Наставно-научном већу
Математичког факултета
Универзитета у Београду

Одлуком Наставно-научног већа Математичког факултета Универзитета у Београду донетом на 366. седници одржаној 22.11.2019. године именовани смо за чланове Комисије за преглед и оцену докторске дисертације **„Утицај класификације текста на примене у обради природних језика“** кандидата Браниславе Шандрих. После прегледа поднетог рукописа подносимо следећи

ИЗВЕШТАЈ

БИОГРАФИЈА КАНДИДАТА

Бранислава Шандрих рођена је 19.08.1991. у Панчеву. Након завршене средње Електротехничке школе „Никола Тесла“ у Панчеву, 2010. године уписала је основне студије на Математичком факултету. Четири године касније, 2014. године, стекла је звање дипломирани математичар на смеру Рачунарство и информатика - просечна оцена 9.48, а годину дана касније стекла је звање мастер математичар на истом смеру са просечном оценом 9.80. Мастер тезу под називом „Рефакторисање кода према обрасцима дизајна“ написала је под руководоством др Владимира Филиповића. Године 2015. уписала је докторске студије на смеру Информатика на Математичком факултету. Положила је све испите предвиђене планом и програмом докторских студија са просечном оценом 10.00.

Бранислава Шандрих је запослена као асистент на Катедри за библиотекарство и информатику при Филолошком факултету у Београду где изводи вежбе на основним и мастер студијама из следећих предмета: Информатика за библиотекаре 1 и 2, Информатички практикум 1, 2, 3 и 4, Дигитални текст 1, Језичке технологије 1 и 2, Мултимедијални документи, Проналажење информација, Напредне методе проналажења информација, Формирање и обликовање садржаја на вебу.

Током студија имала је стручну праксу у компанији SAP SE, Валдорф, Немачка. Учествовала је у пет иностраних студијских боравака. Учесник је једног научно истраживачког пројекта и 3 COST акције: 1) Српски језик и његови ресурси: теорија, опис и примене; 2) COST акција CA16105 European Network for Combining Language Learning with Crowdsourcing Techniques; 3) COST акција CA18231 Multi3Generation: Multi-task, Multilingual, Multimodal Language Generation; 4) COST акција CA16204 – Distant Reading for European Literary History. Током студија добила је већи број награда и стипендија: Стипендија немачке привреде „др Зоран Ђинђић“, Доситијева стипендија Фонда за младе таленте, Награда за најбоље студенте града Панчева (3 пута), Награда за најбоље студенте Математичког факултета и Стипендија Министарства просвете, науке и технолошког развоја (3 пута).

ОБЈАВЉЕНИ НАУЧНИ РАДОВИ

Бранислава Шандрих је до сада објавила 9 радова у научним часописима од чега су 2 у часописима са SCI листе (један у категорији рачунарских наука). У 8 од ових 9 радова Бранислава Шандрих је први аутор, а у 2 рада је једини аутор. Радови означени бројевима од 1 до 6 су релевантни за дисертацију.

1. Branislava Šandrih, Cvetana Krstev, and Ranka Stanković. Two Approaches to Compilation of Bilingual Multi-Word Terminology Lists from Lexical Resources. Natural Language Engineering, 2020. doi 10.1017/S1351324919000615. **(M23, IF 2018 = 1.130, Computer Science & Artificial Intelligence)**

2. Jelena Andonovski, Branislava Šandrih, and Olivera Kitanović. Bilingual Lexical Extraction based on Word Alignment for Improving Corpus Search. *The Electronic Library*, 37 (4), 722–739, 2019. **(M22, IF 2018 = 0.886, Information Science & Library Science)**

3. Branislava Šandrih. SMS Sentiment Classification based on Lexical Features, Emoticons and Informal Abbreviations. Serdica Journal of Computing, 13 (1-2), 81-94, 2019. **(M53)**

4. Бранислава Шандрих, Ранка Станковић и Мирјана Гочанин. Чији је пример? Анализа лексичких обележја на примерима Речника САНУ. *Научни сасстанак слависта у Вукове дане – Српски језик и његови ресурси: теорија, опис и примене*. Међународни славистички центар, Београд, Vol. 48 (3), 299–316, 2019**. (M53**)

5. Branislava Šandrih, and Ranka Stanković, Extraction of Bilingual Terminology using Graphs, Dictionaries and GIZA++, Infotheca – Journal for Digital Humanities, 19 (2), 2019. **(M53)**

6. Бранислава Шандрих и Душко Витас. Квантитативни преглед језика кратких порука. *Научни састанак слависта у Вукове дане – Српски језик и његови ресурси: теорија, опис и примене.* Међународни славистички центар, Београд, Vol. 47 (3), 155–165, 2018. **(M53)**

7. Branislava Šandrih. Informatics for Library and Information Science students with special focus on Python. Infotheca - Journal for Digital Humanities, 18 (1), 2018. **(M53)**

8. Branislava Šandrih, Dušan Tošić, and Vladimir Filipović. Towards Efficient and Unified XML/JSON Conversion – a New Conversion Method. Transactions on Internet Research (TIR) 13 (1), 58–64, 2017. **(M53)**

9. Branislava Šandrih, Vladimir Filipović, Saša Malkov, and Aleksandar Kartelj. Distributed Computing among independent Web Browsers applied to Text and Image Processing. Review of the National Center for Digitization, (31), 30–39, 2017. **(M53)**

ОБЈАВЉЕНА САОПШТЕЊА СА НАУЧНИХ СКУПОВА

Бранислава Шандрих до сада има 9 саопштења са научних скупова штампаних у целини или у изводу. Сва саопштења осим оног под редним бројем 6 су релевантна за дисертацију.

1. Ranka Stanković, Branislava Šandrih, Rada Stijović, Cvetana Krstev, Duško Vitas, and Aleksandra Marković. SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian. In eLex 2019: Smart Lexicography, Sintra, Portugal, 248–269, 2019. **(M33)**

2. Branislava Šandrih, Cvetana Krstev, and Ranka Stanković. Development and Evaluation of Three Named Entity Recognition Systems for Serbian – the Case of Personal Names. In RANLP 2019: Recent Advances in Natural Language Processing, Varna, Bulgaria, 1060–1068, 2019. **(M33)**

3. Branislava Šandrih. Fingerprints in SMS messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting. In 3rd International Conference Computational Linguistics in Bulgaria (CLIB 2018), Department of Computational Linguistics at the Institute for Bulgarian Language with the Bulgarian Academy of Sciences, Sofia, Bulgaria, 203–210, 2018. **(M33)**

4. Cvetana Krstev, Branislava Šandrih, Ranka Stanković, and Miljana Mladenović. Using English Baits to Catch Serbian Multi-Word Terminology. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, France, 2487–2494, 2018. **(M33)**

5. Cvetana Krstev, Duško Vitas, Miloš Utvić, and Branislava Šandrih. The New Clothes for an Old Cookbook. In 8th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2017), Poznan, Poland, 174-178, 2017. **(M33)**

6. Ranka Stanković, Branislava Šandrih, Olivera Kitanović, Ivan Obradović, and Miloš Manić. An E-Learning Approach to Social Sciences. In The 8th International Conference on eLearning (eLearning-2017), Belgrade, Serbia, 26–29, 2017. **(M33)**

7. Peter Dekker, Tanara Zingano Kuhn, Branislava Šandrih, and Rina Zviel-Girshin. Corpus Cleaning via Crowdsourcing for Developing a Learner's Dictionary. In eLex 2019: Smart Lexicography, Sintra, Portugal, 84-85, 2019. **(M34)**

8. Tanara Zingano Kuhn, Peter Dekker, Branislava Šandrih, Rina Zviel-Girshin, Spela Arhar Holdt, and Tanneke Schoonheim. Crowdsourcing Corpus Cleaning for Language Learning Resource Development. In EuroCALL 2019: European Association of Computer Assisted Language Learning, Louvain-la-Neuve, Belgium, p. 159, 2019. **(M34)**

9. Branislava Šandrih. SMS Sentiment Classification based on Emoticons, Informal Abbreviations and other Text Features. In International Quantitative Linguistics Conference (QUALICO 2018). Institute of Library and Information Science / Faculty of Mathematics and Computer Science (University of Wroclaw), Wroclaw, Poland, 2018. **(M34)**

ПРЕДМЕТ ДИСЕРТАЦИЈЕ

Предмет докторске дисертације припада областима обраде природних језика и машинског учења. У фокусу дисертације је проблем класификације текста и његове примене у решавању значајних проблема обраде природних језика. Заједнички оквир за решавање проблема класификације је заснован на прилагођеном простору лингвистичких атрибута који се показао као добра основа за решавање неких проблема из домена обраде природних језика.

Како се терминологија у оквиру научних области данас првенствено успоставља на енглеском језику, јавља се потреба за стандардизованом терминологијом на другим језицима која иде у корак са појавом нових доменских термина. Ручно прављење билингвалног речника појмова захтева интензиван рад доменских и језичких стручњака. У дисертацији је предложен систем за аутоматску екстракцију и валидацију доменске терминологије и његов рад је демонстриран на примеру успостављања билингвалне енглеско-српске листе доменских парова појмова. За потребе класификације добрих/лоших билингвалних парова предложена је метода подржавајућих вектора (енг. Support vector machine) са линеарним и радијалним језгром.

Илустрација употребе речничке одреднице на одговарајућем скупу примера употребе кључна је за добро разумевање значења, како за говорнике, тако и за оне који тек уче нови језик. Примери употребе језика су свугде присутни: у дневној штампи, на друштвеним мрежама, на

форумима, у књижевним делима и слично. Али нису сви примери једнако добри. У дисертацији је предложен приступ у којем се најпре врши формирање скупа података провереног од стране лингвистичких експерата. У том скупу су примери употребе речничких одредница, након вишеструких фаза припреме, обележени као неадекватни или адекватни. Такав скуп података употребљен је као улаз за обучавање стабла одлучивања. Постојање овако обученог бинарног класификационог модела омогућава класификацију нових примера употребе у будућности. Иако су експериментални резултати везани за српски језик, само језгро предложеног приступа је могуће проширити тако да ради и са другим језицима.

Проблем идентификације ауторства бави се препознавањем аутора произвољног текстуалног документа. У овом случају, важно је препознати скуп атрибута који најбоље карактеришу дати документ и који га дискриминишу у односу на остале документе те колекције. У тези је предложена: 1) метода за класификацију СМС порука према томе да ли је њен аутор једна конкретна особа или неко други и 2) метода за класификацију СМС порука према томе да ли је њен аутор особа или је порука аутоматски генерисана од стране рачунара. У циљу добијања што квалитетнијих резултата, употребљен је велики број емпиријски одређених лингвистичких атрибута из група: лексичких атрибута (нпр. бројеви карактера и интерпункцијских симбола); синтаксних атрибута (нпр. емотограми и скраћенице) и стилских атрибута (нпр. употреба малог слова на почетку реченице и учесталост употребе негације).

Анализа ставова и расположења код кратких порука представља својеврстан изазов, јер такве поруке носе значајно мању количину информација у односу на, на пример, форумске дискусије. Други проблем код кратких порука лежи и у њиховој формулацији. Уочен је тренд да аутори кратких порука теже да кроз специфичну употребу знакова интерпункције, понављања слова и симбола, писањем само великим словима, употребом специјализованих скраћеница и емотограма, остварују сличност писаног са говорним језиком. На тај начин, аутори изражавају себе, свој став и своја осећања. У тези су предложени различити класификациони модели попут методе подржавајућих вектора, к-најближих суседа, стабала одлучивања над скупом СМС порука аутора означених у складу са типом расположења које у себи носе. Простор атрибута је прилагођен кратким порукама и састоји се од лексичких, синтаксних и стилских атрибута.

У дисертацији је такође реализован веб сервис који на улазу очекује произвољан текст, а као излаз производи асоцијативни низ парова кључ-вредност, при чему су кључеви називи лингвистичких атрибута, а њихове вредности представљају вредности одговарајућих атрибута. Веб сервис се користи за корак припреме скупа података за обучавање поменутих класификационих метода.

ПРИКАЗ ДИСЕРТАЦИЈЕ

Рукопис има 145 страна и састоји се од 6 поглавља основног текста, списка коришћене литературе од 245 библиографских јединица, листе скраћеница и прилога. Рукопис је писан на енглеском језику и има следећу структуру.

У уводном поглављу су описане математичке основе надгледаног учења са специјалним фокусом на проблем класификације. Ту су размотрени различити модели за класификацију података попут методе подржавајућих вектора, стабала одлучивања, методе градијентног појачавања и друге. Потом су уведени проблеми обраде природних језика и мотивисана је примена класификације текста у њиховом решавању. Специјални фокус је посвећен групама различитих атрибута који се користе за репрезентацију текста. Такође, у овом поглављу је направљен осврт на неке важне теме из области обраде природних језика које су неопходне за

даље разумевање текста тезе, а дат је и кратак увид у расположиве алате и ресурсе за обраду текстова на српском језику.

У другом поглављу је представљен проблем екстракције и валидације билингвалних парова лексема и преглед релевантних резултата из литературе, а потом и технике аутоматског извлачења терминологије из постојећих паралелних текстова на различитим природним језицима. Након тога је описана предложена метода за валидацију добрих/лоших парова билингвалних лексема и преглед експерименталних резултата са дискусијом.

Треће поглавље је везано за проблем класификације добрих примера употребе речничких одредница. Након увођења математичке формулације и прегледа релевантних резултата из литературе, уследио је опис предложене методе засноване на стаблу одлучивања за решавање овог проблема као и преглед добијених резултата са дискусијом.

У четвртом поглављу су представљене различите формулације проблема препознавања ауторства и преглед метода за њихово решавање. Након тога је додатни фокус посвећен специфичностима ових проблема када су у питању кратки текстови због чега је дат предлог употребе прилагођеног скупа лингвистичких атрибута. Потом је уследио опис предложене методе засноване на градијентном појачавању и преглед експерименталних резултата са дискусијом.

Пето поглавље се односи на класификацију кратких порука у односу на став и расположење. Након увођења математичке формулације проблема и прегледа релевантних доприноса из литературе, описан је скуп погодних лингвистичких атрибута над којима се даље примењује класификација. Предложена метода заснована на подржавајућим векторима је експериментално испитана након чега је дата дискусија добијених резултата.

У завршном, закључном поглављу, сумиран је значај примене класификације текста у решавању проблема обраде природних језика као и доприноси тезе на том пољу. Такође су наведени и неки правци будућег истраживачког рада.

НАУЧНИ ДОПРИНОС ДИСЕРТАЦИЈЕ

Сажети опис остварених научних доприноса је следећи:

- Конструисан је погодан скуп лингвистичких атрибута у који се текст пресликава, са циљем побољшања класификације текста у контексту обраде природних језика;
- Предложен је нови приступ који користи методу подржавајућих вектора (енг. Support vector machine) са линеарним и радијалним језгром за аутоматско генерисање доменског, енглеско-српског речника појмова;
- Предложена је техника заснована на стаблима одлучивања за класификацију добрих и лоших примера употребе речничких одредница као значајан корак ка даљем аутоматском креирању савремених електронских речника;
- Предложена је метода градијентног појачавања (енг. Gradient boost) за препознавање аутора у кратким порукама;
- Предложена је метода за анализу ставова и расположења у кратким порукама заснована на методи подржавајућих вектора.

ЗАКЉУЧАК И ПРЕДЛОГ КОМИСИЈЕ

У разматраном рукопису кандидат Бранислава Шандрих је показала темељно, али и широко знање области обраде природних језика. Резултати добијени у тези представљају доприносе области обраде природних језика, а посебно обради српског језика. Иако су демонстрације предложених метода направљене у контексту српског језика, све предложене методе поседују општост која им омогућава да уз уграђивање додатних језичких ресурса, функционишу и на вишејезичном нивоу. Сви представљени појединачни резултати су претходно објављени у међународним научним часописима или представљају саопштења са међународних конференција.

На основу свега наведеног, као и на основу испуњености свих формалних услова, предлажемо Наставно-научном већу Математичког факултета Универзитета у Београду да рукопис *„Утицај класификације текста на примене у обради природних језика“* (*“Impact of Text Classification on Natural Language Processing Applications”)* кандидата Браниславе Шандрих, прихвати као докторску дисертацију и одреди комисију за њену јавну одбрану.

У Београду, 27.04.2020. године

Комисија:

_____
др Александар Картељ, доцент (ментор)

_____
др Гордана Павловић-Лажетић, редовни професор

_____
др Владимир Филиповић, редовни професор

_____
др Цветана Крстев, редовни професор
Филолошки факултет, Универзитет у Београду

_____
др Руслан Митков, редовни професор,
Универзитет у Вулверхемптону, Енглеска
(Dr Ruslan Mitkov, full professor,
University of Wolverhampton, England)