзационим текстовима на српском језику.

Поред тога, рад уводи иновативне приступе обележавању и провери ових ресурса кроз комбинацију аутоматских и ручних метода, што доприноси квалитету података и поузданости резултата. Значајан допринос представља и евалуација примењених модела на корпусима текстова из реалних дигиталних комуникационих окружења, као и адаптација савремених трансформер архитектура српском језику и специфичностима друштвених мрежа.

Дисертација истиче значај укључивања моралних и емоционалних димензија у области обраде природних језика као важног приступа за дубље и свеобухватније разумевање дигиталне комуникације, што представља иновативан допринос у српској лингвистичкој и рачунарској науци.

На крају, резултати дисертације отварају могућности за даље истраживање у правцу оптимизације модела, проширења језичких ресурса и примене у различитим апликацијама дигиталне комуникације, као и унапређења процеса обележавања и развоја етичких стандарда у коришћењу вештачке интелигенције. Овај рад доприноси развоју теоријских и примењених аспеката обраде српског језика, као и унапређењу интердисциплинарних истраживања на пољу природних језика, морала и емоција.

Закључак

Резултати до којих је Милена М. Шошић дошла током истраживања у области обраде природних језика, и који су представљени у рукопису **„Моделовање моралних и емоционалних аспеката језика у класификацији конверзационих текстова“** представљају вредан научни допринос у области обраде природних језика. Кандидаткиња Милена М. Шошић је показала одлично познавање ове области и оспособљеност за обављање самосталног научног рада. Добијени резултати имају значајан потенцијал за даљу примену и наставак истраживања. Имајући у виду све претходно наведено предлажемо Наставно-научном већу Математичког факултета да рукопис **„Моделовање моралних и**

емоционалних аспеката језика у класификацији конверзационих текстова” кандидаткиње Милене М. Шошић прихвати као докторску дисертацију и одреди комисију за њену одбрану.

У Београду, 29.9.2025.

Чланови комисије за оцену

(др Ненад Митић, редовни професор)

(др Младен Николић, ванредни професор)

(др Ранка Станковић, редовни професор,
Универзитет у Београду, Рударско-геолошки факултет)