Класификација и сентимент анализа на COVID-19 твитови поврзани со лажни и вистинити информации

Мила Куч, 171074

Факултет за информатички науки и компјутерско инженерство, Скопје
Проектна задача по предметот Обработка на природните јазици

Февруари, 2021 година

Абстракт **-** Со појавата на пандемијата COVID-19, животот на луѓето во целиот свет ненадејно и драстично се смени, и истото влијаеше за тоа сите луѓе да внимаваат на квалитетот на информациите кои ги добиваат. Целта на овој проект е да се долови важноста и разликата помеѓу вистинските и лажните информации, со помош на NLP анализа на постови од социјалната мрежа Твитер. Дополнително опфаќа и: анализа и соодветни визуелизации на овие податоци, сентимент анализа на истите, како и модел на длабоко учење за соодветна класификација на твитови.

I. Вовед

Во времиња на светска пандемија, информациите и податоците станаа едни од најважните нешта кои во поголемата слика треба да се земат во предвид. Анализата на тие податоци и информации со помош на различни методи вклучувајќи го машинското учење, длабокото учење и различни визуелизации, може да ни помогне да извлечеме клучни поени и забелешки.

Но, која е потребата од ова? Според едно истражување од Универзитетот МИТ [1], на социјалната мрежа Твитер, лажните вести имаат дури 70% поголема веројатност да се споделувани отколку реалните. Лажните информации се реална закана во нашето општество.

Доколку ги имаме тие добро засновани заклучоци, може да се олесни процесот на превенција или решавање на проблеми од поголем размер во светот, наспроти нивно зголемување ако не го превентираме верувањето во лажните вести.

Дополнително и неповрзано со вистински и лажни вести, самата анализа на податоците од социјалните мрежи, може да се види какво е расположението на популацијата, или дел од неа, во врска со многу важни теми. Со порастот на користењето на технологијата и интернетот, особено во времиња кога повеќето луѓе седат дома, голем број од нив се информираат онлајн, а исто така постираат за нивните мислења и чувства на социјалните медиуми.

Целта на овој проект е да се доловат двата аспекта - и самото разликување на вистински и лажни вести, но и анализата на тоа како луѓето се чувствуваат, за потенцијално да се испита во кои подрачја од секојдневниот живот тие имаат проблеми.

II. Слични истражувања

Според истражувањето на Limeng Cui и Dongwon Lee од Pennsylvania State University, [2] кои го формирааа CoAID податочното множество (понатаму искористено во овој проект) најдоа некои загрижувачки факти поврзани со лажните информации и COVID-19. Со самото влошување на состојбата во светот, неточните информации исто така се зголемија. Пример за тоа се лажните лекови за COVID-19; во Аризона, човек починал и неговата жена била хоспитализирана откако парот се обидел да земе Хлорокин како превенција против пандемијата. [3] Друг пример е, 77 телефонски кули биле запалени заради теоријата дека 5G мобилните мрежи го шират вирусот. [4]

Друго истражување од страна на University Of South Florida, Tampa, Northern Illinois University и Ajivar LLC [5] кое се фокусираше повеќе на сентимент анализата и класификацијата на реченици според најактуелните теми, покажува дека кај студентите по пандемијата најважни биле темите за здравје и семејство, најчесто кажани во негативен контекст.

Постојат и многу други истражувања во врска со препознавањето на вистински и лажни вести во социјалните медиуми [6][7][8], но сите тие се сведуваат на

едно нешто - колку е важна ваквата анализа со помош на обработката на природните јазици, без разлика на тоа за која тема се користи.

III. Опис на проблемот

Во потрага по податочно множество, потребно е да се има некаква цел и замисла за обработката на тие податоци. Бидејќи темата е поврзана со лажните и вистинските податоци, како платформа што е целосно отворена за сите, има достапно API и е доволно користена и популарна и од обичниот корисник, но и од портали за вести, ја одбрав социјалната мрежа Twitter.

Неколку клучни прашања или задачи на кои треба да се даде одговор се:

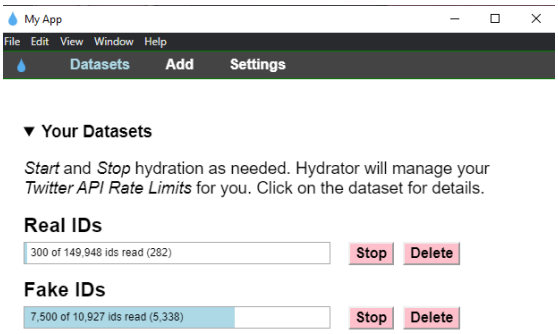
1. Каква била активноста на твитер корисниците во оваа тема во различните месеци? (дали таа се зголемува, намалува,итн.)
2. Кои се најчесто споменуваните зборови во секој месец? - за да ја дознаеме најактуелната тема во тој одреден временски период
3. Кои се најчесто споменувани зборови во лажните и вистинските твитови?
4. На одредено податочно множество да се направи сентимент анализа
5. Да се креира модел на длабоко учење за класифицирање лажни наспроти вистински твитови.

Со одговорите на овие прашања или задачи може да се постигне исцрпна анализа на нашето податочно множество.

IV. Податочно множество

Податочното множество CoAID [9] се состои од 5.216 новости, 296.752 активности на корисници, 958 објави на социјалната платформа Твитер за COVID-19. Самото податочно множество е поделено на неколку дела, т.е. податоци кои се собрани секој втор месец (мај, јули, септември и ноември).
Податочното множество е достапно на Github [10] и од таму може да се преземе.

Во овој проект во предвид ги земав само податоците од твитер, а истите го содржат само ID-то на твитот, но не и текстот и други метаподатоци.
Со таа цел, го користев Hydrator [11], десктоп апликација базирана на Electron, со помош на која може да се преземаат твитови врз основа на нивното ID, а резултатот е во JSON формат. Постапката беше релативно лесна, а пример за истата може да се види на Сл. 1.



Сл.1 Преземање на твитови со помош на Hydrator

Бидејќи потешко е да се работи со JSON податоци, искористив python скрипта со која претвораат во .csv фајлови, а искористив и уште една помошна скрипта за спојување на податоците. [12][13] Со тоа ги добивме финализираните резултати - еден документ со лажните твитови, еден документ со вистинските твитови. (fake.csv и real.csv) [14][15]

V. Анализа на податоците

За манипулација со податоците, се користат неколку библиотеки; NLTK како една од попопуларните и поуспешните во полето на обработката на природните јазици, како и Pandas и Numpy како стандардни библиотеки за работа со податоци.

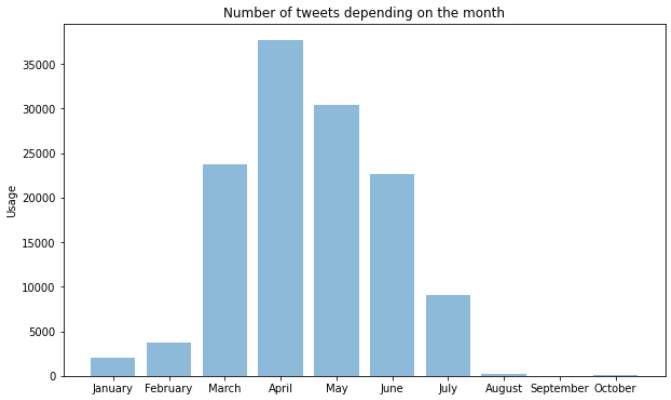
Добиените податочни множества имаат многу колони, а најкористени ќе бидат колоните id, full_text, created_at, lang. Бидејќи работиме стриктно со англиски податоци, тие најпрво се филтрираат по колоната lang. Дополнително, на вистинските информации им се доделува колона за класа 1, а на лажните 0.

Следуваат некои од анализите и заклучоците на самите податоци.

A. Број на твитови според месец

Најпрво, да го одговориме првото прашање - Каква била активноста на твитер корисниците во оваа тема во различните месеци? Ова е важно за да се даде некаква перспектива за тоа какви се податоците во однос на временската рамка, и каква е активноста на корисниците во однос на постирање на COVID-19 твитови.

Една забелешка за самото податочно множество е тоа дека во подоцнежните месеци има значително помалку твитови, што не мора да значи дека тоа е реалноста, туку само помал дел од нив се содржат во склоп на податоците кои ги користиме.
Резултатите можат да се видат на Сл. 2. Во нашето податочно множество се содржат најмногу податоци во месец април, дури 37651 твитови.



Сл.2 Број на твитови во податочното множество во зависност од месецот.

Б. Најфреквентни зборови секој месец

Наоѓањето на најфреквентен збор секој месец е важно за да се определи најважната тема на разговор тогаш.
За оваа анализа потребно е да се создаде вокабулар од зборовите и од вистинските и од лажните вести заедно, како и да е подреден според нивната фреквенција.

Важен дел од овој процес е токенизацијата на зборовите, а овој пат таа беше направена со NLTK Word Tokenize.[16] Дополнително, за да се избегне повторување, зборовите "COVID-19" и "COVID" не беа дел од вокабуларот

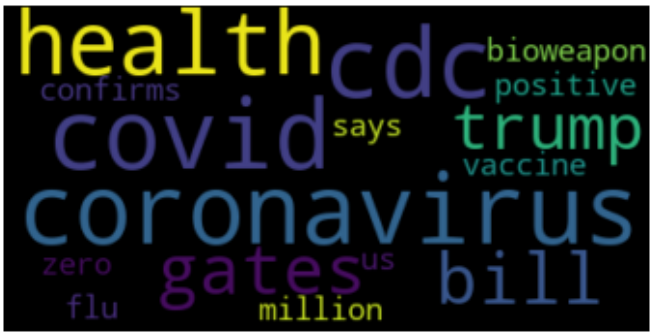
Резултатите можат да се видат на Сл.3, а од сите месеци најмногу пати се појавил зборот people во април, со 3664 пати.

Month	Word	Frequency
January	china	454
February	new	649
March	new	3291
April	people	3664
May	testing	3145
June	people	2315
July	health	1430
August	contact	46
September	vaccine	14
October	trump	79

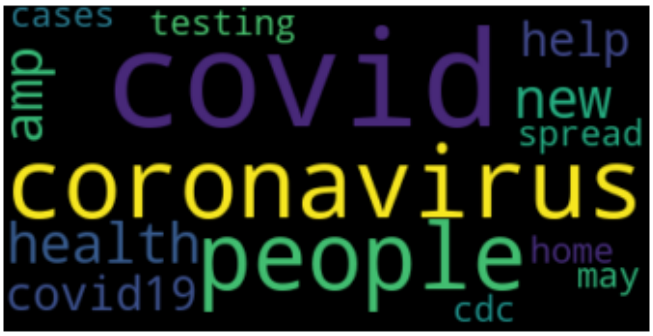
Сл.3 Табела со најпојавуваниот збор во секој месец, и неговата соодветна фреквенција

В. Кои зборови се најзастапени во вистинските и лажните вести

Можеби една од најинтересните и најкорисните анализи е наоѓањето на најкористените зборови во вистинските и лажните твитови, поради тоа што може чисто да се видат кои теми се застапени во секој вид.
Резултатите се исто така интересни, во лажните твитови јасно се застапени повеќе политички теми ("bill", "gates", "trump", "bioweapon"), а во вистинските фокусот е повеќе на здравството и безбедноста на луѓето ("health","people","home"). Истите можат да се видат на Сл.4 и Сл.5.



Сл.4 Најкористени зборови во лажни COVID-19 твитови



Сл.5 Најкористени зборови во вистински COVID-19 твитови

Овие т.н. Word Clouds се генерирани со помош на библиотеката word cloud [17], која овозможува интересни и впечатливи визуелизации на зборови. За создавање на овие визуелизации се искористени 2 сортирани вокабулари слични на тој од претходната задача - едниот со вистински вести, другиот со лажни. Всушност, претставени се првите 20 зборови за секој од нив, со најголемите фреквенции.

VI. Сентимент анализа

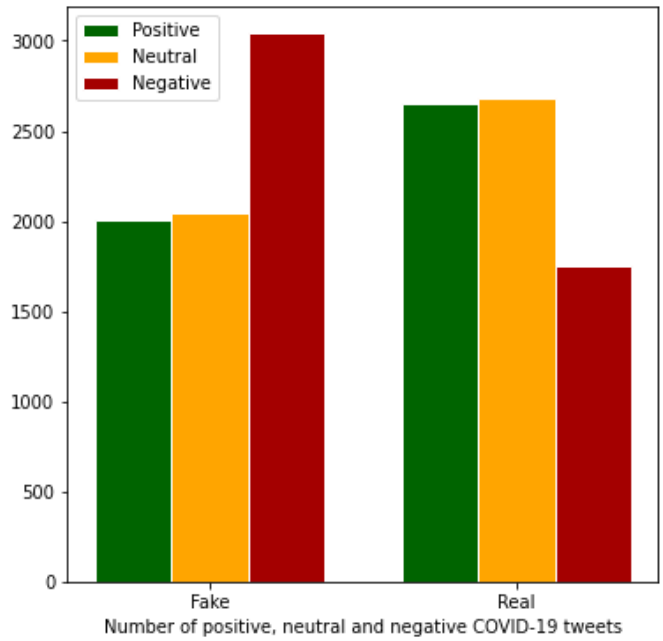
Сентимент анализата е честа NLP задача, која инволвира класифицирање на деловите од текстовите во предефинирани сентименти. Моделите за сентимент анализа се фокусираат на поларитетот (позитивен, негативен, неутрален), но исто така и на самите емоции (лути, среќни, тажни, итн.) [18]

Постојат повеќе пристапи и алатки за сентимент анализа, а еден од најпознатите е VADER (Valence Aware Dictionary and sEntiment Reasoner)[19]. Оваа библиотека е базирана на лексикон и правила за сентимент анализа, и е специфично корисна за сентименти на социјалните медиуми. Библиотеката е достапна и на Github [20]
Овој модел е сензитивен и на поларитет (позитивен, негативен), и на интензитетот на чувството, а финалниот резултат е сума од интензитетот на секој збор во текстот.

На пример, твитот "Contrary to claims in viral social media posts, the novel coronavirus was not man-made nor patented before outbreak" има сентимент резултат од 0.3182 и е лабелиран како позитивен.

Спротивен пример е твитот "Firstpost: Coronavirus Outbreak LIVE Updates: Maharashtra govt to send 100 buses to bring back students stuck in Kota; Rajasthan reports 66 new cases." е лабелиран како негативен, со сентимент резултат од -0.25

На Сл. 6 се прикажани вкупните резултати од сентимент анализата на твитовите, лажни наспроти реални. Може јасно да се забележи од самиот график дека во лажните вести преовладуваат повеќе твитови со негативна, "застрашувачка" конотација, а во реалните твитови тој број е значително помал.



Сл.6 Резултати од сентимент анализата на лажни наспроти реални твитови

VII. Модел за длабоко учење

Целта на длабокото учење е да се истражи како компјутерите можат да ги искористат податоците за да ги развијат репрезентациите соодветни за комплексните интерпретациски задачи. [21]

Во обработката на природните јазици, има повеќе пристапи кога станува збор за длабокото учење. Пред сè, скорешен тренд се зборовните вгнездувања (word embeddings), како и рекурентните невронски мрежи. Во овој дел ќе бидат искомбинирани двата пристапи, за најдобри можни резултати.

Како влез во невронската мрежа назначени се текстовите (full_text колоната), во форма на зборовни вектори, а како излез се лабелите за класата (0 или 1, дали е лажна или вистинска вест соодветно). Тренирачкото, тестирачкото и валидациското множество се поделени во сооднос 80:10:10.

A. BERT Embeddings

Во 2018 година, Google AI тимот направи револуционерна промена во обработката на природните јазици, со воведувањето на BERT (Bidirectional Encoder Representations from Transformers). Заради неговиот високо прагматичен пристап и подобрени перформанси, BERT е една од најсовремените алатки за многу задачи во ова поле.[22]

Во овој проект ја искористив библиотеката Sentence Transformers [23]. Оваа алатка има лесни методи со цел да се добијат соодветни векторски репрезентации на ниво на реченици или параграфи (исто така наречени sentence embeddings). Моделите на кои е базирана оваа мрежа се BERT / RoBERTa / XLM-RoBERTa итн. и се приспособени така што реченициите со слични значења се блиску во векторскиот простор.

Моделот искористен во овој проект е 'paraphrase-distilroberta-base-v1'. Во склоп на овој модел, има и посебен tokenizer, а како резултат од секоја реченица се добива вектор со големина 768. Така, нашите податоци се подготвени за

Б. RNN невронска мрежа со LSTM слој

Иако оваа задача може да се изведе и со BERT модел, решив да го одберам пристапот со класична рекурентна невронска мрежа во Keras со LSTM слој. [24]

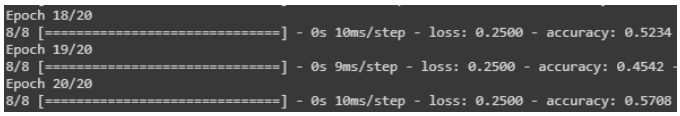
Моделот се состои од влезен слој, LSTM слој, еден скриен слој и излез кој е класата. Преку тренирање на нашите податоци во 20 епохи, со batch_size од 10, се добива прецизност од 0.57 во последниот слој. (Сл.7)

Дополнително, за предвидувањата може да се пресмета RMS како мерка за прецизност. На пример, за едно предвидување изнесува 0.70710, но оваа бројка може да варира во зависност од податоците кои се избрани за тестирање.(Сл.8)

Со ова, ние можеме за кој било твит да ја предвидиме неговата класа, т.е. дали е лажен или вистинит. Овој модел има уште простор за подобрување, но и овие резултати се задоволителни.

```
RMS: 0.7071067811865476
```

Сл.7 RMS - мерка за прецизност



Сл.6 Прецизноста во последните епохи на тренирањето

VIII. Заклучок

Кодот на проектот е поставен на Google Colab [25] Овие модели и пристапи на анализи имаат доста место за подобрување, како и се останато во светот на обработката на природните јазици. Сепак, понекогаш е интересно да погледнеме во светот на податоците и како тие се вклопуваат во реалниот свет, како реални информации, и да извлечеме заклучоци и поуки од истите.

Референци:

[1] Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018;359(6380):1146– 51.

[2]https://arxiv.org/pdf/2006.00885.pdf

[3]https://www.npr.org/sections/coronavirus-live-updates/2020/03/24/820512107/man-dies-woman-hospitalized-after-taking-form-of-chloroquine-to-prevent-covid-19

[4]https://www.businessinsider.com/77-phone-masts-fire-coronavirus-5g-conspiracy-theory-2020-5

[5]Assessing COVID-19 Impacts on College Students via Automated Processing of Free-form Text, arXiv:2012.09369 **[cs.CL]**

[6]N. X. Nyow and H. N. Chua, "Detecting Fake News with Tweets' Properties," 2019 IEEE Conference on Application, Information and Network Security (AINS), Pulau Pinang, Malaysia, 2019, pp. 24-29, doi: 10.1109/AINS47559.2019.8968706.

[7]Global Sentiment Analysis Of COVID-19 Tweets Over Time, arXiv:2010.14234 [cs.CL]

[8]CoronaVis: A Real-time COVID-19 Tweets Data Analyzer and Data Repository, arXiv:2004.13932 **[cs.SI]**

[9] CoAID: COVID-19 Healthcare Misinformation Dataset, Limeng Cui and Dongwon Lee, 2020, 2006.00885, arXiv, cs.SI

[10]https://github.com/cuilimeng/CoAID

[11]https://github.com/DocNow/hydrator

[12]https://drive.google.com/file/d/1Z0xnv55Dizp-E4DxS-cXez5js9ckJopJ/view?usp=sharing

[13]https://drive.google.com/file/d/1bz_AUW3Lco-gPfaskqLRe5Ehwmbs3xlR/view?usp=sharing

[14]https://drive.google.com/file/d/1W3M7XYsxhLuMW29zUeRTR9mq8JwAqx0f/view?usp=sharing

[15]https://drive.google.com/file/d/1pNNEQXpcLlEhGQlAB3nEu0iQvXE8XSBL/view?usp=sharing

[16]https://www.nltk.org/api/nltk.tokenize.html

[17]http://amueller.github.io/word_cloud/

[18]https://monkeylearn.com/sentiment-analysis/

[19]Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[20]https://github.com/cjhutto/vaderSentiment

[21]Richard Socher, Yoshua Bengio, and Christopher D. Manning. 2012. Deep learning for NLP (without magic). In Tutorial Abstracts of ACL 2012 (ACL '12). Association for Computational Linguistics, USA, 5

[22]https://medium.com/analytics-vidhya/bert-word-embeddings-deep-dive-32f6214f02bf

[23]https://github.com/UKPLab/sentence-transformers

[24]https://keras.io/api/layers/recurrent_layers/lstm/

[25]https://colab.research.google.com/drive/1vPxKE7k6KwxhkO92V5eNMU3xXU203Ben?usp=sharing