Универзитет „Св. Кирил и Методиј“ - Скопје

Факултет за информатички науки и компјутерско инженерство

Истражување на јавната свест за аерозагадувањето во Западен Балкан преку анализа на Твитер дискусии и електронски вести со користење на NLP техники и анализа на временски серии

Дипломска работа

Ментор:

доц. д-р Александра Дединец

Изработил:

Ангела Маџар, 181010

Скопје, 2022

Апстракт

Аерозагадувањето претставува сериозна закана по здравјето на жителите на земјите од Западен Балкан, со биомасата како главен загадувач на воздухот во регионот. Оваа дипломска работа се темели врз претпоставката дека активноста на социјалните мрежи, како што е Твитер, се зголемува правопропорционално со ескалацијата на загадувањето на воздухот во текот на зимскиот период. Целта на овој труд е да се истражи јавната свест за загадувањето на воздухот во Македонија, Србија, Босна и Херцеговина и Црна Гора во период од ноември, 2021 до март, 2022, земајќи го предвид потенцијалното влијание на електронските вести врз ставовите и чувствата на поширокаta јавност. Применети се техники за обработка на природни јазици, како што се Sentiment Analysis и Topic Modeling, како и статистички анализи на временски серии со цел да се утврди дали сентиментите изразени во Твитер дискусиите на тема „аерозагадување“ се во согласност со нивоата на ПМ10 честичките измерени од официјалните мерни станици за квалитет на воздух во овие земји. Вакви анализи се извршени и врз написи на електронски вести, во обид да се утврди дали интернет порталите прикажуваат реална слика за состојбата со аерозагадувањето и дали промовираат еколошко однесување. Согласно резултатите, ваквите анализи можат да послужат како мерка за јавната свест на тема аерозагадување. Анализата на содржината на Твитер дискусиите може да открие евентуални проблеми во јавното мислење и да оствари свој придонес при изнаоѓањето начини за нивно решавање.

Клучни зборови: *аерозагадување, ПМ10, Западен Балкан, Twitter, анализа на твитови, сентимент-анализа, topic modeling, крос-корелација*

Содржина

1. Вовед

Аерозагадувањето е сериозен еколошки проблем на кој се должи годишната смртност на 3.7 милиони луѓе на глобално ниво. *Европската агенција за животна средина* проценува дека 90% од европските урбани жители се изложени на штетни загадувачи на воздухот, што укажува на потребата од поттикнување иницијативи за воспоставување контрола врз загадувањето на воздухот. Подеднакво се засегнати и сите земји од регионот на Западен Балкан. Се проценува дека луѓето кои живеат во овие предели ќе имаат живот пократок за 1.3 години како последица на истакнатите нивоа на загадувачи во воздухот. Според *Извештајот за квалитет на воздух (2018)*, суровата стапка на смртност која се припишува на загадениот воздух во Македонија изнесувала 154,7 на 100 000 жители. Во соседните земји, како Србија, изнесувала 200,7 на 100 000; во Црна гора била 110,9 на 100 000, додека пак во Босна и Херцеговина 105,1 на 100 000 жители.

Еден од загадувачите чиишто нивоа најчесто ги надминуваат законските граници во регионот се ПМ10 честичките кои главно се емитуваат при човечки активности како што се: греењето во домаќинствата, транспортот и индустријата. Неконтролираната урбанизација, дивоградбите, лошото планирање и проектирање на градби имаат свој удел во загадувањето на животната средина и на воздухот. Исто така, прекугранично загадување во рамки на Западниот Балкан, како и неговата околина, придонесува кон високите концентрации на овие честички. Сепак, според *Алијансата за здравје и животна средина* (Health and Environment Alliance – HEAL), двата главни извори на аерозагадување во Западниот Балкан се користењето на јаглен како доминантен електроенергетски ресурс при производството на електрична енергија и користење на огрев во домаќинствата за време на грејната сезона. Како резултат на ова, максималната дневна граница на ПМ10 која изнесува 40 μg/m3, а која е утврдена со закон, е надмината помеѓу 120 и 180 дена во годината, најчесто во зимскиот период. Континуирана изложеност на овие честички може да предизвика мноштво здравствени проблеми, често поврзани со срцето и белите дробови.

Целта на овој дипломски труд е да ја истражи јавната свест за аерозагадувањето преку користење квалитативни и квантитативни анализи за споредување на активноста на социјалните мрежи и електронски вести на тема аерозагадување со податоците за ПМ10 честички измерени од официјалните мерни станици за квалитет на воздух. Истражувањето се темели врз претпоставката дека ескалацијата на аерозагадувањето во Западниот Балкан во текот на зимскиот период ќе предизвика интензивирана активност на Твитер, социјална мрежа која луѓето ја користат за јавно да ги споделат своите ставови и чувства на одредена тема. Притоа, се зема предвид и потенцијалното влијание кое содржините презентирани од страна на електронските вести на интернет порталите би

можеле да го имаат врз емоциите и расудувањето на пошироката јавност. За спроведување на ваквите анализи, се користат податочни множества кои се резултат на неделна колекција на твитови и електронски вести во период од ноември, 2021 до март, 2022 година. За да се изведе заклучок за генералната перцепција за загадувањето на воздухот во неколку земји од Западниот Балкан, меѓу кои: Македонија, Србија, Босна и Херцеговина и Црна Гора, во овој труд се користат техники за обработка на природни јазици, како што се Sentiment Analysis и Topic Modeling, како и статистички анализи, односно Cross-correlation на временски серии.

Неодамна, употребата на вакви техники за анализа на куси содржини споделени на социјалните мрежи за микро-блогирање (како што е Твитер), со цел извлекување генерален заклучок за јавното мислење на теми како што е аерозагадувањето, стана популарно поле на истражување за научниците ширум светот. Во литературата, првичните анализи од овој тип се направени користејќи содржини од Weibo (кинеска социјална мрежа која е пандан на Твитер), со употреба на мануелен метод на квалитативна класификација за одвојување на содржините во кои се изразени позитивни и негативни сентименти. Фреквенцијата на вака филтрираните микро-блогови се користела за подобрување на корелацијата со дневниот индекс за квалитет на воздух (AQI), што се покажало како ефикасен метод за делумно следење на аерозагадувањето. Дополнително, се покажало дека анализата на содржината на микро-блоговите поделени по сентименти може да открие значајни сознанија за генералната перцепција на јавноста на оваа тема. Други истражувања го демонстрираат потенцијалот на социјални медиуми за следење на нивоата на ПМ2.5 честички како алтернативен медот во области без системи за следење на аерозагадувањето, така што користат техники за Sentiment Analysis за поделба микро-блоговите по сентименти и ја споредуваат нивната фреквенција со официјални податоци за ПМ2.5 честички користејќи статистички анализи. Покрај тоа, одредени студии сугерираат дека реакцијата на јавноста при одреден екстремен настан, како што е голем шумски пожар кој влијае врз стапката на загадување на воздухот, може да се долови со користење на техники за Topic modeling, за да се извлечат теми од твитовите објавени за време на тој настан. Се на сè, експериментирањето со алгоритми од областа на надгледувано и ненадгледувано учење, може да биде извор на информации за временската еволуција на актуелните теми дискутирани во микро-блогови (како што се твитовите). Исто така, може да открие и евентуални проблеми во јавното мислење и да оствари свој придонес при изнаоѓањето начини за нивно решавање.

Резултатите кои произлегуваат од овој дипломски труд имаат значаен придонес кон постоечката литература, потврдувајќи дека класификацијата на позитивни и негативни сентименти во твитовите и електронските вести, нивната временска дистрибуција и корелација, како и анализата на содржани теми поврзани со аерозагадувањето во

Западниот Балкан, откриваат интересни и значајни сознанија за јавната свест на оваа тема.

Трудот е организиран на следниот начин: по воведот, следи објаснување за податоците користени при истражувањето. Во следното поглавје, претставено е теоретско објаснување на методологиите кои се користат, почнувајќи од Sentiment Analysis, Cross-correlation кај временски серии, па се до различните алгоритми за Topic Modeling. Потоа, претставени се добиените резултати, а на крај следи дискусија за нивно појаснување, како и заклучок во кој се истакнати главните сознанија произлезени од истражувањето и насоки за негово идно подобрување.

2. Податоци

Како студија на случај, се користеа податоци собирани од три различни извори, и тоа: твитови, кратка содржина на електронски вести и официјални податоци за нивоата на ПМ10 честички. Податоците се детално образложени во продолжение на ова поглавје.

2.1. Twitter податоци

За неделна колекција на твитови во чијашто содржина се дискутираат теми поврзани со аерозагадувањето, во период од 01.11.2021, до 28.02.2022, беше развиена едноставна апликација за пребарување користејќи ја библотеката *Tweepy* во програмскиот јазик *Python*. Се користеше стандардното *Twitter API* кое овозможува пребарување на твитови споделени во период од последните седум дена од моментот на пребарување. За да се опфатат македонски твитови на оваа тема, се пребаруваа клучните зборови: „*aerozagaduvanje*“, „*аерозагадување*“, „*zagaduvanje*“, „*загадување*“, „*ПМ10*“ и „*дишеме*“. На сличен начин, за да се опфатат твитови напишани на некој од останатите јазици од земјите од регионот на Западен Балкан (изоставувајќи го албанскиот јазик), се пребаруваа зборовите: „*zagadjenje*“, „*загађење ваздуха*“ и „*zagadjenje vazduha*“.

Иако зборот „загадување“ е широк поим, беше донесен емпириски заклучок дека Twitter корисниците го користат зборот „загадување“ како синоним на зборот „аерозагадување“. Неколкуте твитови во кои што се споменуваше различен тип на загадување беа мануелно отстранети.

2.2. Податоци од електронски вести

Паралелно со колекцијата на твитови, на неделно ниво се вршеше и колекција на кратки содржини на електронски вести (*teasers*) во кои се содржани горе-споменатите клучни зборови, со помош на моќната алатка за web-crawling – *Octoparse*. Во англискиот јазик, *teaser* е краток и илустративен текст кој се наоѓа веднаш под насловот на електронските вести, кој има за цел да го привлече вниманието на читателот (означен со црвена боја на *Слика 1*.). Вакви податоци се собираа од два интернет портали, и тоа *Time.mk* и *Time.rs*, кои се всушност агрегатори на вести кои дневно анализираат 15.000 статии, собрани од 120 различни извори. За агрегирање на вестите, позадински користат алгоритми за кластерирање.

За поедноставно изразување, во остатокот од овој дипломски труд, зборот „вести“ ќе се користи како синоним за „кратка содржина на електронски вести“, односно „teasers“.

Јагленот и мазутот се големи загадувачи, но ќе се проверува квалитетот на увезеното – уверуваат од Министерството за ...

Телма **- пред: 11 часа**

Ќе се проверува квалитетот на увезените енергенси, уверуваат од Министерството за животна средина. На прагот на грејната сезона и тешката зима којашто претстои, надлежните кројат планови како да не се загадува воздухот дополнително, но и што полесно ...

Слика 1. Појаснување на поимот „teaser“, односно „кратка содржина на електронски вести“

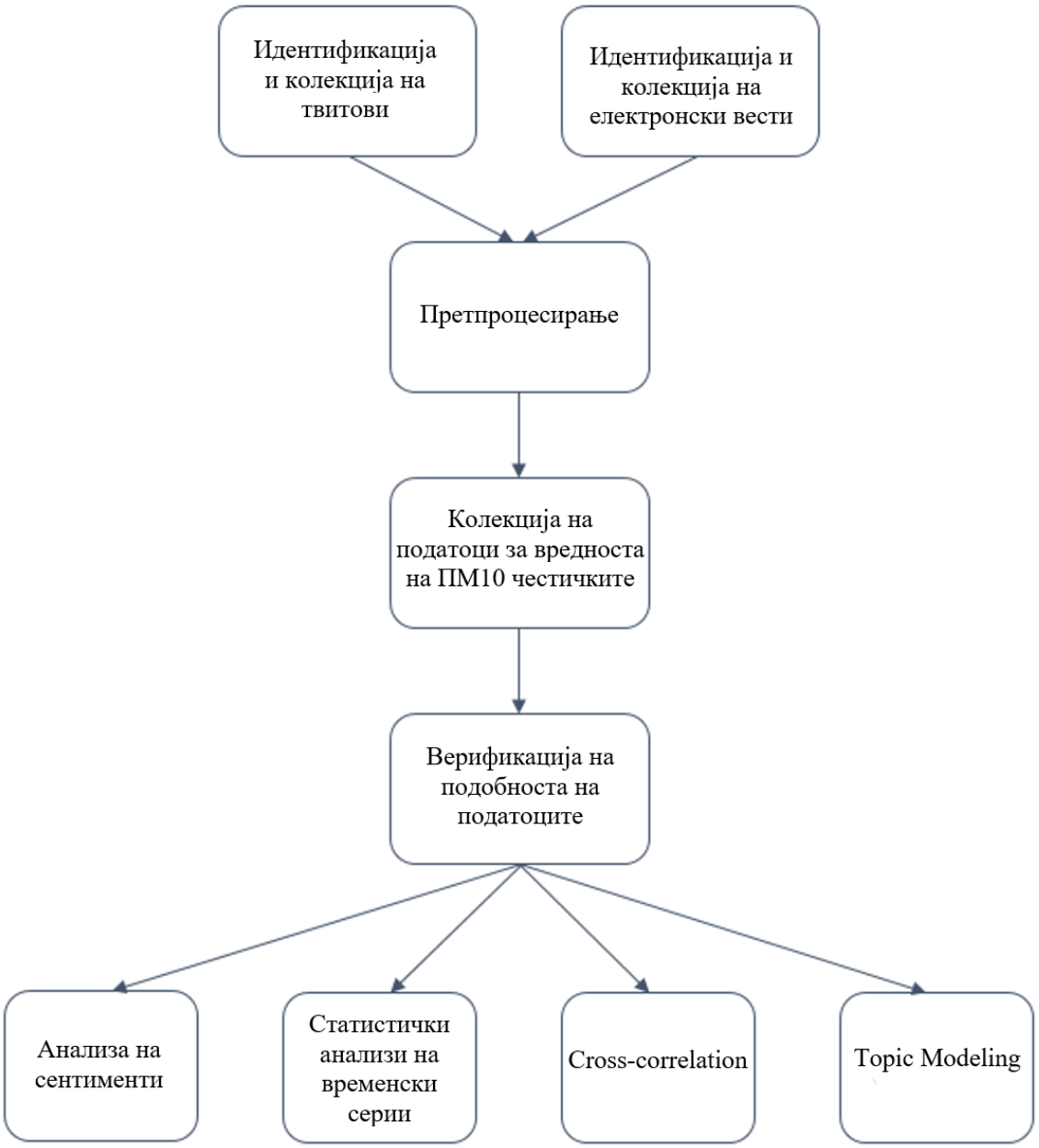
2.3. Официјални ПМ10 податоци

За да се направи споредба помеѓу врвовите и падовите на фреквенцијата на твитови и вести со оние на вредностите на ПМ10 честичките измерени од официјални мерни станици, се собираа податоци за Македонија, Србија, Црна Гора и Босна и Херцеговина од 21, 37, 6 и 16 официјални мерни станици, соодветно. Нивоата на ПМ10 честички измерени на секој час во период од 01.11.2021 до 28.02.2022, беа агрегирани на неделно ниво, како би соодветствувале на колекцијата на твитови и вести. Дополнително, ПМ10 податоците од мерните станици во секоја држава од интерес беа агрегирани за да се опфати целата нејзина територија.

Важно е да се напомене дека податоци за нивоата на ПМ10 честичките во Србија беа достапни во период од 01.12.2021 до 28.02.2022.

3. Методологија

Пред деталниот преглед на користената методологија, на *Слика 2*. е прикажан процесот на колекција на податоци и спроведените анализи врз нив.



Слика 2. Колекција на податоци и процес на анализа

3.1. Sentiment Analysis (Анализа на сентименти)

3.1.1. Што е Sentiment Analysis?

Со зголемената употреба на социјалните мрежи, јавното искажување на ставови и чувства за луѓето станува секојдневие. Оттука, се наметнува потребата од алатка која би можела автоматски и на ефикасен начин да извлече суштинско знаење од големи количества на неорганизиран и неструктуриран текст кој произлегува од социјалните медиуми, микро-блогирањето, форумите, коментарите и слично.

Sentiment Analysis е техника за обработка на природни јазици која идентификува дали во одреден текст се искажани позитивни, негативни или неутрални чувства. Покрај поларноста на текстот (позитивна, негативна, неутрална), може да се одредат и специфични емоции кои преовладуваат во него (лутина, среќа, тага и сл.). Луѓето ги асоцираат зборовите, фразите и реченициte со емоции, а областа на текстуална анализа на сентименти се потпира на алгоритми од машинско учење и податочно рударство за да ги декодира и квантифицира емоциите содржани во текст. Анализата на сентименти е широка област, но воглавно се сведува на два главни пристапи: *лексички пристап* и *пристап на машинско учење*.

3.1.2. Пристапи на Sentiment Analysis

Лексичките пристапи имаат тенденција да креираат лексикон, односно „речник на сентименти“, мапирајќи ги зборовите со сентименти. Вака креираниот речник е доволен за проценка на сентиментот искажан во парче текст (фраза, реченица, параграф). Сентиментите може да бидат категориски, односно позитивни, неутрални, негативни, или пак нумерички, односно опсег на интензитетот на чувствата искажани во текстот. Ваквите пристапи го проценуваат сентиментот на целата реченица врз основа на сентиментот на секој збор во неа. Предноста на лексичките пристапи се сведува токму на „речникот на сентименти“, поради кој се изоставува потребата од тренирање на модел користејќи лабелирани податоци за предвидување на сентиментот.

Од друга страна, пристапите на машинско учење користат претходно лабелирани податоци за да го проценат сентиментот на нови, непознати реченици. Вклучуваат процес на тренирање на модел со употреба на претходно познат текст за предвидување/класификација на сентиментот на целосно нов текст. Со зголемување на количеството податоци, генерално се подобрува и резултатот од предвидувањето/класификацијата, што може да се истакне како предност на пристапите на машинско учење. Сепак, голем недостаток во споредба со лексичките пристапи, е фактот дека се потребни претходно означени податоци за воопшто да се користат модели на машинско учење.

3.1.3. VADER (Valence Aware Dictionary and sEntiment Reasoner)

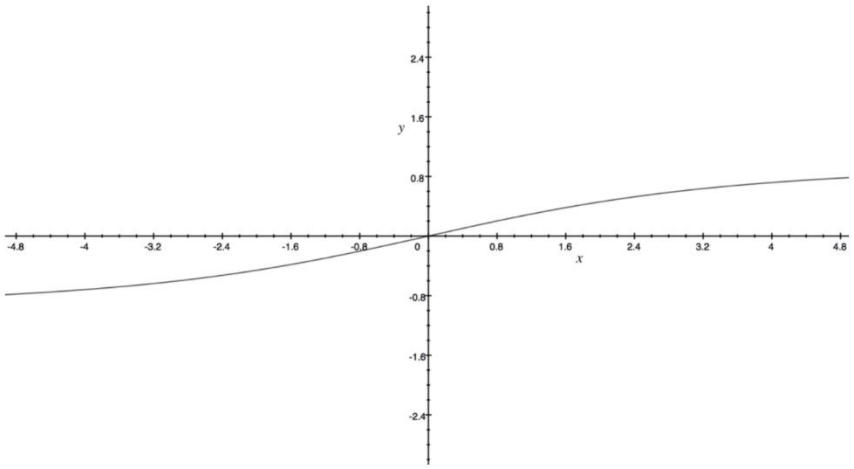
За одредување на чувствата искажани кон аерозагадувањето во податоците собрани од Twitter и од вестите, овој дипломски труд се фокусира на лексички пристап. Поконкретно, акцентот е ставен на *VADER (Valence Aware Dictionary and sEntiment Reasoner)* – лексички метод за анализа на сентименти, наменет и оптимизиран за микро-блогови како што се твитовите. *VADER* анализата на сентименти се потпира на речник во кој лексичките карактеристики се мапирани во интензитети на емоција, како и на пет едноставни хеуристики кои одредуваат како контекстуалните елементи (пример: интерпункциски знаци, големи букви) влијаат врз зголемувањето/намалувањето на сентиментот на текстот.

Под лексичка карактеристика се подразбира сè што се користи при текстуална комуникација, вклучувајќи зборови, емотикони, акроними и сленг. Интензитетот на емоција на парче текст се одредува со сумирање на таквиот интензитет на секој збор во текстот. Истиот се мери на скала од -4 до +4, каде -4 означува најнегативни емоции, 0 означува неутрални, а +4 означува најпозитивни емоции. Моќта на *VADER* лежи во користењето на колективно мислење (*wisdom of the crowd*) при проценка на интензитетот на емоција на секоја лексичка карактеристика. За да се избегне субјективноста при креирање на „речник на сентименти“, креаторите на овој метод интервјуирале вработени лица во Amazon и го пресметале просекот од нивните индивидуални проценки за интензитет на емоција на секоја лексичка карактеристика во речникот.

Во програмскиот јазик *Python*, интензитетот на емоција на реченица (сума на индивидуални интензитети на емоција на секој збор во реченицата) се мапира во ранг [-1, +1], користејќи ја следната нормализација:

$$x = \frac{x}{\sqrt{x^2+\alpha}} \qquad (1)$$

Во (1), x е сумата на интензитети на емоција на зборовите кои се дел од реченицата од интерес, а α е нормализациска константа чија предодредена вредност е еднаква на 15. Графикот на нормализација може да се погледне на *Слика 3.* Воочливо е дека со зголемувањето на бројот на зборови во текстот, се добива интензитет на емоција со вредност блиску до -1 или 1. Оттука, *VADER* дава најдобри резултати кога е применет врз кратки текстови, како што се твитови.



Слика 3. График на нормализација на интензитетите на емоција на ниво на реченица добиени при VADER анализа на сентименти

Петте едноставни хеуристики според кои *VADER* го квантифицира влијанието на контекстуалните елементи во реченицата врз зголемувањето, односно намалувањето на позитивните и негативните сентименти, се објаснети во продолжение. Повторно, ефектот на овие хеуристики е квантифициран потпирајќи се на колективното мислење на луѓе вработени во Amazon.

1. **Интерпункција**
   При споредба на реченициte *“I like it”* и *“I like it !!!”,* човек лесно би утврдил дека втората реченица искажува поинтензивни емоции од првата, поради што и интензитетот на емоција одреден од *VADER* би требало да биде повисок. *VADER* анализата на сентименти го зема ова во предвид така што го засилува, т.е. намалува интензитетот на емоции на реченицата, пропорционално со бројот на *„!“* или *„?“* на крајот од реченицата. На почеток се пресметува интензитетот на емоции без интерпункциските знаци и доколку тој е позитивен, *VADER* додава емпириски пресметана вредност за секој извичник (0.292) и за секој прашалник (0.18). Доколку резултатот е негативен, ваквите вредности се одземаат.

2. **Капитализација**
   Големите букви во текстуална комуникација честопати се користат за искажување интензивирани чувства. Така, во реченицитe *“The food was amazing”* и *“The food was AMAZING”*, *VADER* би го зголемил интензитетот на емоција на зборот напишан со големи букви за 0.733 доколку истиот е позитивен, а би го намалил за исто толку доколку е негативен.

3. **Модификатори на степен на емоција**
   Ако за пример се земат реченициte *“Definitely nice”* и *“Sort of nice”*, прилогот во првата реченица има за цел да го зголеми, а во втората реченица да го намали интензитетот на емоција. За справување со вакви реченици, *VADER* има посебен речник на прилози кои би можеле да го зголемат или намалат интензитетот на емоција за 0.293 на зборот кој следи по нив, во зависност дали тој е проценет како позитивен или негативен.

4. **Промена на поларноста поради сврзникот “but”**
   Честопати, сврзникот *“but”* поврзува две контрастни реченици од кои емоцијата искажана во втората реченица е подоминантна. *VADER* го намалува интензитетот на емоција на зборовите кои се наоѓаат пред сврзникот “*but*” за 50%, а го зголемува за 150% на оние кои следат во подреченицата по него.

5. **Анализа на три-грами за детекција на негација**
   Петтата и последна хеуристика го анализира три-грамот пред лексичката карактеристика за да детектира евентуална негација на поларноста. *VADER* ја доловува негацијата со множење на интензитетот на емоции на лексичката карактеристика со емпириски пресметана вредност еднаква на 0.74.

За поделба на твитовите и вестите според сентиментот кој го искажуваат, во овој дипломски труд се користеше границата (threshold): >= 0.05 за позитивни сентименти, > -0.05 и < 0.05 за неутрални сентименти и <= 0.05 за негативни сентименти.

Пред да се искористи вградената *NLTK VADER* алатка за анализа на сентименти во *Python*, податочните множества беа преведени на англиски јазик користејќи автоматски онлајн преведувач.

3.2. Статистички анализи на временски серии

Кога станува збор за анализа на временски серии, мерењето на сличноста е од голема важност за да се процени случајната врска помеѓу два сигнали во времето. За да се направи споредба на фреквенциите на поделените твитови и вести по сентимент со податоците за PM10, во овој дипломски труд се користеше *Cross-Correlation*.

3.2.1. Cross Correlation Function – CCF (Функција на крос-корелација)

Функцијата на крос-корелација (*CCF*) е мерка за сличност на две серии како функција од релативното поместување на едната серија во однос на другата. Може да се дефинира како корелација помеѓу опсервациите на две временски серии x_t и y_t, одделени за k временски единици (корелацијата помеѓу y_{t+k} и x_t), каде што k се нарекува *lag* (заостанување). Интервалот на доверба се пресметува со следната формула:

$$\pm \frac{2}{\sqrt{n-|k|}} \qquad (2)$$

Во (2), n го претставува бројот на опсервации, а k е lag-от, односно задоцнувањето. Корелацијата се смета за сигнификантна доколку нејзината апсолутна вредност е поголема од $\frac{2}{\sqrt{n-|k|}}$.

CCF се потпира на претпоставката дека податоците се стационарни, односно дека средната вредност и варијансата се константни и непроменливи со текот на времето. Во случаи кога е детектиран силен растечки или опаѓачки тренд, нестационарноста може да се адресира со песметка на прв извод на податоците.

3.2.2. Mann-Kendall-ов тест

За тестирање на стационарноста на различните групи на сентименти добиени при анализа на сентименти на Twitter податоците и податоците од електронски вести, како и добиените ПМ10 податоци, во овој дипломски труд беше искористен непараметарскиот *Mann-Kendall*-ов тест. Нултата хипотеза на овој тест е дека во податоците не постои монотон тренд, додека пак алтернативната хипотеза е дека постои тренд кој може да биде позитивен или негативен. Тестот го одредува трендот, односно монотоноста, така што ја анализира разликата во сигналите во секоја временски точка со секоја наредна точка во времето. Доколку тестот детектира тренд, значи дека вредностите константно се зголемуваат или опаѓаат. *Mann-Kendall*-овиот тест откри опаѓачки тренд во податоците собрани од електронските вести од *Time.rs*. Затоа, пред пресметката на *CCF*, беше пресметан прв извод на податоците за да се постигне нивна стационарност. За останатите податоци, тестот не откри ниту опаѓачки, ниту растечки трендови.

3.3. Topic modeling

3.3.1. Што е Topic modeling?

Во задачите на *Natural Language Understanding* (разбирање на природните јазици) постои хиерархија преку која може да се извлече знаење од текст, почнувајќи од зборови, реченици, параграфи, па се до документи. На ниво на документи, текстот најефикасно може да се разбере преку анализа на темите опфатени во него. *Topic modeling* е техника на ненадгледувано машинско учење која автоматски анализира множество документи (во овој дипломски труд – твитови) за да идентификува апстрактни „теми“ кои најдобро ја опишуваат содржината во документите. Во документ кој елаборира специфична тема, фреквенцијата на појава на поими карактеристични за таа тема ќе биде повисока отколку за други зборови. Сличните поими се групираат во кластер, а темата се одредува врз основа на статистичката веројатност за појава на тие поими. Така, topic моделите откриваат латентни семантички структури кои се појавуваат во неструктурирани податоци, какви што денес преовладуваат на Интернетот. Спротивно од *Topic classification*, која како техника на надгледувано учење има за цел да детектира предефинирани теми во корпусот на документи, *Topic modeling* резултира со множества на поими од кои со дополнителна визуелна инспекција треба да се заклучи темата која го карактеризира секое множество.

За разлика од долги текстови, како што се новинарски статии, кои имаат јасна структура и контекст и користат официјален начин на изразување, кратките текстови, како што се микро-блоговите, често имаат недостаток на структура и контекст, а начинот на изразување вклучува употреба на сленг и кратенки. Поради ова, примената на *topic modeling* за откривање на теми во кратки текстови е значително покомплексна задача.

За да се заклучи што придонесува кон катастрофалната состојба со аерозагадувањето и кому му се препишува вината според јавното мислење, за целите на овој дипломски труд се применети два *topic modeling* пристапи: *Latent Dirichlet Allocation (LDA)* and *Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM)*. Дополнително, направена е евалуација и споредба на нивните перформанси врз корпусот на кратки документи, односно твитови и вести, користејќи ја метриката *Topic Coherence score.* Пред примената на алгоритмите за *topic modeling,* податоците беа претпроцесирани, така што беа отстранети симболите за retweets, интерпункциските знаци, линковите (URLs), емотиконите и непотребните празни места. За да се стави акцент на важните информации во текстовите, од податоците беа отстранети сите *stopwords*, односно зборови кои се често користени во јазикот (како што се сврзниците), но немаат никаква улога при одредувањето на теми. Дополнително, податоците беа *токенизирани*, така што секоја реченица во документите беше поделена на помали делови, наречени *токени*, составени

од само еден збор. Вака добиените токени беа сведени на својата *лема*, т.е. коренот на зборот. На пример, коренот на зборовите eating, eats, eaten е зборот eat.

3.3.2. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation, или скратено LDA, е еден од најпознатите и најчесто користени алгоритми за *topic modeling*. Се заснова на претпоставката дека секој документ е составен од дистрибуција на теми, а секоја тема е составена од дистрибуција на зборови. Имајќи ги документите (твитовите) и зборовите во нив, задачата на алгоритмот е да ги конструира скриените, односно латентните теми, пресметувајќи го придонесот на секоја од нив во документот. Во продолжение следи објаснување на секој збор во називот на алгоритмот, со цел да се стекне интуиција за неговата функција.

- **Latent** –темите во текстот се „скриени“, односно латентни, а целта на алгоритмот е да ги кластерира текстовите со кои располага, на начин интерпретабилен за луѓето.
- **Dirichlet** – моделот користи Дирихлетова распределба како приор за генерирање на дистрибуција на теми во рамки на еден документ и дистрибуција на зборови во рамки на секоја тема.
- **Allocation** – алгоритмот се обидува да алоцира теми на достапните текстови.

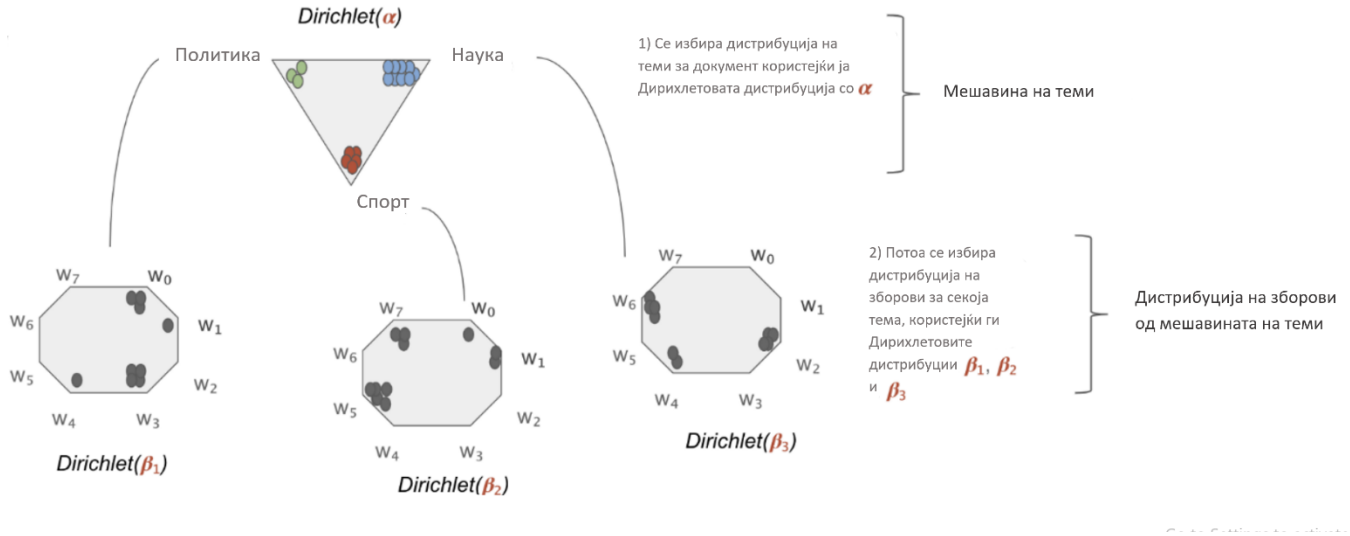
За објаснување на начинот на кој LDA работи, соодветно е да се земе пример за споредба на распределбите на веројатност на множество теми во корпус на документи. Доколку корпусот кој се моделира содржи документи од 3 многу различни области, посакуваниот тип на распределба би бил таков што една од темите ќе има најголема тежина во секоја од дистрибуциите, додека другите две теми ќе имаат помали тежини. Пример за посакувана распределба на три теми (Спорт, Политика и Наука) е:

- **Дистрибуција X: 90% тема спорт**, 5% тема Политика, 5% тема Наука
- **Дистрибуција Y:** 5% тема спорт, **90% тема Политика**, 5% тема Наука
- **Дистрибуција Z:** 5% тема спорт, 5% тема Политика, **90% тема Наука**

Доколку се извлече случајна распределба од оваа Дирихлетова распределба, параметризирана со големи тежини на една тема, веројатно е да се добие дистрибуција која ќе наликува на Дистрибуција X, Дистрибуција Y или Дистрибуција Z.

Слика 4. е илустративен пример кој ги прикажува чекорите на LDA за откривање на теми во секој документ во корпусот. За поедноставна илустрација, зададени се $K = 3$ теми и $N = 8$ збора. Темите се исти како претходно-зададените, односно Спорт, Политика и Наука. Откако сите документи од корпусот се генерирани следејќи ги чекорите

објаснети на *Слика 4.*, дистрибуцијата на зборови се користи за одредување на хиперпараметрите α и β. Честопати, и бројот на итерации е хиперпараметар.



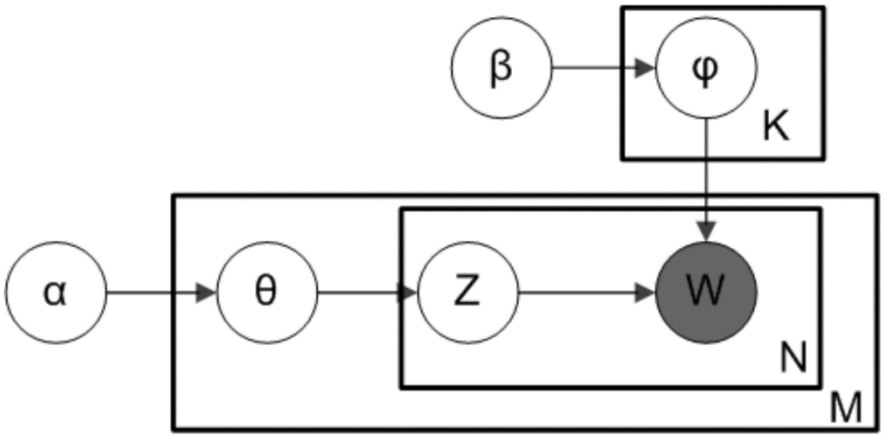
Слика 4. Илустративен приказ на LDA алгоритмот

Бидејќи секоја тема има процентуален придонес во документот, 0.8*Наука, 0.1*Спорт, 0.1*Политика би значело дека 80% од документот содржи зборови поврзани со темата Наука, 10% зборови поврзани со Спорт и 10% зборови поврзани со Политика. На сличен начин, зборовите во секоја тема имаат свој процентуален придонес. Така, на *Слика 4.*, за темата Наука, зборовите w6, w4 и w2 имаат најголем процентуален придонес.

Рангираните списоци на зборови поврзани со даден збор w_n се добиваат со пресметување на збирот на тежината на секоја тема генерирана од LDA помножена со тежината на секој збор w_n содржан во таа тема. Тежината на рангирањето на зборот i се пресметува на следниов начин:

$$w_i = \sum_{j=1..N} w_{ij} * w_{nj} \qquad (3)$$

Во (3), N е бројот на теми, а w_{ij} ја означува тежината на зборот i во темата j. Шематски приказ на илустративниот пример од *Слика 4.* Е прикажан на *Слика 5.*
Од Дирихлетова дистрибуција $Dir(\alpha)$ се извлекува случаен примерок, односно *дистрибуција на теми* во документ од множеството.



Слика 5. Шематски приказ на LDA алгоритмот

Таквата дистрибуција на теми е θ. Од θ, се извлекува тема Z, врз основа на дистрибуцијата на темите. Потоа, од друга Дирихлетова дистрибуција $Dir(\beta)$ се извлекува случаен примерок, односно *дистрибуција на зборови* во темата Z. Ваквата дистрибуција на зборови е φ. Од φ, се избира збор w.

Математички, ова може да се претстави со формулата (4):

$$P(w, Z, \theta, \varphi, \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j\ ;\ \alpha) \prod_{i=1}^{K} P(\varphi_i\ ;\ \beta)\ \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j)\ P(W_j \mid \varphi_{Z_{j,t}}) \qquad (4)$$

За градење на LDA моделот, во овој дипломски труд се користеше *Python* библиотеката *genism*, а за креирање на интерактивна визуелизација на резултатот од моделот се користеше библиотеката *pyLDAvis*, исто така во програмскиот јазик *Python*.

3.3.3. Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM)

Перформансите на LDA се одлични кога се применува врз големи текстови (> 50 збора), но склони се кон опаѓање кога се користи за моделирање на теми во куси текстови, како што се микро-блоговите. Микро-блоговите често се однесуваат на само една тема, што е спротивно на претпоставката на LDA.

Алтернативен алторитам за *topic modeling* кој е модификација на LDA, е GSDMM, чијашто претпоставка е токму дека во секој документ преовладува една, единствена тема, што го прави соодветен за детекција на теми во куси документи, како што се твитови и вести. Дополнителна предност над LDA алгоритмот е тоа што не е потребно да се внесе однапред дефиниран број на теми. Доволно е да се зададе максимален број на теми, а

GSDMM го одредува оптималниот број на теми итерирајќи и преназначувајќи ги на сличен начин како што функционира *Наивниот Баесов Класификатор*, односно документите (твитовите) се доделуваат на кластери земајќи ја предвид највисоката условна веројатност.

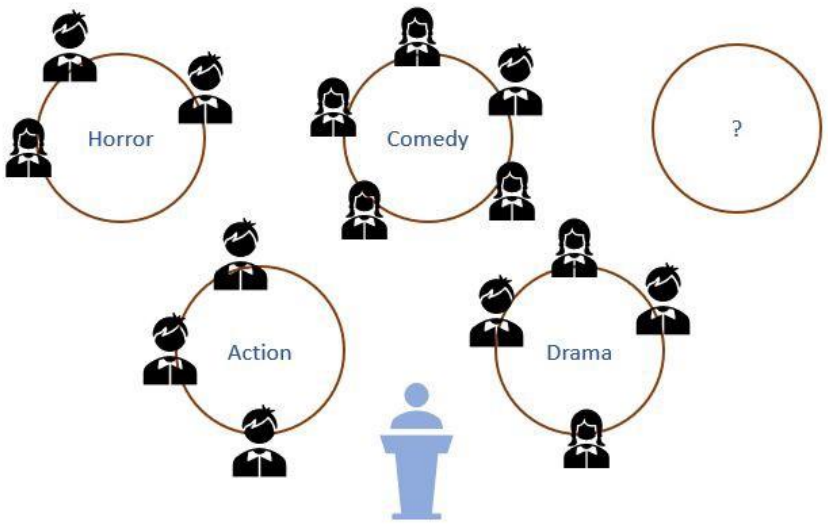
При објаснување на GSDMM, во литературата често се референцира аналогна процедура наречена **Movie Group Process**, чие објаснување следи во продолжение. Нека еден професор држи час по филм со група студенти. На почеток од часот, професорот бара од студентите да запишат куса листа на омилени филмови на лист хартија. *Студентите се аналогија за документите (твитовите), а нивните листи со филмови се аналогија за зборови во документите.* Потоа, студентите се случајно распределени на K маси. Целта е да се кластерираат и групираат студентите, така што оние кои ќе седат на иста маса ќе имаат слични интереси за филм. На крај, професорот повеќепати ги именува студентите еден по еден и секој од нив треба да одбере нова маса, почитувајќи ги следните правила:

1. **Правило 1:** Да избере маса со повеќе студенти.

2. **Правило 2:** Да избере маса каде студентите имаат сличен интерес за филмови.

Со првото правило се подобрува **комплетноста**. Наместо да бидат распределени на различни маси, сите студенти со сличен интерес за филмови ќе седат на иста маса.

Второто правило води кон подобра **хомогеност**, овозможувајќи *само* студенти кои имаат слични интереси седат на иста маса.



Слика 6. Визуелен приказ на Movie Group Process аналогијата

Двата чекори се повторуваат сè додека не се постигне оптимален број на кластери, притоа очекувајќи некои маси да исчезнат, а други да се зголемат. Надежно, крајниот резултат ќе биде оптимален број на кластери на студенти со слични преференции за филм. На овој начин работи GSDMM алгоритмот. *Слика 6.* прикажува визуелна репрезентација на *Movie Group Process* аналогијата.

Слично како LDA, и GSDMM моделот има хиперпараметри α и β, при што α ја контролира веројатноста дека студентот ќе седне на маса која е празна во моментот. Кога $\alpha = 0$, ниту еден студент нема да седне на празна маса. β го контролира афинитетот на студентот кон други студенти со слични интереси за филм. Ниска вредност на β значи дека студентите сакаат да делат маса со студенти со сличен интерес, а висока вредност на β значи дека студентите преферираат да седнат на популарна маса со голем број на студенти, без разлика на нивниот афинитет, наспроти маса со помалку студенти со сличен интерес како нивниот.

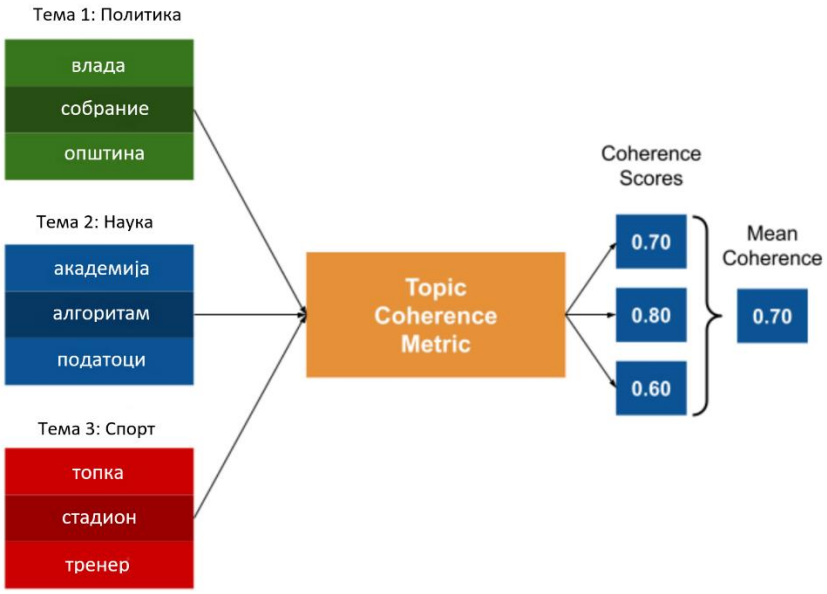
За градење на GSDMM модел, во овој дипломски труд се користеше *Python* пакетот инсталиран директно од следниот Github репозиториум: https://github.com/rwalk/gsdmm.git.

3.3.4. Topic Coherence score

Компјутерите анализираат текстуални податоци на многу поразличен начин од луѓето. Така, при откривање на теми во одреден текст, *topic modeling* алгоритмите се потпираат на математички и статистички пресметки. Но, математички оптималните теми не се секогаш интерпретабилни за човечкото око.

Topic Coherence Score е објективна метрика која се темели на дистрибутивната лингвистичка хипотеза дека зборови со слично значење имаат тенденција да се појавуваат во слични контексти. Темите се сметаат за кохерентни кога сите или повеќето зборови во нив се тесно поврзани. Се мери квалитетот на темата мерејќи го степенот на семантичка сличност помеѓу зборовите со висок процентуален придонес во таа тема. На овој начин се одредува кои теми се семантички интерпретабилни, а кои теми се артефакти на статистички пресметки. Крајниот резултат е средна вредност од резултатите за кохерентност на секоја тема добиена при *topic modeling* (*Слика 7.*).

Во овој дипломски труд, се користеше една од најпопуларните метрики за кохерентност: *C_v Coherence Score.* Оваа метрика креира вектори на зборови од содржина на текст користејќи ги нивните истовремени појави, а потоа го пресметува резултатот користејќи ги мерката за асоцијација *Normalized Pointwise Mutual Information (NPMI)* и *косинусната сличност*. Вредноста на *Topic coherence* за добиените теми во сите податочни множества беше пресметана со употреба на библиотеката *genism* во програмскиот јазик *Python*.



Слика 7. Topic Coherence score на повеќе теми

4. Резултати

4.1. Добиени сентименти

Статистичките податоци за колекцијата на твитови и вести собрана во текот на ова 17-неделно истражување, како и сентиментите добиени со примена на *Анализа на сентименти* врз сите податочни множества, се прикажани во *Табела 1.* Воочлив е доминантниот процент на негативни сентименти во секое множество на податоци, додека неутрален сентимент е најмалку застапен.

Релативно високиот број на *Retweets* укажува на меѓусебна согласност и одобрување на ставот во врска со аерозагадувањето кој преовладува кај Твитер корисниците. Од голема важност е да се напомене дека вкупниот број на твитови не е еднаков на збирот на ретвитови и уникатни твитови (вклучувајќи ги и одговорите на оригиналниот твит). Понекогаш, оригиналниот твит што е споделен ("ретвитуван") не е опфатен во колекцијата на твитови бидејќи датира во минатото, надвор од временскиот опсег на оваа студија. Освен тоа, може да се случи неколку оригинални твитови да бидат плагијати. Затоа, за уникатни се сметаат твитовите со уникатна претходно претпроцесирана содржина (со отстранети симболи за ретвит и специјални знаци). На сличен начин, вестите со автентична содржина се сметаат за уникатни.

Name	Total number	Retweets (%)	Unique (%)	Negative (%)	Positive (%)	Neutral (%)
Macedonian Tweets	1018	33.99	53.93	64.3	19.4	16.2
Time.mk teasers	994		48.89	55.2	39.8	4.9
Western Balkan Tweets	2664	47.56	44.52	54.3	29.09	16.61
Time.rs teasers	709		73.77	64.2	27.5	8.3

Табела 1. Статистички податоци за собраната колекција на твитови и вести

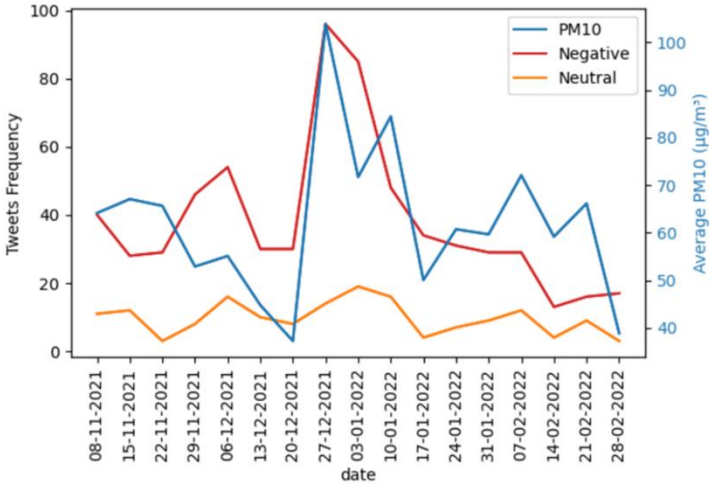
4.2. Cross-correlation

За да се утврди сличноста помеѓу твитер дискусиите, вестите и измерените нивоа на ПМ10 честички во воздухот, неделните фреквенции на твитови и вести класифицирани според сентиментот се претставени на график наспроти официјалните ПМ10 податоци. Различните групи на сентименти добиени од македонските твитови беа споредени со податоците за ПМ10 честички добиени од мерните станици во Македонија. Групите на сентименти добиени од останатите твитови беа споредени со ПМ10 податоците измерени

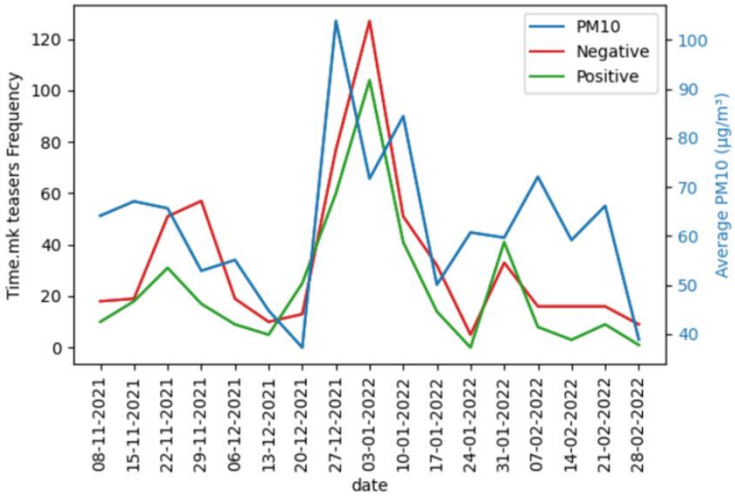
во Србија, Црна Гора и Босна и Херцеговина за да се провери евентуална кореспонденција.

Максималната крос-корелација помеѓу сите категории македонски твитови и ПМ10 честичките во земјата имаше lag (задоцнување) 0, што укажува дека нема временско поместување нанапред или наназад помеѓу нивоата на ПМ10 честичките и твитер дискусиите во земјата. Максималната крос-корелација помеѓу бројот на негативни твитови и ПМ10 честички беше со коефициент од 0.62 ($p = 0.001$), а со коефициент од 0.52 ($p < 0.001$) меѓу бројот на неутрални твитови и ПМ10 честичките (*Слика 8.*). Крос-корелацијата меѓу позитивните твитови и ПМ10 честичките беше незначителна.

Што се однесува до групите на вести од Time.mk, максималната крос-корелација меѓу фреквенцијата на негативни вести и податоците за ПМ10 беше 0.52 ($p = 0.002$), меѓу позитивните вести и податоците за ПМ10 беше 0.52 ($p < 0.001$) (*Слика 9.*), и двете со lag 0; додека помеѓу неутралните написи и податоците за ПМ10 коефициентот беше незначителна.



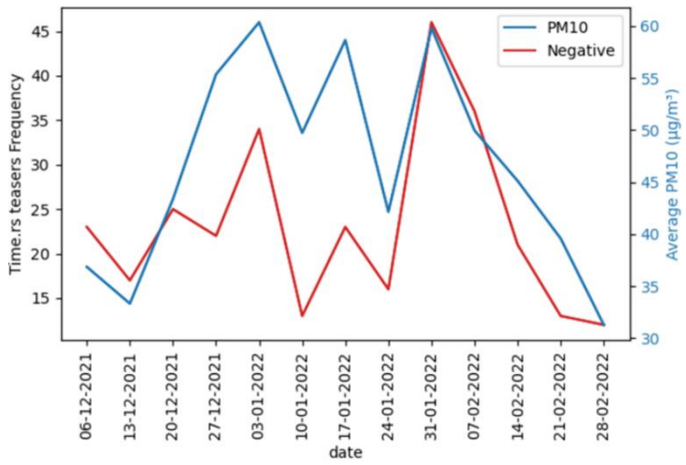
Слика 8. Неделна споредба на официјалните ПМ10 податоци во Македонија и фреквенцијата на македонски твитови



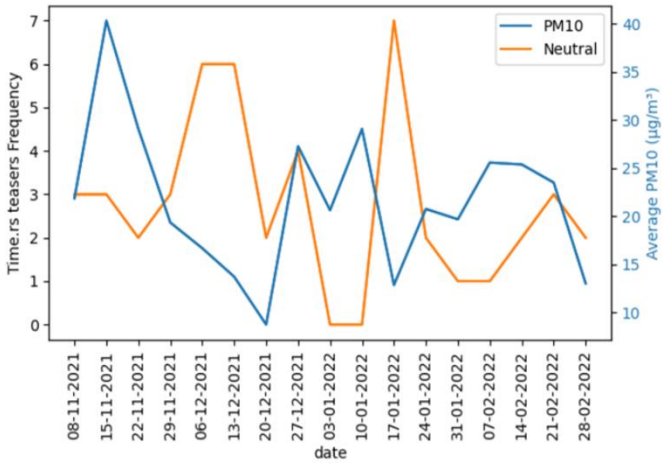
Слика 9. Неделна споредба на официјалните ПМ10 податоци во Македонија и фреквенцијата на Time.mk вести

Не беше откриена значајна крос-корелација помеѓу преостанатите твитови и некои од ПМ10 податоците измерени во Србија, Босна и Херцеговина и Црна Гора.

Во однос на вестите од Time.rs, максималната крос корелација со lag 0 меѓу негативните вести и ПМ10 податоците измерени во Србија изнесуваше 0.66 ($p < 0.001$) (*Слика 10.*), додека помеѓу неутралните вести и ПМ10 податоците измерени во Црна Гора изнесуваше 0.65 ($p < 0.001$) со lag 3. (*Слика 11.*).

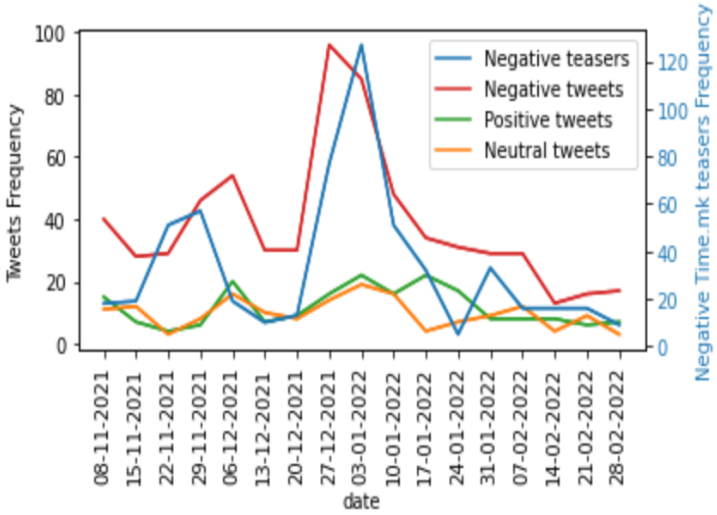


Слика 10. Неделна споредба на официјалните ПМ10 податоци во Србија и фреквенцијата на Time.rs вести

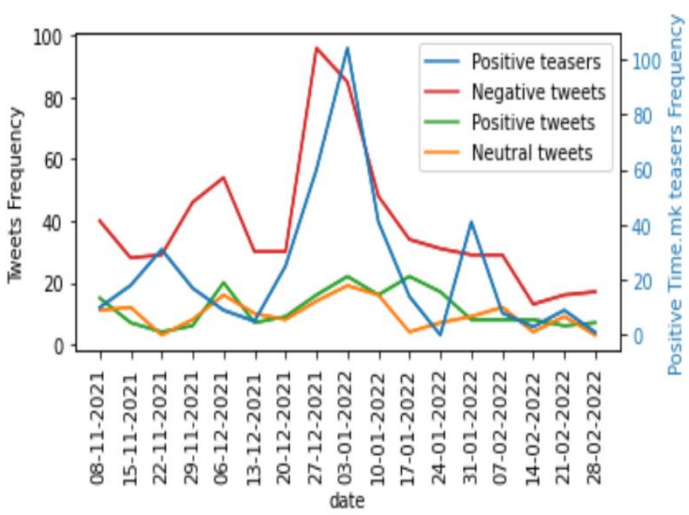


Слика 11. Неделна споредба на официјалните ПМ10 податоци во Црна Гора и фреквенцијата на Time.rs вести

Дополнително, резултатот од тестирањето на корелација помеѓу твитовите и вестите откри максимална крос-корелација меѓу негативните македонски твитови и негативните Time.mk вести еднаква на 0.8 со lag 0 (p = 0.0001), помеѓу позитивните македонски твитови и негативните Time.mk вести еднаква на 0.55 со lag 1 (p = 0.02077) и помеѓу неутралните македонски твитови и негативните Time.mk вести еднаква на 0.53 со lag 0 (п = 0.02854) (*Слика 12.*).



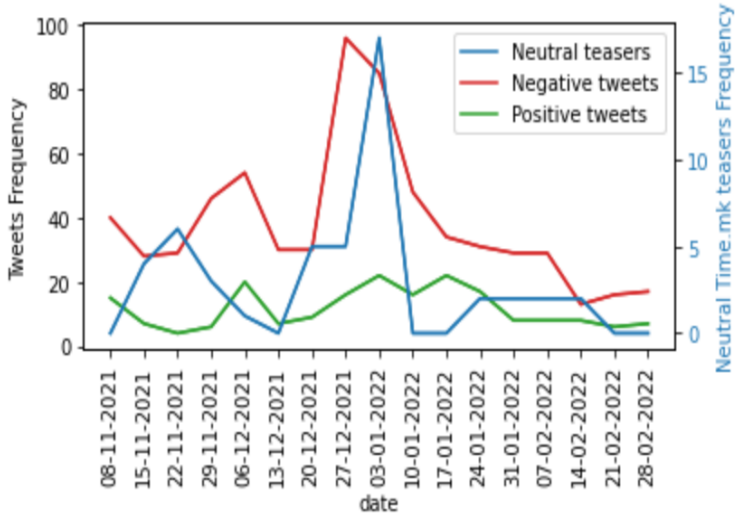
Слика 12. Неделна споредба на фреквенциите на негативни Time.mk вести и фреквенциите на македонски твитови



Слика 13. Неделна споредба на фреквенциите на позитивни Time.mk вести и фреквенциите на македонски твитови

Максималната крос-корелација меѓу негативните македонски твитови и позитивните Time.mk вести е 0.77 со lag 0 (p = 0.00028), помеѓу позитивните македонски твитови и позитивните Time.mk вести е 0.59 со lag 2 (p = 0.01307) и помеѓу неутралните македонски твитови и позитивните Time.mk вести е 0.59 со lag 0 (p = 0.01315) (*Слика 13.*).

Резултатите за крос-корелација меѓу негативни македонски твитови и неутрални Time.mk вести покажа максимална вредност од 0.58 со lag 0 (p = 0.01419), а помеѓу позитивни македонски твитови и неутрални Time.mk вести од 0.55 со lag 2 (p = 0.02319), додека крос-корелацијата помеѓу неутралните македонски твитови и неутралните Time.mk вести не беше сигнификантна (*Слика 14.*).



Слика 14. Неделна споредба на фреквенциите на неутрални Time.mk вести и фреквенциите на македонски твитови

Статистичките тестови потврдија дека не постои сигнификантна крос-корелација помеѓу ниедна група на сентименти од Балканските твитови и ниедна група на сентименти од Time.rs вестите.

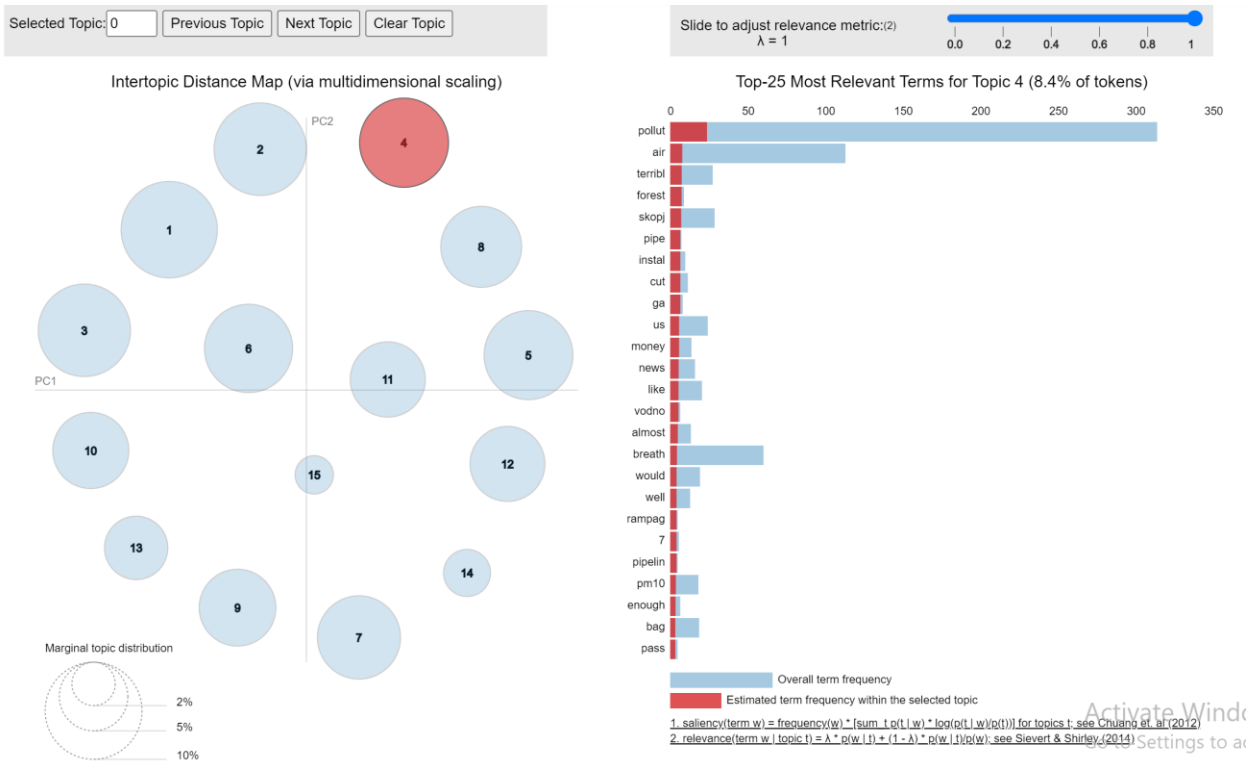
4.3. Добиени теми

Topic modeling е предизвикувачка NLP (Natural Language Processing) задача и изборот на „најдобар модел“ не е поткрепен со факти. Поради контекстот, како и преводот на податоците користени во овој труд, темите може најдобро да ги протолкува човечко око. Сепак, покрај визуелна инспекција на кластерите, за проценка на квалитетот на секој модел се користеше метриката *Topic Coherence*. Процесот на тренирање на LDA и GSDMM вклучуваше и експериментирање со различен број на теми во ранг од 2 до 20, како и различни вредности за хиперпараметрите. Со мал број на теми, се извлекуваа премногу општи теми, резултирајќи со ниска вредност за *Topic Coherence*. Од друга страна, со голем број на теми, извлечените теми се повторуваа, што повторно резултираше со ниска вредност за *Topic Coherence*. Ваквите вредности за *Topic Coherence* сигнализираат дека избраниот број на теми е веројатно погрешен, односно премногу мал или премногу голем.

Според анализите, за GSDMM на македонските негативни твитови беа избрани 9 теми со topic coherence score од 0.51, додека за LDA беа избрани 15 теми со topic coherence score од 0.41. Во однос на податоците од Time.mk, со GSDMM беа избрани 5 теми (topic coherence score = 0.43), а со LDA 10 теми (topic coherence score = 0.41).

За GSDMM на негативни твитови од останатите земји од интерес беа избрани 7 теми (topic coherence score = 0.57), додека за LDA беа избрани 15 теми (topic coherence score = 0.38). За податоците од Time.rs беа избрани 5 теми (topic coherence score = 0.48) со GSDMM и 16 теми (topic coherence score = 0.40) со LDA.

Постоечки студии сугерираат дека традиционалните *topic modeling* алгоритми, како што е LDA, бележат огромна деградација на перформансите при нивна примена врз кратки текстови, што е во согласност со резултатите изнесени во овој дипломски труд. Затоа, акцентот беше ставен на темите добиени со GSDMM, презентирани во *Табела 2.* и *Табела 3.*, заедно со најважните зборови и нивната фреквенција на појавување. За илустративна цел, на *Слика 15.* прикажана е една визуелизација која е излез од LDA алгоритмот.



Слика 15. Пример за визуелизација на теми добиени со LDA алгоритмот и негативните македонски твитови

Topic (GSDMM)	Negative Macedonian Tweets	Topic (GSDMM)	Negative Balkan Tweets
Electricity	pollution (52), crisis (16), increase (16), price (14), air (11), electricity (11), cut (10), energy (9), terrible (9)	Power plant	pollution (407), air (188), cities (62), Belgrade (53), Serbia (48), terrible (45), plant (34), power (31), problem (24)
Transport	pollution (38), air (12), less (11), problem (9), car (7), brt (7), cities (6), need (6), tram (4)	Climate change	pollution (293), air (97), Serbia (83), problem (45), climate (30), change (30), increase (21), death (18), environment (17)
Government	pollution (12), people (11), anything (9), pm10 (7), immediately (6), mayor (6), don't (6), vmro (4), municipality (4)	Industrial air pollution	pollution (49), Sabac (25), miner (19), poison (17), plant (8), industry (7), factories (7), burn (7), suffer (7)
Health	die (8), pollution (6), breathe (6), people (6), lose (6), children (6), poison (4), burn (4), poor (4)	Health	pollution (7), Lazarevac (6), children (5), problem (3), nausea (3), dizziness (3), caugh (3), heart (3), symptom (3)

Табела 2. Теми добиени со примена на GSDMM алгоритмот за Topic modeling на македонските и балканските твитови

Topic (GSDMM)	Negative Time.mk Teasers	Topic (GSDMM)	Negative Time.rs Teasers
Politics	pollution (183), air (176), skopje (88), measure (87), pm10 (41), environment (33), world (38), ministry (24), Arsovska (19)	Health	air (277), pollution (249), Serbia (53), Belgrade (51), Sarajevo (41), environment (38), unhealthy (29), protect (25), health (23)
Health	air (105), reduce (34), cause (20), health (15), death (15), government (13), fight (9), research (9), project (8)	Protest against air pollution	air (101), pollution (92), protest (51), citizen (35), Serbia (34), gather (20), health (20), group (17), change (12)
Landfields	pollution (41), air (32), landfield (14), skopje (10), illegal (9), municipality (8), waste (6), mayor (5), burn (5)	Heating season	pollution (64), air (61), heat (16), mask (8), high (7), season (7), smog (7), increase (6), winter (6)
Protest against industrial air pollution	pollution (35), air (32), protest (17), citizen (14), factories (11), mills (8), winter (7), chimney (6), prevent (4)	Industrial air pollution	air (18), pollution (15), citizen (5), warn (5), politics (5), cause (4), factories (4), reduce (4), need (3)

Табела 3. Теми добиени со примена на GSDMM алгоритмот за Topic modeling на вести од Time.mk и Time.rs

5. Дискусија и заклучок

Спроведените анализи во рамки на овој дипломски труд покажаа дека македонските негативни твитови најдобро ги предвидуваат нивоата на ПМ10 честичките во државата, додека позитивните твитови немаат споредливи врвови и падови со измерените ПМ10 честички. Ова сугерира дека при ескалација на загадувањето на воздухот, јавноста изразува негативни чувства на загриженост на Твитер, што е во согласност со претпоставката дека Твитер дискусиите за загадувањето на воздухот ги рефлектираат измерените вредности на ПМ10 честичките.

Од друга страна, најголема максимална крос-корелација кај вестите од Time.mk беше откриена меѓу негативните вести и податоците за ПМ10 во Македонија, што алудира на тоа дека медиумите во државата споделуваат вистинити и релевантни информации за загаденоста на воздухот. Меѓутоа, не треба да се занемари високата крос-корелација меѓу позитивните вести и ПМ10 податоците, бидејќи имплицира можен обид за потиснување на јавни дискусии за аерозагадувањето од страна на медиумите.

Анализата за крос-корелација за детекција на можна коресподенција помеѓу Time.mk вестите и македонските твитови, покрај очекуваната корелација меѓу негативните твитови и негативните вести, открива и висока корелација на негативните македонски твитови со позитивните и со неутралните вести. Неконзистентноста во сентиментите би можела да значи дека јавноста е навистина свесна за состојбата со аерозагадувањето, потпирајќи се на други, дополнителни извори на информации покрај интернет порталите. Ова сугерира присуство на критичко размислување кај јавноста и желба за пристап до веродостојни податоци. Од друга страна, корелацијата со позитивните и неутралните вести би можела да сигнализира присуство на вести со едукативна содржина за поттикнување на еколошко однесување кај граѓаните, во обид за подигнување на свеста околу загадувањето на воздухот.

Алгоритмите за topic modeling покажаа дека македонската јавност ја припишува вината за аерозагадувањето и здравствените проблеми како негова последица, на производството на електрична енергија и транспортот. Дополнително, јавноста очекува државните институции да преземат мерки за решавање на овој проблем. Сепак, претходни истражувања во Македонија истакнуваат дека уделот кој домаќинствата и транспортот го имаат во аерозагадувањето во земјата е 90% наспроти 5%. Експериментирањето со различен број на теми откри мал број на Твитер дискусии околу негативните ефекти на биомасата, што укажува на ниска колективна свест околу овој загадувач. Големината на кластерите во кои се споменува биомасата е незначителна во споредба со кластерите кои се однесуваат на транспорт и резултираше со ниска кохерентност. Ова имплицира заклучок дека употребата на биомасата во домаќинствата не е меѓу кластерите добиени со највисока оценка на кохерентност, додека аерозагадувањето често се припишува на транспортот. Резултатите укажуваат на ниска

јавна свест околу главните загадувачи на воздухот во Македонија, нагласувајќи ја потребата од подигнување на јавната свест за негативните ефекти од користењето на биомасата, како и потребата од промовирање на еколошко однесување на оваа тема.

Спротивно на претпоставката дека Твитер дискусиите ја одразуваат концентрацијата на ПМ10 во воздухот, не беше откриена сличност помеѓу ниту една група на сентименти од твитовите од Западен Балкан и ПМ10 податоците измерени во Србија, Босна и Херцеговина и Црна Гора. Сепак, повеќето твитови беа класифицирани како негативни, што значи дека негативните чувства изразени во Твитер дискусиите преовладуваат дури и кога нивоата на ПМ10 се умерени или ниски.

Иако бројот на објавени вести на тема аерозагадување бележи тренд на опаѓање со текот на времето, измерената крос-корелација меѓу негативните вести од Time.rs и измерените вредности на ПМ10 честички во Србија беше прилично висока, што ја потврдува хипотезата дека медиумите ја рефлектираат реалната амбиентална состојба.

Примената на topic modeling откри дека голем број од твитовите за загаденост на воздухот се однесуваат на Србија и градови во Србија (Белград, Шабац итн.). На Твитер, јавноста генерално изразува загриженост за здравјето, климатските промени, индустриското загадување на воздухот и загадувањето предизвикано од електраните. И во овој случај, укажувањата се во насока на јавна несвесност за биомасата како главен загадувач на воздухот во Западен Балкан.

Дополнително, темите добиени од Time.mk и Time.rs вестите за аерозагадување откриваат дека негативните емоции најчесто се изразени на тема „проблеми со здравјето“ и „политика“. Ова може да значи дека медиумите ја припишуваат вината за аерозагадувањето и неговите штетни ефекти врз здравјето на луѓето на властите и државните институции во земјите од регионот на Западен Балкан.

Идните насоки за истражување за овој дипломски труд вклучуваат попрецизен превод на собраните податоци за да се подобри отпорноста на *VADER* кон сленг и сарказам, бидејќи контекстот и значењето на сленгот и сарказмот се губат при преводот. Понатаму, со користење на гео-локализирани твитови би се елиминирала можноста за колекција на хрватски твитови при пребарување со клучни зборови за собирање на твитови од Западен Балкан. Исто така, посебна анализа на податоците добиени од секоја мерна станица може да овозможи увид во загадувањето на воздухот во помали реони од секоја земја, имајќи во предвид дека нивоата на ПМ10 честичките варираат во голема мера од аспект на простор и време.

Се на сè, овој дипломски труд ги потврдува бенефитите од употреба на *NLP* техники, како што се *Анализа на сентименти* и *topic modeling*, за анализа на јавните чувства и ставови за важни теми како што е аерозагадувањето, како и за анализа на транспарентноста на медиумите при споделување на информации со јавноста. *Крос-корелацијата* помеѓу различните категории на сентименти изразени во дискусии на

социјални мрежи и измерените податоци за аерозагадување во одредена држава може да послужи како мерка за колективната свест околу аерозагадувањето. Техниките за *topic modeling* може да откријат проблеми во јавното мислење и да остварат свој придонес при изнаоѓањето начини за нивно решавање.

6. Користена литература

1. Colovic Daul, M., M. Kryzanowski, and O. Kujundzic. "Air Pollution and Human Health: The Case of the Western Balkans." UN Environ (2019).
2. Helotonio, Carvalho. "Air pollution-related deaths in Europe – time for action." Journal of Global Health 9.2 (2019).
3. Banja, M., G. Đukanović, and C. A. Belis. "Status of air pollutants and greenhouse gases in the Western Balkans." Publications Office of the European Union, EUR 30113 (2020): 1-53.
4. Meisner, Craig, Dragan Gjorgjev, and Fimka Tozija. "Estimating health impacts and economic costs of air pollution in the Republic of Macedonia." South Eastern European Journal of Public Health (SEEJPH) (2015).
5. Jovanovic, Mica. "Environmental Impact of Illegal Construction, Poor Planning and Design in Western Balkans: A Review."
6. I. Todorović, "HEAL: Biomass is one of main sources of air pollution in Western Balkans", Balkan Green Energy Group, 27-Jan-2022. Available at: https://balkangreenenergynews.com/heal-biomass-is-one-of-main-sources-of-air-pollution-in-western-balkans/#:~:text=The%20European%20Environmental%20Agency%20estimated,challenges%20around%20improving%20woodburning%20technology
7. Environmental Health, "Particulate matter (PM10 and PM2.5)", Environmental Health, 25-Nov-2020. Available at: https://www.health.nsw.gov.au/environment/air/Pages/particulate-matter.aspx
8. Jiang, Wei, et al. "Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter)." PloS one 10.10 (2015): e0141185
9. Hswen, Yulin, et al. "Feasibility of using social media to monitor outdoor air pollution in London, England." Preventive Medicine 121 (2019): 86-93.
10. Sachdeva, Sonya, and Sarah McCaffrey. "Using social media to predict air pollution during California wildfires." Proceedings of the 9th International Conference on Social Media and Society. 2018.
11. Gurajala, Supraja, Suresh Dhaniyala, and Jeanna N. Matthews. "Understanding public response to air quality using tweet analysis." Social Media+ Society 5.3 (2019): 2056305119867656.
12. J. Roesslein, "Tweepy: Twitter for Python!", 2020. Available at: https://github.com/tweepy/tweepy.
13. Twitter, "Standard v1.1", Available at: https://developer.twitter.com/en/docs/twitter-api/v1.
14. Almaqbali, Iqtibas Salim Hilal, et al. "Web Scrapping: Data Extraction from Websites." Journal of Student Research (2019).
15. Karn, Sanjeev Kumar, et al. "News Article Teaser Tweets and How to Generate Them." arXiv preprint arXiv:1807.11535 (2018).
16. I. Trajkovski, "How does TIME.mk work?", Time.mk, 2008. Available at: https://time.mk/info/site
17. I. Trajkovski, Time.rs, 2008. Available at: https://time.rs/

18. Ministry of environment and physical planning – Republic of North Macedonia, “Air Quality Portal”. Available at: https://air.moepp.gov.mk/?page_id=175
19. Republic of Serbia, Open Data Portal, “Air quality - unverified real-time clock data”. Available at: https://data.gov.rs/sr/datasets/kvalitet-vazduha/
20. Environmental Protection Agency of Montenegro, “Measurement data archive”. Available at: http://www.epa.org.me/vazduh/arhiv/7
21. Discomap EEA, “Download of air quality data”. Available at: https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm
22. Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the international AAAI conference on web and social media. Vol. 8. No. 1. 2014.
23. Online Doc Translator, 2021. Available at: https://www.onlinedoctranslator.com/en/
24. Minitab, “Interpret all statistics and graphs for Cross Correlation”, 2022. Available at: https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/cross-correlation/interpret-the-results/all-statistics-and-graphs/
25. Dean, Roger T., and William Dunsmuir. "Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models." Behavior research methods 48.2 (2016): 783-802.
26. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3 Jan (2003): 993-1022.
27. Yin, Jianhua, and Jianyong Wang. "A dirichlet multinomial mixture model-based approach for short text clustering." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.
28. Korzycki, Michał, Izabela Gatkowska, and Wiesław Lubaszewski. "Can the human association norm evaluate machine-made association lists?." Cognitive Approach to Natural Language Processing. Elsevier, 2017. 21-40.
29. Syed, Shaheen, and Marco Spruit. "Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation." 2017 IEEE International conference on data science and advanced analytics (DSAA). IEEE, 2017.
30. Sharifi, Zahra, and Sajjad Shokouhyar. "Promoting consumer's attitude toward refurbished mobile phones: A social media analytics approach." Resources, Conservation and Recycling 167 (2021): 105398.
31. Qiang, Jipeng, et al. "Short text topic modeling techniques, applications, and performance: a survey." IEEE Transactions on Knowledge and Data Engineering (2020).
32. G. Kanevce, A. Dedinec, V. Taseska-Gjorgievska, A.Dedinec. “Transport in Skopje – realities and challenges, Path to green transport”, Third Biennial Update Report on Climate Change, 2017. Available at: https://api.klimatskipromeni.mk/data/rest/file/download/71dea57f28b54b8c5f35e41a364a586d58c97edef47aacf36268b5d4296667ec.pdf
33. https://monkeylearn.com/sentiment-analysis/

34. https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining
35. https://medium.com/@piocalderon/vader-sentiment-analysis-explained-f1c4f9101cd9
36. https://towardsdatascience.com/short-text-topic-modelling-lda-vs-gsdmm-20f1db742e14
37. https://monkeylearn.com/blog/introduction-to-topic-modeling/
38. https://pub.towardsai.net/tweet-topic-modeling-part-3-using-short-text-topic-modeling-on-tweets-bc969a827fef
39. https://towardsdatascience.com/short-text-topic-modeling-70e50a57c883
40. https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05
41. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
42. https://towardsdatascience.com/understanding-topic-coherence-measures-4aa41339634c