Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/23BE04Budjen

АНОТИРАЊЕ И ОДРЕЂИВАЊЕ СЕНТИМЕНАТА ТВИТОВА ВЕЗАНИХ ЗА ПОЛИТИЧКУ СЦЕНУ РЕПУБЛИКЕ СРБИЈЕ

ANNOTATION AND SENTIMENT ANALYSIS OF TWEETS RELATED TO THE POLITICAL SCENE OF THE REPUBLIC OF SERBIA

Владимир Буђен, *Факултет техничких наука, Нови Сад*

Област – ЕЛЕКТРОТЕХНИКА И РАЧУНАРСТВО

Кратак садржај – *У раду је приказано одређивање сентимента твитова и креирање алата који одређује позицију аутора твитова на политичком компасу. Циљ рада је асистеннција корисницима при политичким изборима. За одређивање сентимента твита коришћен је BERTić модел. Добијени сентимент, у комбинацији са темом твита, је коришћен као улаз у модел SVM и Random Forest моделе који служе за поларизацију аутора твита.*

Кључне речи: *твитер, сентимент, политика, вештачка интелигенција, BERT*

Abstract – *This paper describes tweet sentiment analysis and the development of a tool for placing tweet authors on a political map. The purpose of this system is to assist its users in political elections. BERTić model is used for tweet sentiment analysis. Sentiment and tweet's theme are used as inputs for SVM and Random-forest, which are used for user polarization.*

Keywords *tweeter, politics, artificial intelligence, sentiment, BERT*

1. УВОД

Друштвена мрежа „Twitter" својим корисницима нуди лак и јефтин начин да поделе лично мишљење великом броју људи. Због тога велики број корисника ове мреже чине политичке партије и удружења, као и њихови чланови. Објаве на твитеру су ограничене бројем карактера и због тога су погодне за анализу.

Једна од информација коју бисмо могли да сазнамо из твитова политички оријентисаних профила јесте позиција тих профила на политичком компасу. Та информација може да помогне гласачком телу да има реалнији увид у тачну политичку оријентацију странака. Потреба за тим постоји јер многе партије, због жеље да привуку што више људи различитих идеологије, не спомињу своје конкретно место на политичком компасу. Такође, постоје случајеви да странке потенцирају нестандардне идеје у односу на оријентацију којом се представљају.

Како би се аутоматизовао процес идентификације политичке оријентације аутора твитова, потребно је користити технике вештачке интелигенције. Као циљно обележје коришћени су сентимент твита, тема твита и позиција аутора твита на политичком компасу. Скуп података коришћен за обучавање ових модела је прикупљен са твитера и ручно анотиран од стране аутора рада. Текстови твитова су трансформисани у број који представља њихов сентимент. Налози корисника су искључиво са територије Републике Србије, што представља додатни изазов при одређивању сентимената, за шта је коришћен претренирани вишејезични BERT модел [1].

За предикцију сентимента твита коришћен је BERT, док су за поларизацију коришћени метод потпорних вектора (енг. *Support Vector Machine,* SVM) и модел насумичне шуме (енг. *Random Forest*). Излаз из SVM модела је позиција аутора твита са прецизношћу 0.1 на скали од -1 до 1, док је прецизност *Random Forest*-а 0.25.

Евалуацијом решења пројекта се може видети како политичке оријентације пишу твитове чији сентимент зависи од дате теме. Постоје теме у којим већина корисника испољавају исти сентимент, оне у којима нема корелације између оријентације и сентимента и оне чијим сентиментом се лако одреди позиција на политичком компасу (и обрнуто).

Тешко је директно поредити резултате приказане у овом раду са сличним радовима услед тога што користе различите скупове података. Међутим, за теме које се преклапају као и у овом раду, добијени су слични закључци о корелацији те теме и политичке оријентације корисника.

2. ПРЕТХОДНА РЕШЕЊА

У овом поглављу бавићемо се радовима сличне тематике, као и самом еволуцијом области и значајним резултатима и увидима.

Први значајан рад је „*Predicting the Political Alignment of Twitter Users"* [2]. Аутори овог рада се баве анализом твитова у циљу одређивања политичке оријентације корисника по економској скали на подручју Сједињених Америчких Држава у периоду од 14.9.2010. до 1.11.2010. Анализа твитова је извршена на два начина: на основу садржаја и на основу међукорисничких интеракција. На релевантност твитова је утицала појава хештегова (енг. *hashtag*). За класификацију је коришћен SVM, што је један од разлога због чега је тај модел коришћен и у експерименту који овај рад описује. Овај рад пријављује тачност од 0,92.

НАПОМЕНА:

Овај рад проистекао је из мастер рада чији ментор је била др Јелена Сливка, ванр. проф.

Још један значајан рад из ове области је „*A Sentiment Analysis System of Spanish Tweets and Its Application in Colombia 2014 Presidential Election*“ [3]. Идеја рада је базирана на анализи политичке струје у Колумбији 2014. године, с циљем предвиђања резултата председничких избора. Овај рад такође користи хештегове, али поред тога користи и ручно изабране кључне речи. Такође, један део података, везан за кориснике, је ручно анотиран. Подацима је додељен сентимент од стране волонтера и кроз јавне анкете, што је позитивно утицало на резултате. Због малог броја хештегова у политичким твитовима на српском језику, идеја да се користе изабране кључне речи је коришћена и у експерименту везаном за овај рад. Хештегови који су коришћени у наведеним радовима послужили су за бирање кључних речи у овом раду. Рад [3] пријављује F1 меру од 0,58.

3. МЕТОД

У наредним поглављима изложени су скуп података и начин на који је спроведен експеримент.

3.1. Скуп података

Почетна фаза експеримента је креирање упита који ће бити кориштени за прикупљање жељених твитова. Упити у себи садрже речи које желимо да се појаве у твитовима, као и кориснике које желимо да буду аутори. Речи које су кориштене у твитовима су груписане по темама. Листа аутора и тема је ручно креирана. Узор за одабир тема су радови „*Predicting the Political Alignment of Twitter Users“* [2] и „A Sentiment Analysis System of Spanish Tweets and Its Application in Colombia 2014 Presidential Election“ [3]. Конкретне теме коришћене за експеримент су:

- Косово
- Албанија
- Војска
- Црква
- ЛГБТ
- Београд
- Полиција
- Корупција
- Европа
- Украјина
- Путин
- Русија
- Нато
- Америка

Такође, политичка позиција аутора твитова (самим тим и њихових твитова) је ручно креирана. Она је представљена вредностима на социјалној и на економској скали. Вредност је број у распону од -1 до 1 са кораком 0.1. Политичка позиција је битна за фазу обучавања модела. Подаци су подељени на обучавајући и тест скуп у односу 1:5.

3.2. Одређивање сентимента твита

Прикупљање података даје информације о теми, садржају и осталим корисним информацијама. Проблем је у томе што о одређеној теми различити корисници говоре у различитом контексту, што представља битну информацију при одређивању политичког опредељења. Зато је потребно добити информацију да ли се тема у садржају спомиње у позитивном, негативном или неутралном контексту. За ово се користе претренирани модели за одређивање сентимента.

У овом експерименту коришћена је *fine-tune*-ована верзија Бертића [4], специјализована за одређивање сентимента твитова. Овај модел је доступан под називом *EMBEDDIA/bertic-tweetsentiment* на *huggingface* репозиторијуму [5]. Његов улаз је текст писан латиничним писмом.

Због овога је потребно ћирилични текст превести у латинични, за шта је коришћена библиотека „*SrbAi*“ [6]. Излаз из модела је сентимент (енг. *sentiment_label*) и сигурност у добијени резултат изражен бројем (енг. *Sentiment_score*).

3.3. Обучавање модела

Прикупљени подаци коришћени су за обучавање више модела применом два приступа. Оба приступа креирају посебне моделе за економску и за социјалну скалу. За креирање свих модела коришћена је „*scikit-learn*“ [7] библиотека.

Први приступ користи SVM, чији је вектор улаза сачињен од сентимената тема сваког појединачног твита. За позитиван сентимент је дат број 1, негативан -1, а неутралан сентимент 0.5. За све теме које нису везане за изабрани твит, унета је вредност 0. Разлог због којег је одабран овакав коефицијент за неутралан сентимент твита јесте могућност разликовања оних корисника који су имали неутрално мишљење о некој теми, од корисника који се о њој нису изјаснили.

Пошто су сви твитови везани за тачно једну тему улаз у модел је *one-hot* вектор. Излаз, као и у првом приступу, јесте вредност на политичкој скали (један од бројева између -1 и 1 са скоком 0,1). Исти принцип је коришћен при креирању *Random Forest*-а. Једина разлика јесте у томе што су узете вредности на скали са кораком 0,5 уместо 0,1.

Други приступ се разликује од претходног по самом начину формирања вектора, који се овог пута везује за корисника уместо за сам твит. Наиме, вектор се формира за сваког корисника, а као вредности садржи корисниково укупно мишљење на задату тему, при чему теме представљају димензије вектора. Корисниково мишљење на једну тему одређује се по формули:

$$\frac{k_{pos} * n_{pos} + k_{neg} * n_{neg} + k_{neu} * n_{neu}}{n},$$

где k представља коефицијент, а n укупан број твитова одговарајућег сентимента. Вредности коефицијента су 1 за позитиван сентимент, -1 за негативан и 0.5 за неутралан. Овако формирани вектори, употребљени су за обучавање две SVM (један за економски аспект и други за социјални аспект) [8]. За оптимизацију параметара коришћена је класа *GridSearchCV* из *scikit-learn* [7] библиотеке.

4. РЕЗУЛТАТИ И ДИСКУСИЈА

У овом поглављу су приказани експлоративна анализа, добијени резултати експеримента и предочене потенцијалне мане и предложена побољшања.

4.1. Експлоративна анализа

Обрада сентимента је показала да већина твитова политичке природе има негативан сентимент. Конкретно, од 3192 преузета твита, 1926 су негативног, 979 су позитивног, а 287 неутралног сентимента. Након вршења

анализе сентимента, могуће је видети да су неке теме поларизованије од других. Примери таквих тема су:

- Путин
- LGBT
- Русија
- Војска

4.2. Резултати експеримента

Резултати експеримента су приказани у табелама 4.1 – 4.3, где С представља праве координате на социјалној скали, Е представља праве координате на економској скали, С.П представа предикцију координата на социјалној скали, Е.П представа предикцију координата на економској скали, С.Г представа грешку на социјалној скали, Е.Г представа грешку на економској скали. Грешка се рачуна као апсолутна вредност разлике координата. У табелама су приказани насумично одабрани корисници из тест скупа података.

Резултат првог приступа при коришћењу SVM је приказан у табели 4.1. Просечна вредност грешке за економску скалу је 0,27, док је за социјалну скалу вредност 0,57.

Табела 4.1 *Мере евалуације првог приступа и коришћењем svm*

Аутор	С	Е	С.П	Е.П	С.Г	Е.Г
Марко Ђурић	0,9	0,3	0,3	0,2	0,6	**0,1**
Ана Брнабић	0,5	0,1	0,2	0	0,3	**0,1**
Небојша Стефановић	0,8	0,5	0,4	0,2	0,4	0,3
Александар Шапић	0,4	0	0,1	0,1	0,3	**0,1**

Први приступ при коришћењу *Random Forest* модела је тачно погодио половине квадранта у 48,27% података на економској и 37,93% на социјалној скали. Резултати овог приступа су приказани у табели 4.2.

Табела 4.2 *мере евалуације првог приступа и коришћењем Random forest ансамбла*

Аутор	С	Е	С.П	Е.П	С.Г	Е.Г
Марко Ђурић	0,9	0,3	0,3	0,2	0,6	**0,1**
Ана Брнабић	0,5	0,1	0,2	0	0,3	**0,1**
Небојша Стефановић	0,8	0,5	0,4	0,2	0,4	0,3
Александар Шапић	0,4	0	0,1	0,1	0,3	**0,1**

Други приступ, у ком су вектори формирани по упросеченим сентиментима тема појединачних корисника је одступао на економској скали за 0,28 у просеку, док је за социјалну скалу просечно одступање износило 0,39. Резултати овог приступа су приказани у табели 4.3.

Табела 4.3 *мере евалуације другог приступа*

Аутор	С	Е	С.П	Е.П	С.Г	Е.Г
Марко Ђурић	0,49	0,33	0,9	0,3	0,41	0,03
Ана Брнабић	0,54	0,45	0,5	0,1	0,04	0,35
Небојша Стефановић	0,51	0,31	0,8	0,5	0,29	0,19
Александар Шапић	0,14	0,19	0,4	0	0,26	0,19

4.3. Анализа грешака и потенцијална побољшања

Највећа мана система јесте недовољно велики скуп података. Укупан број твитова је 3191. Ово је мали број када се подели на 14 значајних тема. У просеку, то чини 228 твитова по теми. Последица овога је недовољно тренирани модел, као и немогућност балансирања скупа података.

Други извор грешака је то што претрага твитова по темама, за проналажење теме има једини услов да се тема спомиње у тексту, независно од контекста. Ово доводи до тога да тема није примарна ствар у твиту. Пошто се сентимент врши над целим садржајем твита, веза између теме и сентимента је врло слаба. Потенцијално решење за овај проблем јесте коришћење хештегова за избор теме. Овакво решење би било могуће у случају да постоји већа количина података и да их корисници више користе.

Садржај твита често садржи више реченица од којих су неке позитивног, а неке негативног сентимента. Један од начина на који ово може да се реши јесте филтрирањем података, тако да се користе само они твитови, чија вредност сигурности у сентимент (*sentiment_score*) је преко 0,9. Ово долази са ценом смањења скупа података.

Једна од највећих слабости система јесте и та што су вредности позиција на политичком компасу ручно анотирана од стране само једне особе, која притом није уско специјализована за то. Најбољи начин да се ово реши, јесте да се за ово ангажује више стручњака и да се вредности добију упросечавањем вредности које они задају.

4.4. Употребљивост система

Алат развијен овим експериментом се користи у ситуацијама када је потребно добити политичку позицију аутора твитова. Његова велика предност је у томе што је намењен српској популацији. Иако тренутно ради само са српским језиком, могућност проширења на друге језике није комплексна. Друга ствар која чини овај алат другачијим од осталих јесте квантификовани излаз, који је могуће користити као улаз за креирање других модела.

У поређењу са резултатима добијеним од стране људи, модел се показао бољим код људи са недовољним знањем о политици. Доказ о овоме јесте мала грешка при тестирању. Треба имати у виду да је и сам тест скуп ручно анотиран. Због овога модел ни у ком случају не може да превазиђе резултате који би дали људи специјализовани за политичке науке.

Имајући то у виду, тренутна употребљивост модела не улази у професионални домен. Помаже корисницима који нису у стању да сами одреде политичку позицију твита, али никако не може да замени мишљења професионалаца.

5. ЗАКЉУЧАК

У овом раду представљен је систем за одређивање политичке позиције аутора твитова. Мотивација за креирање овог система је била спознаја да би такав систем помогао корисницима да лакше донесу одлуку при гласању на политичким изборима. Систем је им-

плементиран из модула за одређивање сентимента твита и модула који садржи модел за препознавање политичке оријентације аутора твита.

Модул за одређивање сентимента твита је имплементиран коришћењем BERT-*ић* модела. Његовим коришћењем су добијени сентименти твитова. Комбинација добијеног сентимента и теме твита су информације потребне за одређивање политичке позиције. Тај процес се врши у другом модулу коришћењем модела SVM. Излаз из система су пар бројева који представљају координате позиције на политичком компасу.

Приступ у ком су вектори формирани по упросеченим сентиментима одређених тема појединачних корисника је дао најбоље резултате. Он је одступао на економској скали за 0,28 у просеку, док је за социјалну скалу просечно одступање износило 0,39. Директног поређења са другим радовима не може бити јер је коришћен нови скуп података са српског подручја, док су постојећи радови са подручја Сједињених Америчких Држава и Републике Колумбије. Још једна од препрека у поређењу резултата јесте та што су остали радови вршили бинарну класификацију, док је у овом раду вршена мултикласна класификација. Ово је разлог коришћења различитих мера евалуације. Међутим, поређења ради, рад [20] пријављује F1 меру од 0,58, а рад [19] тачност од 0,92.

Овом систему остаје простора за унапређење. Највише побољшање је могуће добити проширивањем скупа података, што би отворило могућност бољег балансирања скупа података по темама и корисницима, самим тим бољих резултата.

6. ЛИТЕРАТУРА

[1] J. D. M.-W. C. K. L. K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“.

[2] B. G. J. R. A. F. F. M. Michael D. Conover, Predicting the Political Alignment of Twitter Users.

[3] E. L.-G. Jhon Adrian Ceron-Guzman, A Sentiment Analysis System of Spanish Tweets and Its Application in Colombia 2014 Presidential Election.

[4] D. L. Nikola Ljubešić, „BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian“.

[5] „EMBEDDIA/bertic-tweetsentimen,“ [На мрежи]. Available: https://huggingface.co/EMBEDDIA/bertic-tweetsentimen.

[6] „SrbAi,“ [На мрежи]. Available: https://github.com/Serbian-AI-Society/SrbAI.

[7] „Scikit-learn,“ [На мрежи]. Available: https://scikit-learn.org/stable/.

[8] М. Кнежевић, „Позиционирање корисника друштвене мреже Twitter на мапи политичког спектра помоћу корисничких твитова“.

Кратка биографија:

Владимир Буђен рођен је 9.5.1998. године у Новом Саду, где је стекао своје основно и средње образовање. Школске 2017/18 године се уписисује на Факултет техничких наука на студијски програм Рачунарство и аутоматика, који је завршио школске 2021. године. Исте године и на истом факултету уписује мастер студије, конкретно, програм Електронско пословање. Положио је све испите предвиђене планом и програмом и стекао услов за одбрану завршног рада.