Зборник радова Факултета техничких наука, Нови Сад

УДК: 004.85
DOI: https://doi.org/10.24867/20BE06Knezevic

ПОЗИЦИОНИРАЊЕ КОРИСНИКА ДРУШТВЕНЕ МРЕЖЕ ТВИТЕР НА МАПИ ПОЛИТИЧКОГ СПЕКТРА ПОМОЋУ КОРИСНИЧКИХ ТВИТОВА

POSITIONING TWITTER USERS ON THE POLITICAL SPECTRUM BASED ON THE CONTENTS OF THEIR TWEETS

Милан Кнежевић, *Факултет техничких наука, Нови Сад*

Област – RAČUNARSTVO I AUTOMATIKA

Кратак Садржај – *У раду је представљен приступ за одређивање политичке оријентације корисника друштвене мреже Твитер базиран на његовим/њеним јавно доступним твитовима. Приступ је заснован на машинском учењу уз употребу Support Vector Machines класификаотра, BERT језичког модела и библиотеке Selenium.*

Кључне речи: *Твитер, политичка оријентација, SVM, BERT*

Abstract – *This paper presents an approach to determine the political orientation of Twitter users based on the contents of their public available tweets. The approach is based on machine learning with the use of the Support Vector Machines classifier, BERT language model, and the Selenium library.*

Keywords: *Twitter, political orientation, SVM, BERT*

1. УВОД

Узевши у обзир јачање друштвених мрежа у претходној и текућој деценији, оне су временом постале инструмент и поље политичке борбе и супротстављених ставова између различитих политичких субјеката.

Проблем је дефинисан на следећи начин: Одредити положај *Twitter* корисника на мапи политичког спектра помоћу његових твитова. Наравно једна од важнијих напомена пре самог почетка коју треба споменути јесте да се радило искључиво о политичким дешавањима и актуелностима које се тичу Републике Србије и збивањима у Републици Србији.

Као и код свих друштвених проблема, не постоји јединствено решење, већ постоји више прегршт различитих приступа проблему. С тим у вези то ову тему чини доста занимљивијом и отворенијом за шире разматрање, као и за полемисање о самој суштини проблема, не само за инжењере већ и разне друге професије које се баве друштвеним дисциплинама.

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био dr Александар Ковачевић, ред. проф.

2. СРОДНА ИСТРАЖИВАЊЕ

Као донекле сличан подухват на који сам се угледао приликом израде јесте рад истраживача: Јупенг Гуа, Тинг Чена, Јижу Суна и Бинђу Ванга под насловом *Детекција идеологије Twitter корисника помоћу анализе повезаности* (*Ideology Detection for Twitter Users via Link Analysis*) [1]. Наиме горе поменути истраживачи су користећи праћења, помињања и ретвитовања линкова (follow, mention, retweet links), покушали да прилагоде једнодимензионалну односно пак бинарну класификацију корисника да ли је присталица или либералне идеологије или конзервативне идеологије.

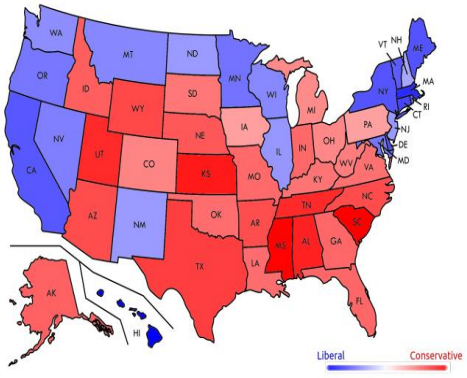
Чињеница је да је експеримент рађен за друштво и политичке прилике Сједињених Америчких Држава где постоје искључиво два јака политичка субјекта а то су Демократска партија који представљају припаднике либералне оријентације, док Републиканска партија представља припаднике конзервативне струје. Самим тим њихов проблем јесте доста поједностављен из више разлога.

Један од тих јесте да у моменту писања рада, САД броје преко 300 милиона становника, самим тим је и обим информација доступних на *Twitter* платформи много већи него ли на нашем поднебљу.

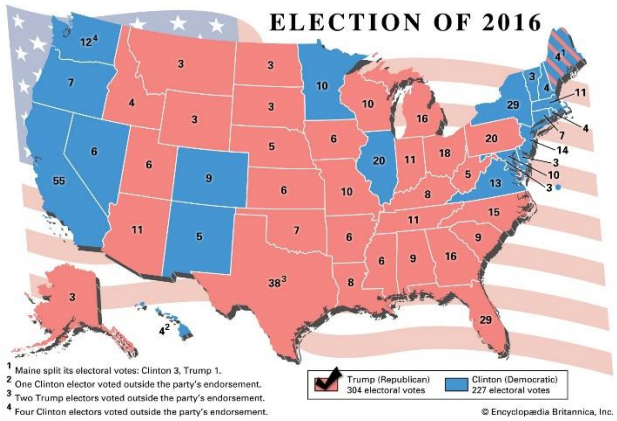
Друга много важнија ствар јесте што је у њиховом случају проблем бинарне природе, док је у мом случају дводимензионалне природе са више дискретних вредности.

Треће анализирају се само праћења, помињања и ритвитови, од којих је ритвит можда и најрелевантнији податак од свих из разлога што га корисници највише користе када желе да поделе твит с којим се у већој мери или у потпуности слажу, док праћења не значе нужно слагање са истим погледима корисника којег пратите, већ можете да га пратите из перспективе информисања и помињање може да се односи у негативном контексту ако некога критикујете приликом помињања.

Уз све наведене ставке претходног рада, приступи овог мастер рада су за бар нијансу комплекснији од претходно наведеног и помало релевантнији што се тиче логичке перцепције, али у толико и теже природе за добити прецизнији резултат.



Слика 1 - *Резултати наведеног рада*



Слика 2 – *Резултати председничких избора у САД из 2016. године*

3. АЛАТИ КОРИШЋЕНИ У РАДУ

На почетку овог поглавља најбоље је набројати их хронолошким редоследом коришћења.

Коришћени алати:

1. *Selenium WebDriver [2]*
2. *BERT*-ић [3]
3. *SrbAI [4]*
4. *Support Vector Machine – SVM [5]*
5. *Matplotlib [6]*

Целокупан рад је израђен у *Python* програмском језику који је иначе најчешће коришћен програмски језик за решавање проблема из области *Data Science*-а. Разлог овоме јесте што је овај наведени језик врло једноставан за коришћење и садржи широку подршку многих алата у које спадају и горе поменути.

Selenium WebDriver је алат који је коришћен у виду *python* библиотеке као *Scraper*. *Scraper* у преводу значи стругач што у потпуности одговара сврси у којој је примењен приликом реализације задатка. Наиме уз помоћ Selenium-а се 'стружу' подаци познавањем правила структуирања XML фајлова тј. у овом случају његовог подскупа односно HTML фајлова јер је у питању садржај интернет странице друштвене мреже *Twitter*.

(Bidirectional Encoder Representation for Transformer (BERT)) је *NLP* модел развијен од стране компаније Гугл 2018. године. NLP (*Natural Language Processing*) је грана вештачке интелигенције која се труди да омогући рачунарима да стекну вештину разумевања писаних и изговорених речи што ближе нивоу људског поимања коришћења истих у језицима. BERT-ова сврха је да одреди сентимент о теми о којој се говори у анализираној реченици. BERT-ић само једна од многих варијанти BERT-а који је намењен за Српски, Хрватски и Словеначки језик.

SrbAI је пројекат прикупљања алгоритама и модела за процесирање српског језика у јединствену *Python* библиотеку. Библиотека треба да садржи како основне методе за процесирање српског, попут стеммера, препознавање врста речи (*part-of-speech tagging*), негација, до напреднијих функционалности, попут препознавање именованих ентитета (*named entity tagging*), класификације, итд.

Support vector machines (SVMs) су метод надгледаног учења које служе за класификацију и регресију. Класификација је процес сврставања свих података у одређене класе података које имају својствене карактеристике. Регресија је сличан процес класификацији, односно она је покушај њене надградње. Разлика између класификације и регресије јесте у томе што као резултат даје меру колико нешто јесте нека класа и колико нешто није та класа у бинарном случају или у вишекласном случају колико је то нешто свака од могућих класа.

Matplotlib јесте широко распростањена *Python* библиотека која служи за креирање статичких, анимираних и интерактивних визуелизација у *Python*-у. Једноставна библиотека са много могућности за презентовање суштине добијених резултата као и додатно описивање истих. Приликом израде овог рада постојала је огромна потреба за визуелизацијом резултата јер су добијени резултати били дводимензионални, стога се много лакше закључују информације визуелним путем.

4. ИМПЛЕМЕНТАЦИЈА РЕШЕЊА

4.1. Фаза прикупљања и анотације података

С обзиром на то да за проблем који се решава у овом раду не постоје јавно доступни подаци, било је потребно прикупити податке са Twitter-а. Направљена је листа упита који је наш програм уз помоћ алата *Selenium*-а. Упити су садржали ознаке корисника (*handle*) које смо узели за тренинг скуп и речи везане за одређену тему (различити граматички облици те речи). Информације о томе којој теми те речи припадају смо сачували у мапи. Теме које смо користили су:

1) Албанија
2) Америка
3) Београд
4) Црква
5) Европа
6) Корупција
7) Косово
8) LGBT
9) NATO
10) Полиција
11) Путин

12) Русија
13) Украјина
14) Војска

Када се упит изврши, програм би преузимао информације о приказаним твитовима као што су:

- Корисничко име
- Датум објаве твита
- Језик/Писмо на ком је твит написан
- Садржај твита

Пошто је познато у тренинг скупу који корисник се налази где на политичкој мапи, потребно је додати још две колоне а то су позиције на мапи политичког спектра односно вредност у распону [-1,1]. Једна оса представља економску политику корисника где -1 предстаља став близак екстремним левичарима док вредност 1 представља став близак екстремним деничарима. Друга оса представља критеријум социјалне политике где -1 вредност представља да ли је неко близак либералном становишту, док 1 представља становиште ауторитарне природе.

Остаје још само да се одреди сентимент преузетих твитова. За овај корак смо користили претренирани BERT модел (*EMBEDDIA/bertic-tweetsentiment*) за твитове на Српском језику, док за твитове на Енглеском смо користили стандардни BERT. *SrbAI* алат је искоришћен као конвертер из азбуке у абецеду односно из ћирилице у латиницу. И као крајњи резултат ове фазе рада добијамо следеће редове у табели са следећим колонама:

- Корисничко име
- Ознаку корисника
- Датум објаве твита
- Језик/писмо на ком је твит написан
- Садржај твита
- Тему
- Сентимент
- Оцену валидности сентимента
- Позицију на економској скали у распону од -1 до 1 са кораком 0.1
- Позицију на социјалној скали у распону од -1 до 1 са кораком 0.1

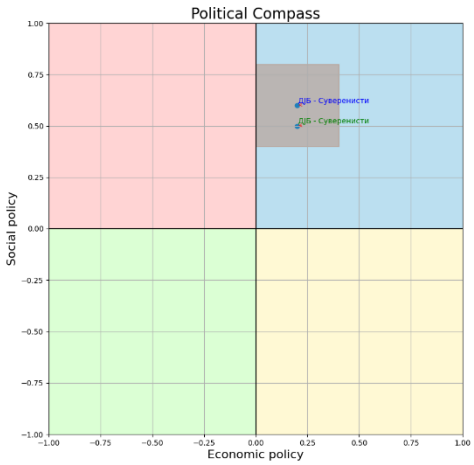
4.2. Фаза позиционирања корисника на мапи политичког спектра

Поступак је био обучавање 2 *Support Vector Regressor*-а (један за економски аспект и други за социјални аспект).

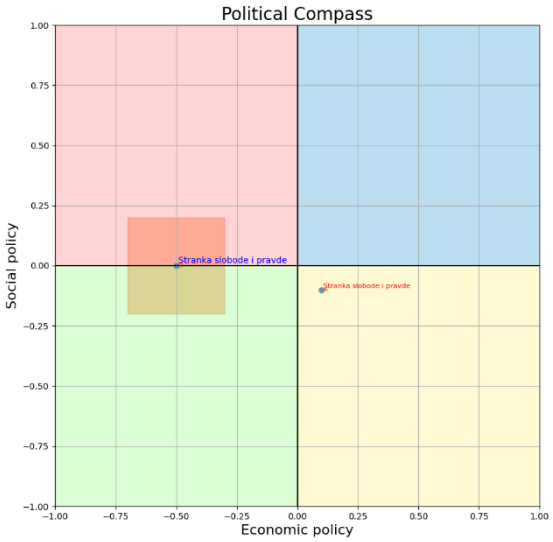
Наиме пре самог тренирања података коришћен је један од првих параметара, а то је параметар за горње ограничење броја твитова једног корисника по једној теми (MAX_TWEETS_PER_TOPIC). Разлог постојања овог параметра лежи у томе да један корисник може да твитује много о једној теми и да практично утиче на предикцију регресора у смислу да ће имати превише истих виђења једног корисника на дату тему. Други параметар који је стандардан приликом сваког обучавања јесте однос тренинг и тест скупа (TRAIN_SET_PERCENTAGE). У пракси често је случај да када се повећа овај проценат исход буде бољи, међутим у овом приступу је супротно понашање, односно мора се наћи одређена мера. Разлог овоме јесте што се у овом приступу коначан положај на мапи политичког спектра рачуна као медијана свих положаја *tweet*-ова једног корисника који су прошли кроз предикцију. Стога је и сасвим логично да што више tweet-ова буде у тест скупу, то ће медијана бити прецизнија. С друге стране не сме се ни радити много на уштрб тренинг скупа јер и то с друге стране утиче на перформасе из очитих разлога. На крају постоји и трећи параметар који одређује околину односно дозвољено окружење у којем би координате требало да се налазе:

(RADIUS_AROUND_REAL_POSITION).

Овај фактор је уведен из разлога што је врло мала вероватноћа да ће погодити корисника у обе координате. Он највише утиче на прецизност и може се рећи да практично одређује проценат тачности овог приступа.



Слика 3 – *Пример поготка*



Слика 4 - *Пример промашаја*

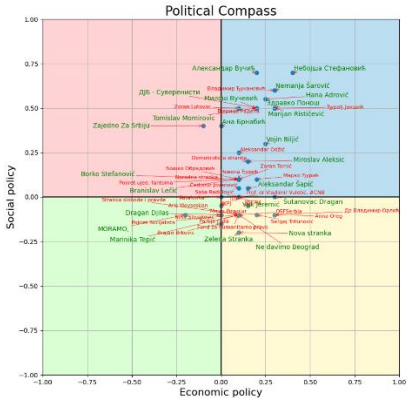
5. РЕЗУЛТАТИ И ДИСКУСИЈА

Резултати претходно описаног приступа су добијени на основу следећих вредности параметара:

1) MAX_TWEETS_PER_TOPIC = 30
2) TRAIN_SET_PERCENTAGE = 0.7
3) RADIUS_AROUND_REAL_POSITION = 0.4

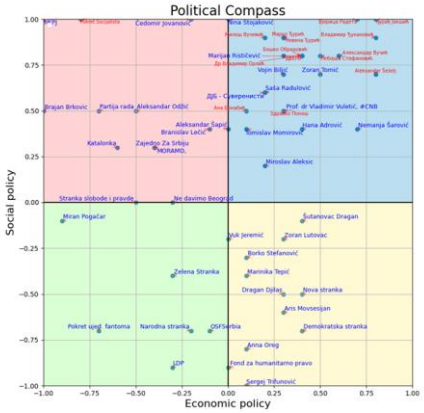
сумирани су у кратким цртама у следећим бројевима:

- Укупан број тестираних корисника: 55
- Укупан број погодака: 24
- Укупан проценат поготка: 43,67%



Слика 5 - *Добијене позиције корисника*

СЛИКА 5 на апстрактном нивоу представља резултате. Наиме на њој су приказани корисници са добијеним координатама и у зависности од тога да ли је и добијена позиција упала у окружење око реалне позиције зависи боја корисника. Зелени су сматрани погодком, док су црвени сматрани грешком. Да бисмо видели боље резултате, најбоље да упоредимо позиције са СЛИКОМ 6 која представља реалне позиције корисника. Највише предикција јесте концентрисано око координатног почетка што и има смисла ако се узме у обзир да медијана увек вуче све кориснике ка средини. Оно што јесте мана скупа података јесте то што и није нађено доста екстрема свих врста како би имали више дистрибуиране позиције.



Слика 6 - *Реалне позиције корисника*

Занимљиво јесте то што се по проценама очекивало да ће се нагињати више ка доњој половини спектра јер су корисници тог дела доста гласноговорнији на овој платформи у односу на кориснике горњег спектра, али вероватно има утицаја и теме које су биране у разматрање приликом скупљања података за тренирање.

6. ЗАКЉУЧАК

У овом раду представљен је један од метода за позиционирање корисника на мапи политичког спектра помоћу ставова и мишљења који корисници износе преко својих профила на друштвеним мрежама. Што се резултата тиче не треба да нас чуди овакав исход из више разлога: мали број података постоји који подлеже овом проблему, алат за сентимент су радили на српском језику који сигурно није усавршен, итд. Али, из истог разлога треба да постоји разлог за оптимизам јер се сигурно резултат од 42% може подићи. Оваква истраживања јавног мњења су корисна, а тек ће и бити за опипавање реакције јавности неке земље на неку од рискантнијих одлука власти те земље у којој би се потенцијално истраживање радило.

7. ЛИТЕРАТУРА

[1] *Ideology Detection for Twitter Users via Link Analysis: Yupeng Gu, Ting Chen, Yizhou Sun and Bingyu Wang (*http://web.cs.ucla.edu/~yzsun/papers/2017_SBP_Ideology*)*

[2] *Selenium WebDriver (*https://www.selenium.dev/documentation/webdriver*)*

[3] BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian, Nikola Ljubešić, Davor Lauc (https://aclanthology.org/2021.bsnlp-1.5.pdf)

[4] SrbAI - Python biblioteka za procesiranje srpskog jezika (https://github.com/Serbian-AI-Society/SrbAI)

[5] *Support Vector Machines – Scikit (*https://scikit-learn.org/stable/modules/svm.html*)*

[6] *MatPlotLib (*https://matplotlib.org*)*

Кратка биографија:

Милан Кнежевић рођен је у Новом Саду 1998. године. Дипломирао је на смеру Рачунарство и аутоматика на Факултету техничких наука у Новом Саду 2021.
контакт: mile.knezevic98@gmail.com