ВЕЋУ ЗА СТУДИЈЕ ПРИ УНИВЕРЗИТЕТУ У БЕОГРАДУ

На редовној седници Већа за студије при Универзитету у Београду одржаној 23. марта 2026. године, Одлуком бр. 06 Број: 06-4574/IV/981/2-26 ЈКЈ / именовани смо за чланове Комисије за оцену научне заснованости предложене теме докторске дисертације под насловом: **„Хибридни модел за препознавање увредљивог говора у кратким текстовима написаним на природним језицима са ограниченим ресурсима“** и испуњености услова кандидата **Јокић Данка** и предложених ментора **др Ранка Станковић** и **др Јелене Граовац.**

На основу поднете документације уз Пријаву теме кандидата **Данка Јокић** Комисија подноси Већу за студије при Универзитету у Београду следећи:

И З В Е Ш Т А Ј
О ОЦЕНИ НАУЧНЕ ЗАСНОВАНОСТИ ТЕМЕ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ И ИСПУЊЕНОСТИ УСЛОВА КАНДИДАТА И МЕНТОРА

1. Биографија кандидата

Данка Јокић је рођена у Ваљеву, 23.12.2973. године, где је завршила основну школу и Ваљевску гимназију (1992.). Дипломирала је у року на Електротехничком факултету у Београду 1997. године са просечном оценом 8.90. 2005. године је уписала партнерске постдипломске студије HEC Paris и Економског факултета у Београду. Студије је завршила као друга у класи од 25 студената са просеком 9.05 и стекла HEC диплому мастера менаџмента и економије 2007. године. Магистарску тезу »Примена друштвених мрежа у процесима регрутације и селекције кандидата за посао у компанијама високих технологија« одбранила је на Економском факултету 2016. године и стекла звање магистра наука пословног управљања. Поседује преко 27 година радног искуства у различитим областима информационих технологија и организације рада у нафтној, банкарској и ИКТ индустрији. Више од 20 година ради на менаџерским и консултантским позицијама у домаћим и страним компанијама. Током своје каријере водила је велике пројекте имплементације информационих система у банкарском, јавном и сектору телекомуникација. Поседује међународно признат сертификат за управљање пројектима PMP (енгл. Project Management Professional). Од фебруара 2024. запослена је у компанији НЛБ ДигИТ д.о.о. на пословима водећег научника за податке у области обраде природног језика.

Уписала је докторске студије при Универзитету у Београду, модул Интелигентни системи, 2019. године.

Активан је члан Друштва за језичке ресурсе и технологије – JeРТех, где учествује у развоју система и алата за обраду српског језика.

Положени испити на докторским студијама

Предмет	Оцена	Шифра предмета	ЕСПБ
Рачунарска визија	10	ИСТЕ4	10
Методе и технике вештачке интелигенције	10	ИС001	12
Процесирање природног језика	10	ИСТЕ2	10
Интелигентна анализа података	9	ИСПР3	9
Интелигентни едукативни системи	9	ИСПР6	11
Статистичке методе у вештачкој интелигенцији	10	ИСВИ4	9
Екстракција информација из текста	10	ИСПР4	11
Машинско учење	10	ИСВИ7	11
Семантички Веб	10	ИСПР1	9
Израда и одбрана Приступног рада за докторску дисертацију	10	ИС011	30
Просечна оцена/Укупно	9.8		122

2. Библиографија кандидата

(категорисано према категоризацији надлежног Министарства, објављени или прихваћени за штампу)

I. Саопштења са међународног скупа штампана у целини (**М33**):

Torunoğlu-Selamet, D., Arslan, D., Wilkens, R., He, W., Eryiğit, D., Pickard, T., ... **Jokic, D**., … & Xie, Z. (2026). A Parallel Cross-Lingual Benchmark for Multimodal Idiomaticity Understanding. In Proceedings of the Fifteenth Language Resources and Evaluation Conference (рад прихваћен за штампу).

Jokić, D., & Stanković**,** R. (2025). Exploring the Synergy Between LLMs and Knowledge Graphs for Advanced Abusive Speech Detection in Serbian. In: *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig Thematic Collection of Papers* (Vol. 1, pp. 395–410). Belgrade: Faculty of Philology, University of Belgrade. https://doi.org/10.18485/judig.2025.1.ch23

Jokić, D., Stanković, R., & Todorović, B. Š. (2024). Abusive speech detection in Serbian using machine learning. In *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security* (pp. 153-163). https://aclanthology.org/2024.nlpaics-1.18/

Jokić, D., Stanković, R., Krstev, C., & Šandrih, B. (2021). A Twitter Corpus and lexicon for abusive speech detection in Serbian. In 3rd Conference on Language, Data and Knowledge (LDK 2021) (pp. 13-1). Schloss Dagstuhl–Leibniz-Zentrum für Informatik. https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.LDK.2021.13

Stanković, R., Mitrović, J., **Jokić, D.**, & Krstev, C. (2020). Multi-word expressions for abusive speech detection in Serbian. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (pp. 74-84). https://aclanthology.org/2020.mwe-1.10/

Саопштење са међународног скупа штампано у изводу (**M34**)

Jokić, D., Stanković, R., & Jaćimović, J. (2024). *Knowledge Graphs in the Era of Large Language Models: Opportunities and Challenges.* In: **Book of Abstracts of the International Conference South Slavic Languages in the Digital Environment JuDig** (pp. 60). Belgrade.

II. Саопштења са скупа националног значаја штампана у изводу (**M64**)

Jokić, D., Stanković, R., & Kovačević, A. (2025). *Fake News Detection in Serbian using Large Language Models and Knowledge Graphs*. In: **Book of Abstracts of the Artificial Intelligence Conference** (pp. 131-132). Belgrade: Serbian Academy of Sciences and Arts (SASA).

Jokić, D. (2023). *Primena algoritama mašinskog učenja za detekciju uvredljivog govora na srpskom jeziku*. In: **Book of Abstracts of the Artificial Intelligence Conference** (pp. 50–51). Belgrade: Serbian Academy of Sciences and Arts (SASA).

3. Предмет и циљеви докторске дисертације

3.1. Предмет докторске дисертације

Са развојем интернета и све већом употребом масовних онлајн медија и друштвених мрежа, откривање увредљивог језика добија на значају као важно подручје истраживања првенствено због свог утицаја на безбедност на мрежи. Увредљив језик се појављује у различитим облицима, укључујући изразе мржње и увреде, токсичност, мизогинију, расизам, сексизам, сајбер малтретирање итд., и у великој мери угрожава онлајн окружење друштвених медија (Mozafari et al., 2022). Говор мржње је постао главни проблем за све врсте онлајн платформи на којима се појављује све већа количина садржаја који генеришу корисници: од коментара на вестима на вебу, преко друштвених мрежа, до чет-ова на игрицама у реалном времену (Saleem et al., 2017).

Како вишејезичне платформе друштвених медија подстичу своје кориснике да комуницирају на свом матерњем језику, неопходно је развити аутоматско управљање овим великим системима, укључујући и алате за откривање говора мржње и увредљивог говора за све језике (Mozafari et al., 2024). Велики кораци у том правцу су направљени са језицима богатим ресурсима као што су енглески, немачки и шпански. Са друге стране, српски и други језици са ограниченим ресурсима остају по страни због ограничене количине анотираних података и недостатка лексичких ресурса. Прикупљање и означавање података је посао који захтева много ресурса, а сложена, субјективна и имплицитна природа увредљивог говора отежава поузданост процеса анотације (Caselli et al., 2021; Waseem, 2016). За обраду природног језика посебан изазов представљају

богатство граматичких облика и морфолошка сложеност српског језика, и стога је неопходно стално обогаћивање постојећих језичких ресурса и алата (Stanković et al., 2020). Пошто творци увредљивог говора често користе суптилне тактике, као што су намерне грешке у куцању или сленг, мешање језика (енгл. *code-switching*) и имплицитне увреде (Ali et al., 2025), а и сам језик се развија током времена, истраживање које обухвата ове различите аспекте доприноси робусности основних модела када се суоче са разноврсним текстуалним подацима.

Пренос знања из језика са великом количином ресурса у језике са мање развијеним ресурсима путем међујезичког и мета-учења се користи и као техника за откривање увредљивог говора (Eronen et al., 2022; Mozafari et al., 2022; Ranasinghe et al., 2024). За поуздану имплементацију, морамо осигурати да су развијени модели објашњиви, да покривају податке ван домена на којима су обучавани и да су праведни за различите демографске групе (Song et al., 2024). Технике као што су SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) и LIME (Ribeiro et al., 2016) за објашњење резултата, као и стратегије ублажавања пристрасности модела, се све више примењују на моделе за препознавање говора мржње, укључујући и оне засноване на великим језичким моделима (BJM) (Kibriya et al., 2024; Piot & Parapar, 2025). Ове технике у комбинацији са језичким ресурсима као што су су лексикони и графови знања (Rajabi & Etminani, 2024) могу допринети премошћавању јаза између техничких перформанси и практичне поузданости.

У овој докторској дисертацији истражују се могућности примене метода машинског учења, лексичких ресурса, база знања и великих језичких модела (Devlin et al., 2019; Mihailo Škorić, 2024), заједно са моделима за препознавање сентимента и/или емоција, као и ироније и сарказма, за препознавање увредљивог говора у кратким текстовима на српском језику. Поред означавања целих текстуалних јединица, истражиће се и идентификовање делова текста који садржи овакав говор (Paraschiv et al., 2024; Pavlopoulos et al., 2021; Ranasinghe et al., 2024), као и препознавање мета према којима је овај говор усмерен (Mathew et al., 2021; Zampieri et al., 2022). Посебан акценат ће бити на препознавању имплицитног увредљивог говора, најчешће израженог кроз примену стилских фигура ироније и сарказма (Cabrera et al., 2025; Mladenović et al., 2017; Ranasinghe, 2020). Додатно ће се применити и модел за препознавање сентимента и/или емоција (Batanović et al., 2016; Šošić et al., 2026), као додатна компонента система за препознавање увредљивог говора.

Циљ овог истраживања је допринос не само у домену језичких технологија већ и примена у безбедносним системима на друштвеним мрежама, порталима и форумима на интернету, видео игрицама, платформама за учење, интерним порталима компаније, државним институцијама итд.

3.2. Циљеви докторске дисертације

Општи циљ ове дисертације је развој методологије за аутоматско препознавање увредљивог говора на природним језицима са ограниченим ресурсима, са фокусом на српски језик. Ова методологија обухвата процесе креирања језичких ресурса, база знања и хибридних модела машинског учења, као и примена метода објашњиве вештачке интелигенције (ВИ) у сврху прецизнијег препознавања увредљивог текста.

У специфичне циљеве спадају:

- Израда ручно анотираног скупа података, који садржи кратке текстове означене на нивоу целог текста и његових делова као увредљиве или неувредљиве.

- Израда ручно анотираног скупа података, који садржи кратке текстове означене на нивоу целог текста маркерима (етикетама) за иронију, сарказам или одсуство обе стилске фигуре.
- Примена метода машинског учења, укључујући најсавременије ВЈМ, за препознавање увредљивог говора:
    - на нивоу целог текста,
    - на нивоу речи и делова реченице.
- Примена трансферног учења кроз употребу скупова података, који садрже примере увредљивог говора на другим језицима, за детекцију увредљивог говора на српском језику.
- Испитивање објашњивости препознавања увредљивог говора применом SHAP и LIME техника.
- Развој онтологије увредљивог говора, као формалне репрезентације концепата увредљивог говора и њихових односа.
- Израда лексикона увредљивог говора и његова интеграција са онтологијом, као и евалуација његове примене за препознавање увредљивог говора. Лексикон ће садржати и фразе чије појединачне компоненте нису уредљиве, док цела фраза јесте.
- Примена онтологије и лексикона увредљивог говора за идентификацију мета према којима је говор усмерен.
- Испитивање могућности генерисања увредљивог говора применом ВЈМ у циљу обогаћивања садржаја лексикона увредљивог говора.
- Примена развијеног скупа података са примерима ироније и сарказма, за детекцију имплицитног увредљивог говора.
- Развој хибридног модела, који укључује моделе машинског учења, лексичке ресурсе и базе знања, у циљу препознавања увредљивог говора у кратким текстовима на српском језику.

Циљ докторске дисертације је креирање ефикасног и надоградивог система који интегрише постојеће и будуће језичке моделе, лексичке ресурсе и базе знања у један робустан систем за препознавање увредљивог говора.

4. Хипотезе

На основу анализе релевантне и расположиве литературе, као и уочених отворених питања у области теме докторске дисертације, формулисане су следеће хипотезе:

Х1 Применом најједноставнијих модела машинског учења заснованих искључиво на садржају, као што је врећа речи (енгл. Bag of Words) не може се постићи најбољи резултат (енгл. SOTA скраћено од State-Of-The Art), односно поузано препознати увредљив говор у кратким текстовима. Овај модел би се користио и као основни модел за поређење са осталим напреднијим моделима.

Х2 Примена лексикона и онтологије увредљивог говора доводи до унапређења резултата модела, како по питању финије класификације увредљивог говора, тако и за прецизнију детекцију имплицитног увредљивог говора.

Х3 Примена трансферног учења утиче на побољшање резултата класификације увредљивог говора у поређењу са обучавањем модела само са скупом за обучавање модела на српском језику.

Х4 Додавање модула за препознавање сентимента и/или емоција доводи до побољшања резултата система за препознавање увредљивог говора.

X5 Додавање модула за препознавање ироније и сарказма омогућава прецизнију детекцију увредљивог садржаја у тексту.

X6 Најбољи резултат се добија применом хибридног модела, који обухвата модел машинског учења заснован на трансформер архитектури, лексичке ресурсе, базу знања, модул за препознавање сентимента и/или емоција и модул за препознавање ироније и сарказма.

5. План рада

Истраживање у оквиру предложене докторске дисертације извршиће се у три фазе и састојаће се од следећих корака:

Фаза припреме

(i) Изучавање литературе претходних релевантних истраживања и утврђивање теоријског оквира.

(ii) Истраживање најсавременијих технологија примене метода машинског учења и ВЈМ на проблеме препознавања увредљивог говора у кратким текстовима.

(iii) Преглед постојећих ресурса и технологија који би се применили у овом истраживачком раду.

Фаза развоја

(iv) Припрема и проналажење анотираних корпуса за задатке препознавања увредљивог говора, анализе сентимента и емоција и препознавања ироније и сарказма.

(v) Дефинисање структуре и развој лексикона увредљивог говора.

(vi) Генерисање увредљивог говора применом ВЈМ ради обогаћивања лексикона примерима увредљивог говора.

(vii) Дизајн и реализација онтологије увредљивог говора.

(viii) Развијање графа знања увредљивог говора.

(ix) Развијање модела за препознавање увредљивог говора на основу лексикона увредљивих речи.

(x) Развијање модела за препознавање увредљивог говора у кратким текстовима на српском језику применом метода машинског учења, укључујући ВЈМ.

(xi) Развијање објашњивог модела за препознавање увредљивог говора, који детектује делове текста који су увредљиви.

(xii) Развој модела који коришћењем онтологије и графа знања препознаје мету (категорију) увредљивог говора.

(xiii) Припрема или проналажење одговарајућег класификатора за препознавање сентимента и/или емоција у кратким текстовима на српском језику.

(xiv) Припрема одговарајућег класификатора за препознавање ироније и сарказма у кратким текстовима на српском језику.

(xv) Пројектовање и развој хибридног система за детекцију увредљивог говора заснованог на примени машинског учења, лексичких ресурса и база знања, уз подршку модула за препознавање сентимента и/или емоција и класификатора за детекцију ироније и сарказма.

Финална фаза

(xvi) Формална евалуација предложене онтологије како би се потврдила њена конзистентност и функционалност у домену препознавања увредљивог говора.
(xvii) Евалуација појединачних делова система и избор решења за примену.
(xviii) Спровођење студије аблације (енгл. *ablation study*) и избор оптималног решења.
(xix) Компаративна анализа перформанси хибридног система и појединачних модула за препознавање увредљивог говора.
(xx) Дискусија и извођење закључака на основу добијених резултата.

6. Методе које се користе у истраживању

Више метода ће се користити за решавање постављених проблема у овом раду:

I. У првој фази ће преовлађивати дескриптивна метода за прикупљање, систематизацију и хармонизацију литературе, ресурса, модела и постојећих софтверских решења. Примениће се компаративна и аналитичка метода за упоређивање и интерпретацију резултата из претходних истраживања, уз основну статистичку обраду података.

II. У другој фази ће се примењивати методе рачунарске лингвистике за анотацију корпуса, у комбинацији са методама машинског учења (укључујући дубоко учење), ради развоја и финог подешавања језичких модела. За развој композитних система ће се користити методе вероватноће и статистике, посебно у задацима анализе препознавања увредљивог говора, као и детекције иронија и сарказма. Језички модели ће се затим користити за креирање векторских репрезентација.За развој онтологије примениће се интегрисана методологија развоја онтологија, приступ развоју заснован на захтевима, који је уобичајен у праксама софтверског инжењерства (Devedzić, 2002).

III. У финалној фази, евалуација развијених модела вршиће се помоћу аутоматске и ручне методе евалуације, уз квантитативну и квалитативну анализу добијених резултата. Применом статистичких тестова значајности и компаративних метода утврђиваће се евентуално побољшање у односу на претходна решења и вршиће се формулисање коначних закључака. У случају задатака који ће се сводити на класификацију (на пример, препознавање увредљивог говора или иронија и сарказма) користиће се мере прецизност, одзив и хармонијска мера F1, а за задатке који ће се сводити на регресију (на пример, додела вредности сентимента) користиће се метод максималне ентропије. Посебан сегмент истраживања обухватиће формалну евалуацију предложене онтологије како би се потврдила њена конзистентност, обухватност и функционална применљивост у домену препознавања увредљивог говора

7. Мултидисциплинарност теме

Мултидисциплинарност теме огледа се како у различитим научним методама, тако и у ресурсима који ће се користити и доменима примене добијених резултата. Обрада природног језика представља спој више дисциплина, пре свега рачунарства, лингвистике и статистике, па ће и ово истраживање обухватити методе корпусне и рачунарске лингвистике, потом примену машинског учења (укључујући дубоко учење), као и статистичке приступе. На тај начин, рад доприноси различитим областима: лингвисти добијају проширене ресурсе и моделе за анализу, док истраживачи у рачунарству могу да користе развијене моделе и алате за изградњу интелигентних система, нарочито у области безбедности. Стога ће резултати истраживања, изузев доприноса научној заједници, имати и практичну вредност у алатима за препознавање

увредљивог говора, задатку за који до сада не постоје отворена решења за српски језик.

8. Очекивани научни допринос докторске дисертације

Истраживање ће, иако усмерено превасходно на развој хибридног модела за препознавање увредљивог говора на српском језику, донети методолошке и технолошке доприносе који су применљиви и на остале језике са недовољно развијеним лингвистичким ресурсима. Онтологија увредљивог говора, које ће бити развијена приликом рада на овој докторској тези, биће пројектована да може да се користи и за друге језике. У ширем смислу, то подразумева да ће као резултат овог рада настати вредни семантички ресурси у оквиру екосистема семантичког веба за унапређење аутоматског препознавања увредљивог говора на језицима са ограниченим ресурсима, што ће допринети безбеднијем онлајн окружењу.

Истовремено, употреба генеративних модела за синтетичко проширење лексичких ресурса унапредиће наше разумевање могућности (и ограничења) оваквих модела у процесу обраде природног језика. Посебан значај имаће резултати који показују у којим случајевима је ово проширење корисно и где се могу јавити проблеми попут непрецизних дефиниција, погрешног контекста или културолошки неодговарајућих примера.

Конкретно, очекивани научни допринос чине:

- Развој методолошког и технолошког оквира за препознавање увредљивог говора на српском језику, на нивоу целог текста, са могућношћу примене на друге језике са недовољно развијеним лингвистичким ресурсима.
- Развој јавно доступних ресурса за српски језик, који до сада нису постојали или нису били расположиви у отвореном приступу: ручно анотираног скупа података, лексикона увредљивог говора и модела за препознавање увредљивог говора.
- Развој методолошког и технолошког оквира за препознавање увредљивих делова текста у кратким текстовима.
- Успостављање оквира, анализа и вредновање употребе генеративних модела за синтетичко проширење лексичких ресурса, уз идентификацију кључних предности, ограничења и потенцијалних ризика.
- Развој побољшаног концептуалног модела увредљивог говора, који доноси шири спектар мета говора мржње, укључује векторске репрезентације речи и више нивоа анотације истог скупа података.
- Развој методолошког и технолошког оквира за објашњиво препознавање увредљивог говора у кратким текстовима на српском језику.
- Анализа и вредновање употребе међу-језичких угњеждених репрезентација речи и трансферног учења за пренос знања из језика богатијих ресурсима или из језика из сличне језичке групе на српски језик.
- Развој методолошког и технолошког оквира за хибридни модел за детекцију увредљивог говора, система који се састоји од модела машинског учења, у комбинацији са лексичким ресурсима, базом знања. модулом за препознавање сентимента и емоција, као и класификатором за детекцију ироније и сарказма.

9. Библиографски подаци релевантни за докторску дисертацију

Ali, S., Blackburn, J., & Stringhini, G. (2025). *Evolving Hate Speech Online: An Adaptive Framework for Detection and Mitigation* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2502.10921

Batanović, V., Nikolić, B., & Milosavljević, M. (2016). Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2688–2696). European Language Resources Association (ELRA). https://aclanthology.org/L16-1427

Cabrera, A., Lei, L., & Ortega, A. (2025). *Transfer Learning via Lexical Relatedness: A Sarcasm and Hate Speech Case Study* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2508.16555

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for Abusive Language Detection in English. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 17–25). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.woah-1.3

Devedzić, V. (2002). Understanding ontological engineering. *Communications of the ACM*, *45*(4), 136–144. https://doi.org/10.1145/505248.506002

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Eronen, J., Ptaszynski, M., Masui, F., Arata, M., Leliwa, G., & Wroczynski, M. (2022). Transfer language selection for zero-shot cross-lingual abusive language detection. *Information Processing & Management*, *59*(4), 102981. https://doi.org/10.1016/j.ipm.2022.102981

Kibriya, H., Siddiqa, A., Khan, W. Z., & Khan, M. K. (2024). Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification. *Computers and Electrical Engineering*, *116*, 109153. https://doi.org/10.1016/j.compeleceng.2024.109153

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1705.07874

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), Article 17. https://doi.org/10.1609/aaai.v35i17.17745

Mihailo Škorić. (2024). Нови језички модели за српски језик. *Infoteka*, *год. 24, бр. 1, фебруар 2024*, 1–22.

Mladenović, M., Krstev, C., Mitrović, J., & Stanković, R. (2017). Using Lexical Resources for Irony and Sarcasm Classification. *Proceedings of the 8th Balkan Conference in Informatics*, 1–8. BCI '17: 8th Balkan Conference in Informatics. https://doi.org/10.1145/3136273.3136298

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2022). Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning. *IEEE Access*, *10*, 14880–14896. https://doi.org/10.1109/access.2022.3147588

Mozafari, M., Mnassri, K., Farahbakhsh, R., & Crespi, N. (2024). Offensive language detection in low resource languages: A use case of Persian language. *PLOS ONE*, *19*(6), e0304166. https://doi.org/10.1371/journal.pone.0304166

Paraschiv, A., Ion, T. A., & Dascalu, M. (2024). Offensive Text Span Detection in Romanian Comments Using Large Language Models. *Information*, *15*(1), Article 1. https://doi.org/10.3390/info15010008

Pavlopoulos, J., Sorensen, J., Laugier, L., & Androutsopoulos, I. (2021). SemEval-2021 Task 5: Toxic Spans Detection. In A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 59–69). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.semeval-1.6

Piot, P., & Parapar, J. (2025). Towards Efficient and Explainable Hate Speech Detection via Model Distillation. *European Conference on Information Retrieval*, 376–392.

Rajabi, E., & Etminani, K. (2024). Knowledge-graph-based explainable AI: A systematic review. *Journal of Information Science*, *50*(4), 1019–1029. https://doi.org/10.1177/01655515221112844

Ranasinghe, T. (2020). *Models of Irony detection in Natural Language Processing*. *11*(3).

Ranasinghe, T., Anuradha, I., Premasiri, D., Silva, K., Hettiarachchi, H., Uyangodage, L., & Zampieri, M. (2024). SOLD: Sinhala offensive language dataset. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-024-09723-1

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. https://doi.org/10.18653/v1/N16-3020

Saleem, H. M., Dillon, K., Benesch, S., & Ruths, D. (2017). A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *ArXiv*. https://www.semanticscholar.org/paper/A-Web-of-Hate%3A-Tackling-Hateful-Speech-in-Online-Saleem-Dillon/249d1b13cbffdcab9ccef4a83126ee222641dc82

Song, P., Ojo, A., & Curry, E. (2024). *Towards Trustworthy Foundation Models: A Survey*. Elsevier BV. https://doi.org/10.2139/ssrn.4985376

Šošić, M., Graovac, J., & Stanković, R. (2026). Building an emotion lexicon for Serbian using curated language resources. *Language Resources and Evaluation*, *60*(1), 9. https://doi.org/10.1007/s10579-025-09894-5

Stanković, R., Mitrović, J., Jokić, D., & Krstev, C. (2020). Multi-word Expressions for Abusive Speech Detection in Serbian. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova, & A. Savary (Eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (pp. 74–84). Association for Computational Linguistics. https://aclanthology.org/2020.mwe-1.10

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor, A. Oh, O. Tsur, & S. Volkova (Eds.), *Proceedings of the First Workshop on NLP and Computational Social Science* (pp. 138–142). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5618

Zampieri, M., Ranasinghe, T., Chaudhari, M., Gaikwad, S., Krishna, P., Nene, M., & Paygude, S. (2022). *Predicting the Type and Target of Offensive Social Media Posts in Marathi* (arXiv:2211.12570). arXiv. https://doi.org/10.48550/arXiv.2211.12570

10. Закључак и предлог комисије

На основу изнетих података Комисија сматра да је предложена тема докторске дисертације "**Хибридни модел за препознавање увредљивог говора у кратким текстовима написаним на природним језицима са ограниченим ресурсима**" научно заснована и актуелна и да кандидат Данка Јокић, мастер инжењер електротехнике и магистар наука пословног управљања, испуњава услове за рад на овој докторској тези. Комисија предлаже Већу за студије при Универзитету у Београду да прихвати тему и кандидату Данки Јокић одобри израду докторске дисертације под наведеним насловом. За менторе се предлажу др Ранка Станковић, редовни професор, Универзитета у Београду, Рударско-геолошки факултет (ужа научна област: математика и информатика) и др Јелена Граовац, ванредни професор, Универзитета у Београду Математички факултет (ужа научна област: рачунарство и информатика).

Београд, 3.4.2026. године

Потпис чланова комисије

Др Владан Девецић, академик, редовни професор,
Универзитет у Београду - Факултет организационих наука
(софтверско инжењерство)

Др Ана Ковачевић, редовни професор,
Универзитет у Београду - Факултет безбедности
(информатика)

Др Михаило Шкорић, научни сарадник,
Универзитет у Београду - Рударско-геолошки факултет
(информационе технологије)