УДК 616.98:578.834]:004.89:577

Евгенија КРАЈЧЕВСКА*
Александра ДЕДИНЕЦ**
Љупчо КОЦАРЕВ***

СЕНТИМЕНТАЛНА АНАЛИЗА НА ТВИТЕР-ПОДАТОЦИ ЗА ВРЕМЕ НА ПАНДЕМИЈАТА НА КОВИД-19 ВО РЕПУБЛИКА МАКЕДОНИЈА ВО СПОРЕДБА СО СВЕТОТ

Апстракт

Ниеден настан во последните неколку децении не ги потресе луѓето на глобално ниво како пандемијата од ковид-19 која зема замав од почетокот на 2020 година. Првпат во историјата на човештвото голем број држави беа ставени под карантин како резултат од зголемениот број заболени лица и неможноста на здравствените системи да ја совладаат пандемијата со ваков размер.

Од друга страна, масивните ограничувања на економијата може да имаат исто толку здравствени последици како и самиот вирус – милиони луѓе може да ги загубат своите работни места и приходи, да станат депресивни, да се зголеми употребата на наркотични средства како цигари, алкохол, дрога, поради страв да се одложуваат посетите на лекарите и во болниците, особено за луѓето со хронични и сериозни медицински состојби кои не се поврзани со вирусот. [1] Во овој труд се истражуваат токму состојбата и мислењето на луѓето во овој период на пандемијата, анализирајќи ги податоците од Твитер-платформата, преку

* Факултет за информатички науки и компјутерско инженерство, Скопје, РС Македонија

** Факултет за информатички науки и компјутерско инженерство, Скопје, РС Македонија

*** Македонска академија на науките и уметностите

користењето на методите за рударење на мислење, или повеќе познато како сентиментална анализа. Дополнително, направена е споредба на мислењето кај локалното население, со глобалните мислења и ставови.

Клучни зборови: *сентиментална анализа, ковид-19, корона-вирус*

Вовед

Појавата на вирусот во Вухан, во Кина, во декември 2019 година, не претставуваше закана за светот сè додека не почна интензивно да се шири, така што беа засегнати голем број држави. Меѓу првите земји кои детектираа позитивни случаи беа Тајланд, Јапонија и Јужна Кореја, подоцна се јавуваат во САД, Австралија и Европа.

Во Македонија, првиот случај е откриен на 26 февруари 2020 година. Со зголемувањето на бројот на заразените, неопходни беа мерки за претпазливост, кои вклучуваа забрана за патување во заразените земји, воведување карантин и полициски часови, како и забрана за собири, јавни настани, спортски натпревари и слично. Сето ова ја зголеми свесноста на луѓето за опасноста на вирусот, стана причина за паника, загриженост и социјално дистанцирање, што воедно влијаеше и на психичкото здравје на луѓето.

Влијанието на пандемијата врз здравјето на луѓето е анализирано од различни аспекти во неколку научни студии [2] [3]. Дополнително, направени се и анализи за мислењето на луѓето во различни држави во светот со помош на информациите од социјалните мрежи [4] [5]. Во овој труд, сентиментална анализа на мислењето на луѓето, на глобално ниво и на ниво на Македонија, е направена со користење на податоците од Твитер, која може да даде значајни информации за психичката состојба на луѓето, како и за нивните ставови поврзани со настанатата ситуација. Притоа, за сентименталната анализа развиени се два модела, така што првиот е со употреба на лексикон, а во вториот пристап се користи проблемот на класификацијата на податоците. Во првото поглавје од овој труд е опишан начинот на собирање и процесуирање на податоците на глобално и на локално ниво, како и временскиот период на кој се однесуваат. Потоа, следува основна анализа и

споредба на добиените множества, ри што ги презентираме најчесто користените зборови и ознаки (*тагови*), како и истовремени појавувања на одделни термини. Во третото поглавје, преку два различни пристапи, се прави сентиментална анализа, при што се разгледува значењето на пораките (твитовите) и во каква конотација најчесто се користат. Конечно, во последното поглавје, ги претставуваме заклучоците од спроведената анализа, односно сличностите и разликите помеѓу мислењата за ситуацијата во Македонија наспроти погледите на јавноста од целиот свет.

Собирање и процесуирање на податоците

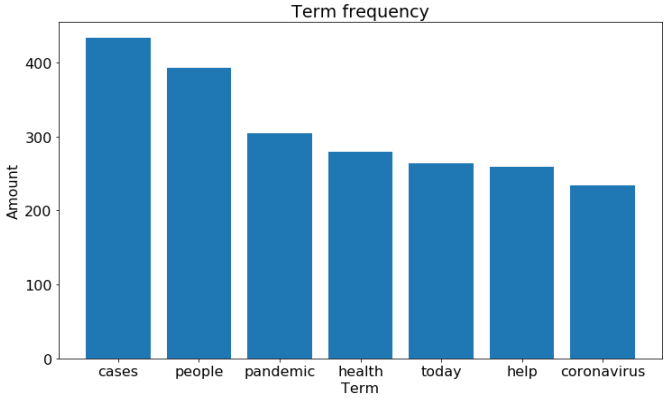
Новонастанатата ситуација, во однос на мислењето на народот, најдобро се следи на социјалните мрежи, па токму затоа изворот на податоци со кои ги анализираме нивните ставови и главните теми е преку социјалната платформа *Твитер*. За да можеме да пристапиме до податоците мора да регистрираме апликација преку постојниот профил. Постојат голем број библиотеки кои го олеснуваат пристапот, а една од нив е библиотеката *Питон* (*Python*) – *Твиѝи* (*Tweepy*). За генерирање на множеството може да се користат филтри како јазикот на кој се напишани, државата, односно локацијата на авторот на пораката (твитот), клучни зборови, ознаки (тагови) и слично. Пребарувањето не ги зема предвид големите и малите букви, па така ја користиме ознаката (тагот) „covid19“ за глобалното множество, а терминот „корона“ и филтер за македонски јазик за локалното множество имајќи предвид дека на нашите простори не е актуализирана употребата на ознаки (тагови) кај пораките (твитовите). Глобалното множество е составено од 8 000 пораки (твитови), додека локалното множество од 722. Разликата во големината се должи на помалата популарност на оваа социјална мрежа во Македонија, во споредба со *Фејсбук* и *Инстаграм*, кои интензивно се користат. Со цел да се опфатат поголем број податоци за различни случувања и да се задржи диверзитетот, податоците се собирани во рамки од еден месец, почнувајќи од 27-ми март, бидејќи не е дозволен пристап до пораките (твитовите) постари од една недела.

Процесуирањето на ваквите податоци се однесува на нивното прочистување кое се одвива во неколку чекори. Се отстрануваат интерпункциските знаци, емотикони, предлози и сврзници кои не носат значење, специјални знаци, врски (линкови), други корисници кои, можеби, се спомнуваат во рамките на пораката (твитот) и, конечно, се прави *токенизација* и трансформација на текстот во мали букви. Презентацијата на множеството на ваков начин ни дозволува да бидат спроведени следните две анализи.

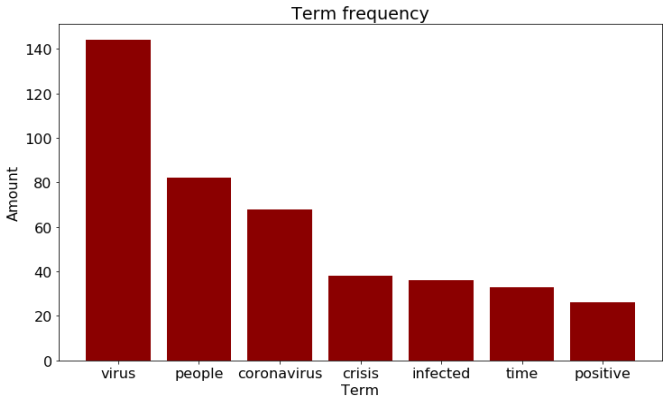
Основна анализа на добиените множества

Еден од основните начини за да добиеме генерален преглед на множеството е да ги испитаме најчесто користените термини и ознаки (тагови), односно нивните фреквенции на појавување. Користиме едноставен речник во кој ги броиме појавувањата на секој од термините во различни пораки (твитови). За глобалното множество го добиваме следниот приказ, прикажан на слика 1. Додека, за локалното множество, го добиваме приказот на најчесто користени термини на слика 2. Можеме да забележиме дека заедничките термини се поврзани со известувања за бројот на заразените, како и за кризата околу пандемијата. Преку најчестите термини од глобалното множество, може да се заклучи дека главна тема е состојбата на луѓето, новостите околу развојот на вирусот и како се снаоѓаат здравствените системи, воедно и бројот на смртните случаи. Кај локалното множество, пак, се истакнуваат термини кои посочуваат интерес за позитивните случаи, како и кризата што владее во рамките на државата.

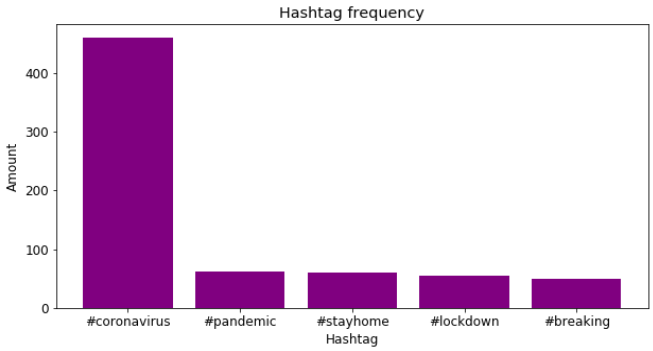
Во однос на ознаките (таговите), на глобално ниво, наизменично се користат ознаката (тагот) „covid19“, според кој е генерирано множеството, и „coronavirus“, како што можеме да забележиме на слика 3, додека на локално ниво, мал е бројот на употребените ознаки (тагови), за да може да се донесе конкретен заклучок.



Слика 1: Најчесто користени термини во пораки (твитови), глобално



Слика 2: Најчесто користени термини во македонски пораки (твитови)



Слика 3: Најчесто користени ознаки (тагови) во пораки (твитови), глобално

Појавувањето на одделни термини не ни дава детално објаснување за што, всушност, се однесува пораката (твитот). Токму затоа, го разгледуваме заедничкото појавување на два термина, за подобро да го разбереме контекстот. Градиме горнотриаголна матрица, каде што полето [X][Y] го означува бројот на пораките (твитовите), каде што терминот X се среќава заедно со терминот Y. Најчесто користени термини, заедно, на глобално ниво, се: [((‘cases’, ‘total’), 204), ((‘cases’, ‘deaths’), 170), ((‘cases’, ‘confirmed’), 121), ((‘deaths’, ‘total’), 108), ((‘home’, ‘stay’), 79)], додека за локалното множество добиваме: [((‘corona’, ‘virus’), 122), ((‘corona’, ‘people’), 48), ((‘corona’, ‘crisis’), 29), ((‘corona’, ‘time’), 21), ((‘affairs’, ‘minister’), 20)]. Оттаму, ја согледуваме корелацијата на издвоените најчесто користени термини и добиваме појасна слика за значењето.

Сентиментална анализа

Овој тип анализа има широка примена во анализирањето на текстовите, каде што се истакнуваат личното мислење и намерата на авторот. Постојат два главни пристапи на проблемот за автоматско извлекување на мислењето. Првиот пристап е со употреба на лексикон, при што за секој документ, односно во нашиот случај порака (твит), се пресметува семантичката ориентација во однос на зборовите или фразите кои се користат. Другиот пристап се однесува на класификација, односно тренирање на класификатори од одредено множество на лабелирани инстанции, на текст или на реченици, со помош на методи од машинско учење. Во прилог се подготвени анализи според двата пристапа.

А. *Семантичка ориентација*

Семантичката ориентација на даден збор се дефинира како разлика помеѓу неговите асоцијации со позитивни и со негативни зборови. Сакаме да пресметаме колку зборот е поврзан со добри или со лоши термини. Ваквата поврзаност можеме да ја пресметаме со *Pointwise Mutual Information* (PMI) која, за терминте t1 и t2, се пресметува како

$$PMI(t1, t2) = log\left(\frac{P(t1 \wedge t2)}{P(t1)P(t2)}\right) \qquad (1)$$

Во одредени трудови [6], семантичката ориентација на зборот се пресметува во однос на неколку термини, но ова множество може да се прошири со користењето на лексикони. Пример за еден таков лексикон е лексиконот на Минкинг Ху (Minqing Hu) и Бинг Лиу (Bing Liu) [7], каде што, ако за позитивните термини користиме V+, а за негативните V–, тогаш семантичката ориентација на терминот t ќе се пресмета на следниов начин

$$SO(t) = \sum_{t \epsilon V+} PMI(t, t') - \sum_{t \epsilon V-} PMI(t, t') \quad (2)$$

За дадено множество од документи (твитови) D, дефинираме DF (Document Frequency) за одреден термин, како честота на појавување на терминот во множеството од документи. Истата дефиниција се однесува и на заедничкото појавување на два термина t1 и t2. Според ова, веројатностите $P(t)$ и $P(t1 \wedge t2)$ може да се пресметаат како:

$$P(t) = \frac{DF(t)}{|D|} \quad (3)$$

$$P(t1 \wedge t2) = \frac{DF(t1 \wedge t2)}{|D|} \quad (4)$$

Семантичката ориентација на еден термин ќе има позитивна, т. е. негативна вредност, ако терминот често се јавува заедно со позитивни, т. е. со негативни термини од лексиконот. Вредноста ќе биде нула ако терминот подеднакво се сретнува заедно со позитивни и со негативни термини или, пак, не се сретнува воопшто со термини од лексиконите.

Во глобалното множество за термини, кои се користат во позитивен контекст, претставени заедно со информацијата за семантичката ориентација, добиваме: [(‘donations’, 38,48), (‘almighty’, 33,52), (‘hospital’, 33,51), (‘entitled’, 32,97), (‘hope’, 32,95), (‘ambassador’, 31,9), (‘appreciate’, 31,52), (‘healthy’, 30,8)], додека на локално ниво такви се термините: [(‘conversation’, 26,07), (‘bring’, 24,49, (‘holding’, 24,16), (‘medical’, 24,07), (‘aware’, 22,58), (‘awaits’, 22,58), (‘like’, 20,85), (‘better’, 20,15)]. Во глобалното множество за

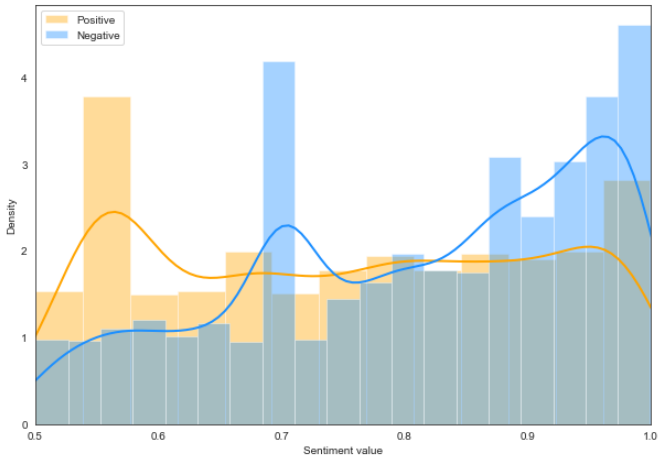
термини, кои најчесто се асоцираат со други негативни термини, добиваме: [(‘failing’, –44,07), (‘every’, –45,98), (‘dream’, –47,7), (‘evil’, –52,48), (‘attacks’, –57,21), (‘americans’, 58,64), (‘already’, –59,26), (’alone’, –68,74)], додека за локалното множество добиваме: [(‘measures’, –33,78), (‘biggest’, –33,99), (‘corona’, –35,03), (‘burning’, –35,08), (‘fraud’, –36,22), (‘call’, –36,65), (‘blame’, –44,56), (‘believe’, –46,23)]. Исто така, можеме да ја провериме семантичката ориентација на најчесто користените термини. За глобалното множество, тоа се: (cases, –29,46), (covid, –34,038), (people, –23,77) (pandemic, –19,92), (health, –30,97), додека на локално ниво: (virus, 1,01), (people, –0,79), (coronavirus, –23,94), (infected, –3,66), (crisis, –12,01).

Најчесто користените термини на глобално ниво имаат негативна семантичка ориентација, па ова ни е показател дека често биле користени во негативен контекст, додека за локалното множество првите два термина се неутрални, но останатите се, исто така, со негативна конотација.

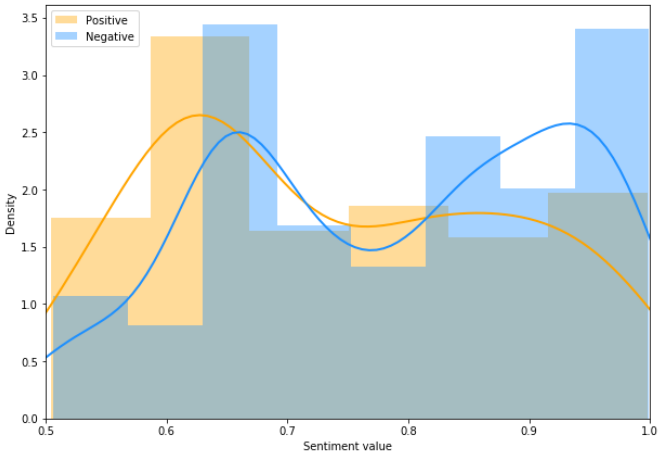
Б. *Класификација*

Кај методи на машинско учење, кога станува збор за анализа на текст, често се користи класификација. За разлика од претходниот пристап, овде мора да имаме почетно множество со означени податоци за тренирање на одреден класификатор. Бидејќи работиме со пораки (твитови), можеме да го употребиме множеството „twitter-samples“ од nltk-библиотеката. Податоците ги користиме во сооднос 70 % за тренирање и 30 % за тест-множеството, и на овој начин тренираме Бајесов класификатор, кој е еден од најчесто употребуваните класификатори за сентиментална анализа и, генерално, дава добри резултати [8].

Секоја порака (твит) ѝ припаѓа на една од двете класи кои се однесуваат на сентимент – позитивен или негативен, со одредена веројатност. За глобалното множество ја добиваме следната класификација: негативни 5 466 пораки (твита) и позитивни – 2534 пораки (твита). На слика 4 е прикажана густината на распределбата на вредностите во однос на сентиментот.



Слика 4: Густина на распределбата на сентиментот кај глобалното множество



Слика 5: Густина на распределбата на сентиментот кај локалното множество

Важно е да се напомене дека постојат податоци кои се доделени во позитивната или негативната класа, но со веројатност во рангот од [0,5;0,6], па наместо да ги доделиме во едната или во другата класа, можеме да ги сметаме за неутрални. На глобално ниво, поголем број од примероците ѝ припаѓаат на негативната класа, додека позитивните примероци се сконцетрирани во горе-

споменатиот ранг, па дискутабилно е дали да ги сместиме во оваа класа.

На слика 5 прикажана е густината на распределба на класите од локалното множество и се забележува поголемата сличност меѓу распределбите, но ова може да се должи на помалиот број податоци во ова множество. Сепак, класификацијата на локалното множество е во истиот сооднос како и во глобалното множество, односно негативните примероци се застапени двапати повеќе од позитивните: негативни – 500 пораки (твита) и позитивни – 222 пораки (твита). Повторно, голем број од пораките (твитовите) кои се класификувани како позитивни, со помала веројатност ѝ припаѓаат на класата, немаат значително влијание и може да се сметаат за неутрални. За негативните пораки (твитови), пак, сме сигурни дека играат значајна улога во анализата поради тоа што голем број се предвидени со повисока веројатност.

Заклучок

Целта на ова истражување е да добиеме преглед за комуникацијата на социјалните мрежи во Македонија, но и во светот, од психолошки аспект, за време на пандемијата. Преку спроведените анализи јасно може да се види дека најчесто употребуваните термини поврзани со вирусот се често користени во негативен контекст, без разлика дали се истакнува разочараноста од лошиот начин на совладување на пандемијата од страна на властите, недоличното однесување на луѓето или проблемите со здравствениот систем. И покрај тоа што секоја земја на свој начин се бори со пандемијата, резултатите покажуваат незадоволство, страв и вознемиреност, не само во Македонија туку низ целиот свет.

Денес, луѓето се послободни бидејќи карантинот е укинат во голем број држави, но континуираниот раст на новозаболени лица покажува дека пандемијата не е целосно под контрола и има тенденција повторно да земе замав. За совладување на пандемијата работат голем број научни центри насочени кон пронаѓањето вакцини [9].

Во услови кога вакцините ќе бидат достапни низ целиот свет, може да се спроведат слични анализи за мислењето на популацијата за вакцинацијата за ковид-19 имајќи предвид дека во последните неколку години постои движење против задолжителната вакцинација на деца.

Литература

[1] Autumn, Kujawa, Green, Haley, Compas, Bruce, Dickey, Lindsay and Pegg, Samantha. (2020). "Exposure to COVID-19 Pandemic Stress: Associations with Depression and Anxiety in Emerging Adults in the U.S." PsyArXiv. June 29.

[2] Brooks, Samantha K. (2020). *The psychological impact of quarantine and how to reduce it: rapid review of the evidence*, PhD, The Lancet, Volume 395, Issue 10227, 14 – 20 March 2020, pages 912–920.

[3] Drias, Habiba H., Drias, Yassine. (2020). *Mining Twitter Data on COVID-19 for Sentiment analysis and frequent patterns Discovery*, May 2020.

[4] Duan, Li, Zhu, Gang. (2020). *Psychological interventions for people affected by the COVID-19 epidemic*, The Lancet, Psychiatry, Volume 7, Issue 4, April 2020, pages 300–302.

[5] Joyce, Brandon and Deng, Jing. *Sentiment Analysis Using Naive Bayes Approach with Weighted Reviews*, UNC GreensboroGreensboro, NC 27412, USA.

[6] Kamaran, Hussein Manguri, Rebaz, Najeeb Ramadhan, Pshko, Rasul Mohammed Amin. (2020). *Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks*, May 2020.

[7] Minqing, Hu and Bing, Liu. (2004). *Mining and Summarizing Customer Reviews*. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2004), Aug. 22 – 25, 2004, Seattle, Washington, USA.

[8] Turney, Peter D. *Thumbs Up or Thumbs Down?* Semantic Orientation Applied to Unsupervised Classification of Reviews, National Research Council of Canada.

[9] Lurie, Nicole M.D., M.S.P.H., Saville, Melanie, M.D., Hatchett, Richard, M.D., and Halton, Jane, A.O., P.S.M. *Developing Covid-19 Vaccines at Pandemic Speed*.