UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

NETWORK SCIENCE

FAKE NEWS DETECTION

Student	Badge Number
Martin Collado	2039907
Sandra Martinez	2043786
Asmaa Mirkhan	2054616

Professor: Tomaso Erseghe

Academic year 2021-2022

Subtract 3

Introduction 4

Methods 5

Techniques 5

Data Gathering 5

Data Cleaning 5

Classification Models 5

Deployment 5

Datasets 5

Type of Analysis 6

Sentiment Analysis 6

Semantic Analysis 6

Hashtag Analysis 7

Mention and Retweet Analysis 7

Retweet and Favorite Analysis 7

Location Analysis 7

Popularity Analysis 7

Network Science Analysis 7

Results 8

Data Gathering 8

Sources of data 8

Accessing the data 9

Cleaning the data 10

Data Analysis 13

Polarity Analysis 13

Objectivity Analysis 14

Semantic analysis 15

Word Number Analysis 15

N-Gram Analysis 15

Hashtag analysis 20

Mention and retweet analysis 22

Retweet and favourite analysis 25

Location analysis 27

Popularity analysis 28

Network Analysis 32

Accounts Network Analysis 32

Hashtag Network Analysis 35

Classification Models 41

Machine Learning Based Solution 41

Deep Learning Based Solution 42

Deployment 44

Contributions 47

Conclusions 48

References 49

Subtract

In the wake of the American Presidential Election in 2016, the spread of fake news has been a subject of increased discussion which has increased more and more during Covid-19 pandemic. In addition, the wide spreading of social media sites increases the importance of the issue and the large number of news and the dangers posed by fake news have revealed the importance of developing a system to detect the fake news automatically. News on uncontrolled agency sites may mislead people, spread rumors quickly and cause diplomatic - and health related - crises by publishing wrong information. Many models have been proposed for the issue of fake news detection in other articles by applying different methods. In this research, by reviewing news in the dataset we used we first extracted a list of features according to network science basic concepts then we correlated these features and finally we gave these features to various machine learning and deep learning algorithms and compared the accuracy rates. By realizing the importance of the relation between the extracted features from news, our research shows the features of graphs in both fake and real news. In conclusion, by understanding the seriousness of the threats that the fake news will cause, a model with an online website that helps to reduce the dangers is proposed.

1. Introduction

In recent years, social media sites have become widespread and popular and the number of people using it has been large and continues to grow. People now prefer news pages on social media sites to follow news and those pages have become the only news source for a large number of people. As a result, it has become easy to spread fake news and rumors and maliciously mislead people. In order to avoid those dangers, systems that automatically detect fake news are needed. The importance of this issue has increased after the elections held in the USA in 2016 and solving the problem has become a serious challenge [1]. In addition Covid-19 pandemic in 2020 showed us again the importance of this field and the necessity of creating automatic systems that can distinguish the fake news and notify the user about them [2]. The difficulty with this is that fake news has no specific characteristics, well-written news can be believable even if it is shown to people, so it is a serious challenge for a machine to detect fake news.

In this research, “Fake News” is news tagged as false after manual verification, “Real News” is news tagged as true after manual verification

A lot of models have been proposed to solve this problem, some of them are completely based on machine learning, some of them combine machine learning and human-factor to give the output [3]. On the other hand, some proposed models classify news based only on the content of the news [4], while some models only classify the social signals coming to the news [5]. Lastly, some models also categorize by both analyzing content and evaluating social signals as well. Since different methods yield different results, we got inspired from all of these approaches and combined them with network science concepts and reported our results. In addition we deployed our model on a website so the users can use it to predict whether the news is real or fake.

2. Methods

In this part of the report we introduce the different techniques and analysis we have worked with in the project and the datasets we have used to perform the results.

2.1. Techniques

Data Gathering

First, using data gathering techniques, we process the original dataset to obtain more information and relevant features from the tweets so that we can then carry out our analysis.

Data Cleaning

After gathering the data we started a cleaning process in which we divided the tweets depending on if they were retweets or real tweets, so each class can be managed in a dedicated way. We also removed the links from tweets by searching about 'http\S+' and 'www\S+' patterns because the occurrence of links in plain texts can produce unpredictable results in text processing algorithms.

Classification Models

We used various algorithms of machine learning and deep learning and got an accuracy of 88% for the machine learning based solution and an accuracy of 95% for the deep learning based solution.

Deployment

We deployed our model by pushing our model to GitHub with related inference scripts and deployed it using Heroku.

2.2. Datasets

In this project we have been working with a dataset collected by LCS2 research group in IIIT-Delhi university, which is divided in two csv files corresponding each of them to a class of data:

- *Genuine.csv*: A dataset of 2001 real news tweets.

- *Fake.csv*: A dataset of 2001 fake news tweets.

2.3. Type of Analysis

We analyzed the dataset according to the followings:

- Sentiment analysis
- Semantic analysis
- Hashtag analysis
- Mention, retweet analysis
- Retweet and favorite analysis
- Location analysis
- Popularity analysis
- Network Science Analysis

Sentiment Analysis

We investigated the polarity and objectivity in each tweet and compared the rates of polarity and objectivity between real and fake news. We mean by polarity a specific score in the range [-1,1] representing the degree in which a sentence's sentiment is considered as negative (-1), neutral (0) or positive (1). While objectivity, a specific score in the range [0,1] representing the degree in which a sentence's objectivity is considered as objective (0), neutral (0.5) or subjective (1).

Semantic Analysis

We analyzed the texts according to the words and the structure of their sentences. We did lemmatization and then, n-gram analysis and reported results about the most frequented words and phrases and average number of words in each tweet.

- Lemmatization is looking beyond word reduction and considering a language's full vocabulary to apply a morphological analysis to words, aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

- N-grams are contiguous sequences of n-items in a sentence. N can be 1, 2 or any other positive integers.

Hashtag Analysis

We searched about hashtags related signals like the most repeated hashtags in both real and fake news and average number of hashtags in each tweet.

Mention and Retweet Analysis

We searched about mention related signals like the most repeated mentions in both real and fake news, average number of mentions in each tweet and case of retweet and the relation between it and the reliability of the tweet.

Retweet and Favorite Analysis

We searched about the average number of RTs and favorites for each class, taking into account the values obtained in these fields for each of the tweets and we compared the amount of tweets that are below or over the average.

Location Analysis

We searched about the most meaningful locations from where the tweets were created in both real and fake news and the distribution of tweets over locations.

Popularity Analysis

We searched about the relationship between the frequency of appearance of users in the dataset and their popularity in twitter considering their followers.

Network Science Analysis

We researched about advanced concepts of network science Betweenness Centrality, Clustering Coefficient and Community Detection and applied them on our datasets to get further insights about the subject.

3. Results

This part of the report describes all the work carried out on the project, explaining in detail the analyses and processes that have been carried out and the results obtained.

3.1. Data Gathering

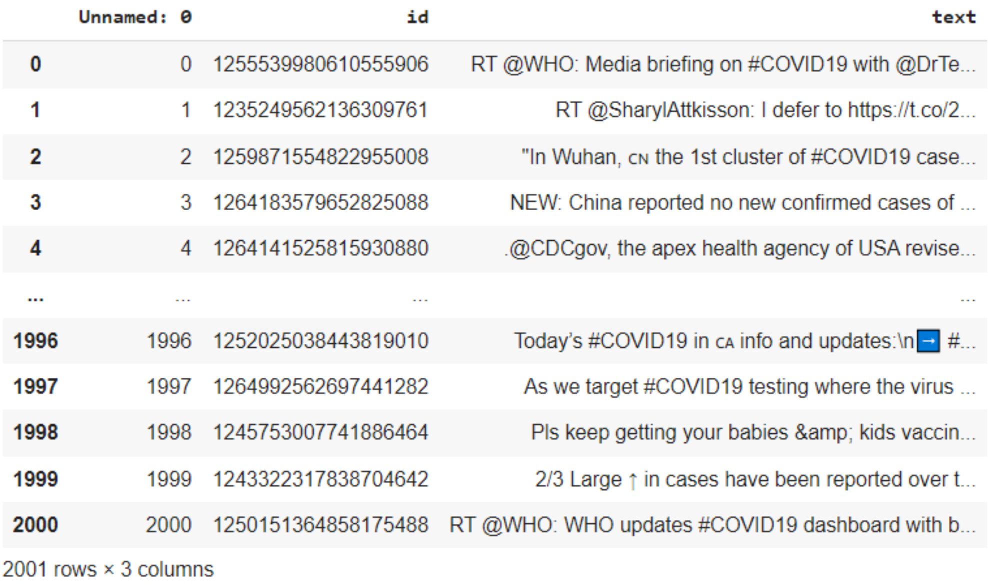
The aim of this section is to explain the data gathering process carried out to obtain relevant and manageable data from raw data sources.

Sources of data

For this Project, we used datasets retrieved by “Laboratory for Computation Social Systems (LCS2) research group” [6], led by Dr. Tanmoy Chakraborty and Dr. Md. Shad Akhtar at IIIT-Delhi university. In march 2021 they developed a project [7] analyzing fake news in Twitter and two datasets were created:

- Genuine.csv: A dataset of 2001 tweets containing verified news content.
- Fake.csv: A dataset of 2001 tweets containing fake news content.

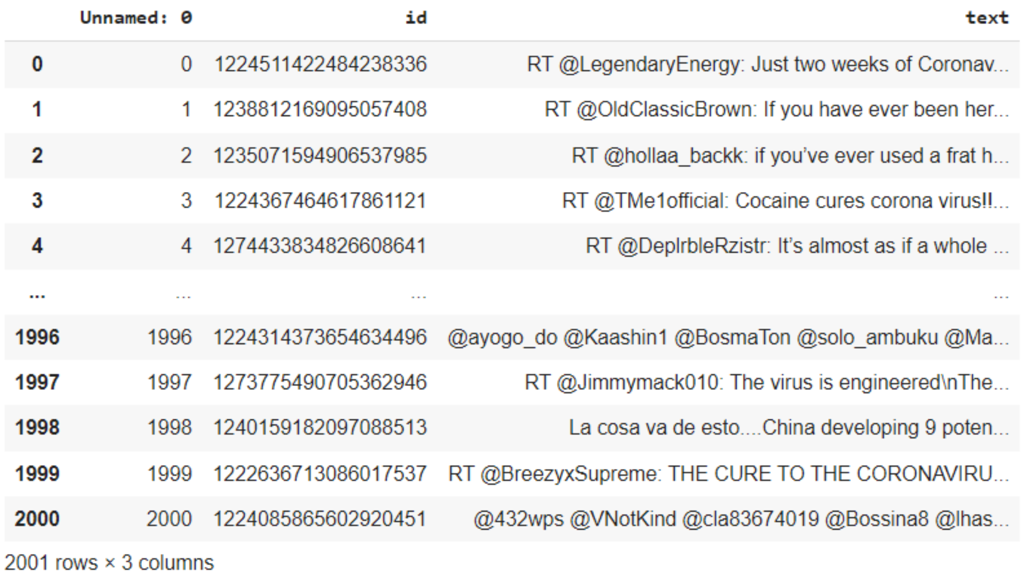
Both datasets have a list with the ids of the tweets and the text that forms the content of each of them. We concluded that this is not enough information for making a complete analysis, so we have created new datasets, using tweets’ ids to retrieve more information.



	Unnamed: 0	id	text
0	0	1255539980610555906	RT @WHO: Media briefing on #COVID19 with @DrTe...
1	1	1235249562136309761	RT @SharylAttkisson: I defer to https://t.co/2...
2	2	1259871554822955008	"In Wuhan, CN the 1st cluster of #COVID19 case...
3	3	1264183579652825088	NEW: China reported no new confirmed cases of ...
4	4	1264141525815930880	.@CDCgov, the apex health agency of USA revise...
...	...	...	...
1996	1996	1252025038443819010	Today’s #COVID19 in CA info and updates:\n➡ #...
1997	1997	1264992562697441282	As we target #COVID19 testing where the virus ...
1998	1998	1245753007741886464	Pls keep getting your babies &amp; kids vaccin...
1999	1999	1243322317838704642	2/3 Large ↑ in cases have been reported over t...
2000	2000	1250151364858175488	RT @WHO: WHO updates #COVID19 dashboard with b...

2001 rows × 3 columns

Figure 1. Original real news dataset



	Unnamed: 0	id	text
0	0	1224511422484238336	RT @LegendaryEnergy: Just two weeks of Coronav...
1	1	1238812169095057408	RT @OldClassicBrown: If you have ever been her...
2	2	1235071594906537985	RT @hollaa_backk: if you’ve ever used a frat h...
3	3	1224367464617861121	RT @TMe1official: Cocaine cures corona virus!!...
4	4	1274433834826608641	RT @DeplrbleRzistr: It’s almost as if a whole ...
...	...	...	...
1996	1996	1224314373654634496	@ayogo_do @Kaashin1 @BosmaTon @solo_ambuku @Ma...
1997	1997	1273775490705362946	RT @Jimmymack010: The virus is engineered\nThe...
1998	1998	1240159182097088513	La cosa va de esto....China developing 9 poten...
1999	1999	1222636713086017537	RT @BreezyxSupreme: THE CURE TO THE CORONAVIRU...
2000	2000	1224085865602920451	@432wps @VNotKind @cla83674019 @Bossina8 @lhas...

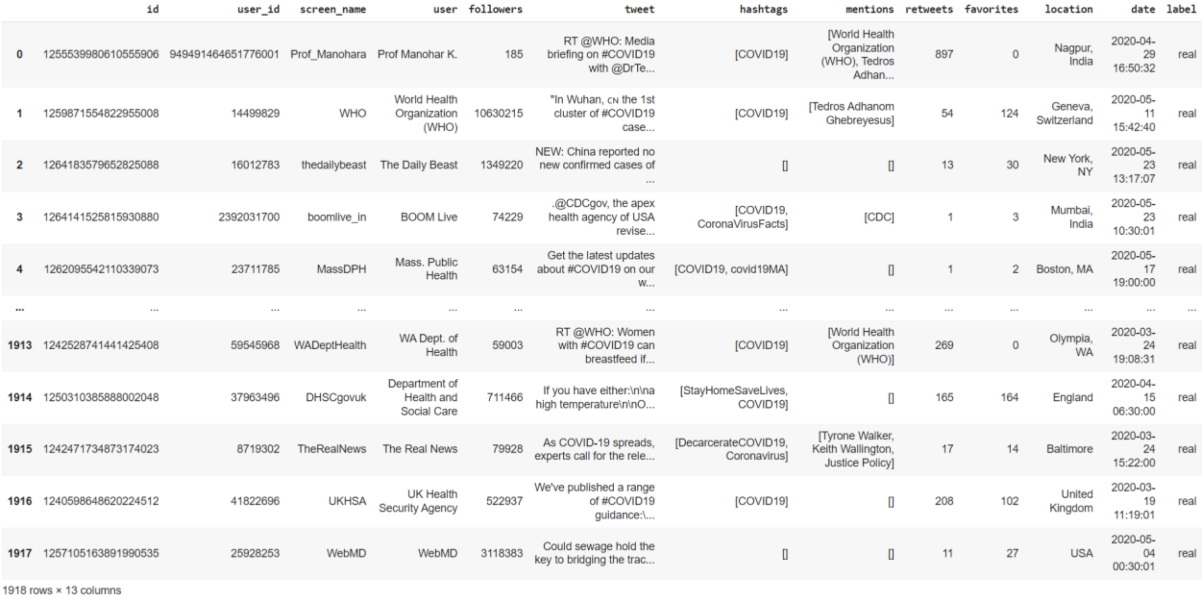
2001 rows × 3 columns

Figure 2. Original fake news dataset

Accessing the data

Data gathering was done using Python’s library called **Tweepy** and the **Twitter API** with a developer account to access the data. Considering the id of each tweet as a starting point, we performed a search in the API from which we obtained the complete status object of the tweet. Once this data was retrieved, we selected the most relevant information for our analysis and created two data frames (one for each dataset) based on this content.

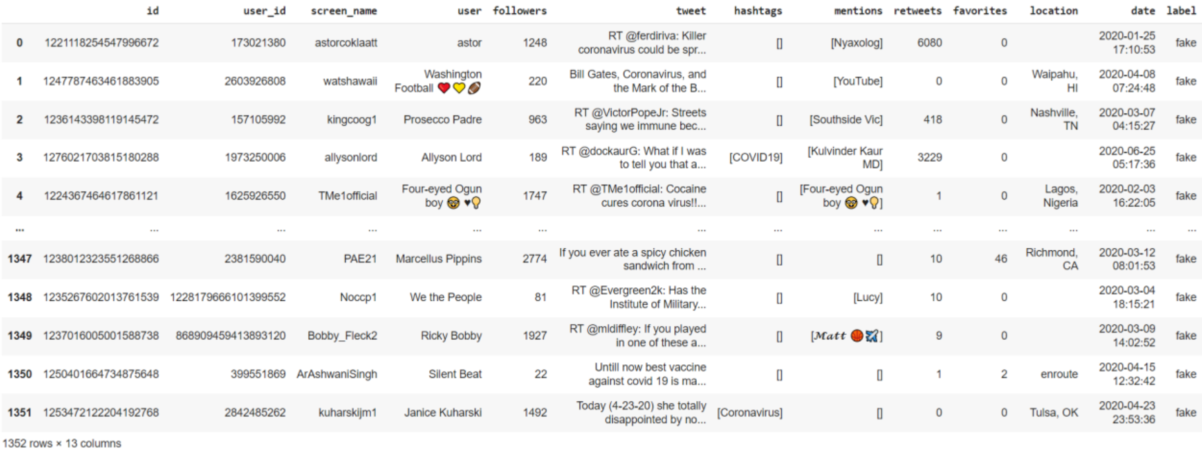
The parameters more relevant of the tweets for our analysis are the id, user id, screen_name, name, its followers, tweet content, hashtags, retweets, favorites, mentions location and date of creation. After collecting all this data in the data frame, a new parameter was added, labeling the data into “fake” or “real” new, for further analysis. At this point the data frames have the following structure.



	id	user_id	screen_name	user	followers	tweet	hashtags	mentions	retweets	favorites	location	date	label
0	1255539980610555906	949491464651776001	Prof_Manohara	Prof Manohar K.	185	RT @WHO: Media briefing on #COVID19 with @DrTe...	[COVID19]	[World Health Organization (WHO), Tedros Adhan...	897	0	Nagpur, India	2020-04-29 16:50:32	real
1	1259871554822955008	14499829	WHO	World Health Organization (WHO)	10630215	"In Wuhan, ᴄɴ the 1st cluster of #COVID19 case...	[COVID19]	[Tedros Adhanom Ghebreyesus]	54	124	Geneva, Switzerland	2020-05-11 15:42:40	real
2	1264183579652825088	16012783	thedailybeast	The Daily Beast	1349220	NEW: China reported no new confirmed cases of ...	[]	[]	13	30	New York, NY	2020-05-23 13:17:07	real
3	1264141525815930880	2392031700	boomlive_in	BOOM Live	74229	.@CDCgov, the apex health agency of USA revise...	[COVID19, CoronaVirusFacts]	[CDC]	1	3	Mumbai, India	2020-05-23 10:30:01	real
4	1262095542110339073	23711785	MassDPH	Mass. Public Health	63154	Get the latest updates about #COVID19 on our w...	[COVID19, covid19MA]	[]	1	2	Boston, MA	2020-05-17 19:00:00	real
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1913	1242528741441425408	59545968	WADeptHealth	WA Dept. of Health	59003	RT @WHO: Women with #COVID19 can breastfeed if...	[COVID19]	[World Health Organization (WHO)]	269	0	Olympia, WA	2020-03-24 19:08:31	real
1914	1250310385888002048	37963496	DHSCgovuk	Department of Health and Social Care	711466	If you have either:\n\na high temperature\n\nO...	[StayHomeSaveLives, COVID19]	[]	165	164	England	2020-04-15 06:30:00	real
1915	1242471734873174023	8719302	TheRealNews	The Real News	79928	As COVID-19 spreads, experts call for the rele...	[DecarcerateCOVID19, Coronavirus]	[Tyrone Walker, Keith Wallington, Justice Policy]	17	14	Baltimore	2020-03-24 15:22:00	real
1916	1240598648620224512	41822696	UKHSA	UK Health Security Agency	522937	We've published a range of #COVID19 guidance:\...	[COVID19]	[]	208	102	United Kingdom	2020-03-19 11:19:01	real
1917	1257105163891990535	25928253	WebMD	WebMD	3118383	Could sewage hold the key to bridging the trac...	[]	[]	11	27	USA	2020-05-04 00:30:01	real

1918 rows × 13 columns

Figure 3. Labeled real news dataset with new features



	id	user_id	screen_name	user	followers	tweet	hashtags	mentions	retweets	favorites	location	date	label
0	1221118254547996672	173021380	astorcoklaatt	astor	1248	RT @ferdiriva: Killer coronavirus could be spr...	[]	[Nyaxolog]	6080	0		2020-01-25 17:10:53	fake
1	1247787463461883905	2603926808	watshawaii	Washington Football ❤💛🏈	220	Bill Gates, Coronavirus, and the Mark of the B...	[]	[YouTube]	0	0	Waipahu, HI	2020-04-08 07:24:48	fake
2	1236143398119145472	157105992	kingcoog1	Prosecco Padre	963	RT @VictorPopeJr: Streets saying we immune bec...	[]	[Southside Vic]	418	0	Nashville, TN	2020-03-07 04:15:27	fake
3	1276021703815180288	1973250006	allysonlord	Allyson Lord	189	RT @dockaurG: What if I was to tell you that a...	[COVID19]	[Kulvinder Kaur MD]	3229	0		2020-06-25 05:17:36	fake
4	1224367464617861121	1625926550	TMe1official	Four-eyed Ogun boy 🤓♥💡	1747	RT @TMe1official: Cocaine cures corona virus!!...	[]	[Four-eyed Ogun boy 🤓♥💡]	1	0	Lagos, Nigeria	2020-02-03 16:22:05	fake
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1347	1238012323551268866	2381590040	PAE21	Marcellus Pippins	2774	If you ever ate a spicy chicken sandwich from ...	[]	[]	10	46	Richmond, CA	2020-03-12 08:01:53	fake
1348	1235267602013761539	1228179666101399552	Noccp1	We the People	81	RT @Evergreen2k: Has the Institute of Military...	[]	[Lucy]	10	0		2020-03-04 18:15:21	fake
1349	1237016005001588738	868909459413893120	Bobby_Fleck2	Ricky Bobby	1927	RT @mldiffley: If you played in one of these a...	[]	[𝓜𝓪𝓽𝓽 🏀✈]	9	0		2020-03-09 14:02:52	fake
1350	1250401664734875648	399551869	ArAshwaniSingh	Silent Beat	22	Untill now best vaccine against covid 19 is ma...	[]	[]	1	2	enroute	2020-04-15 12:32:42	fake
1351	1253472122204192768	2842485262	kuharskijm1	Janice Kuharski	1492	Today (4-23-20) she totally disappointed by no...	[Coronavirus]	[]	0	0	Tulsa, OK	2020-04-23 23:53:36	fake

1352 rows × 13 columns

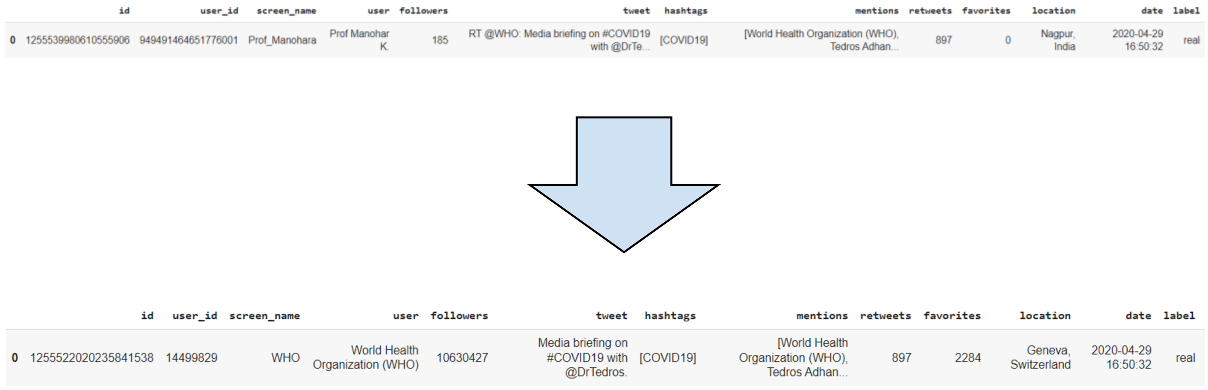
Figure 4. Labeled fake news dataset with new features

Cleaning the data

Before starting with the analysis indeed, it is important to make sure that the data has the best adapted structure for that purpose. Therefore, a data cleaning process has been carried out, consisting of two main steps:

On the one hand, tweets have been divided according to whether they are original tweets or retweets of other tweets. This step is important because RTs have different structure than tweets, and they need to be filtered separately.

Once we divided the data, we cleaned up certain parameters of tweets that turned out to be RTs of others. We considered that the information of the retweeter is not as relevant as the content of the original tweet, so in these cases a process was carried out in which the RT features were replaced by the original ones. Let's take a closer look at an example:



	id	user_id	screen_name	user	followers	tweet	hashtags	mentions	retweets	favorites	location	date	label
0	1255539980610555906	949491464651776001	Prof_Manohara	Prof Manohar K.	185	RT @WHO: Media briefing on #COVID19 with @DrTe...	[COVID19]	[World Health Organization (WHO), Tedros Adhan...	897	0	Nagpur, India	2020-04-29 16:50:32	real

	id	user_id	screen_name	user	followers	tweet	hashtags	mentions	retweets	favorites	location	date	label
0	1255522020235841538	14499829	WHO	World Health Organization (WHO)	10630427	Media briefing on #COVID19 with @DrTedros.	[COVID19]	[World Health Organization (WHO), Tedros Adhan...	897	2284	Geneva, Switzerland	2020-04-29 16:50:32	real

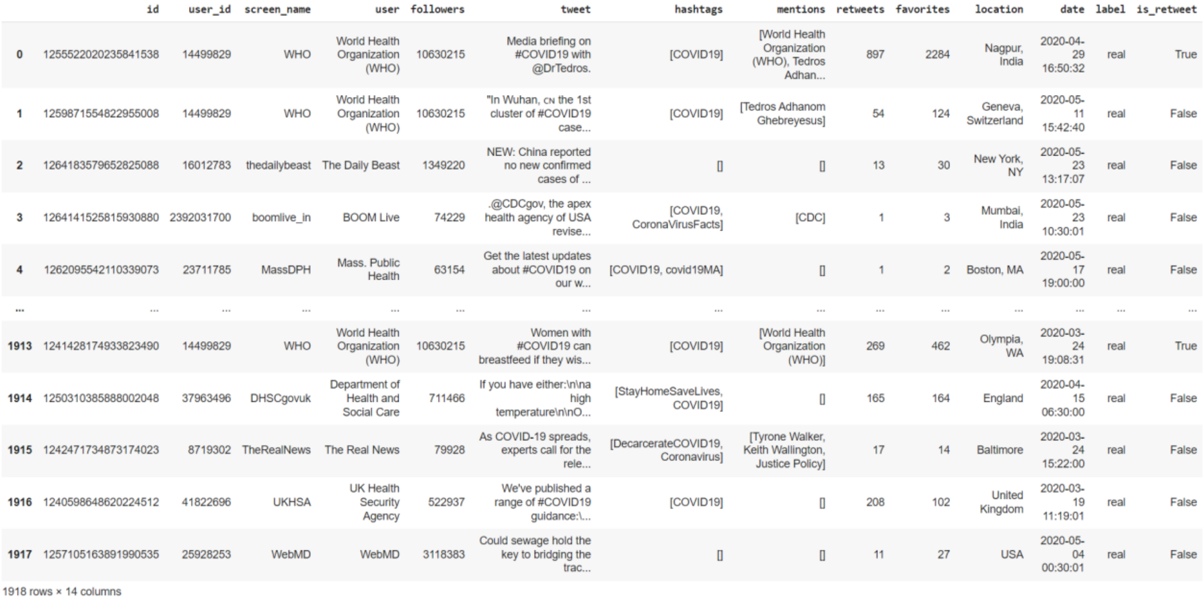
Figure 5. Checking if a sample is a retweet

As we can see above, the tweet recovered from the dataset is a RT that the user "Prof Manohar K." did to a tweet written by the user "World Health Organization (WHO)". For further analysis, we concluded that the original tweet had more information to give, so we substitute the row with the original tweet features.

- Tweet's id changes for the original one.
- User's id, screen_name, name and its followers number now referenced the writer of the tweet, not the retweeter.
- Tweet content has been modified and the section "RT @" has been removed.
- Favorites amount has been updated, as the value is taken from the original tweet.
- Location from where the tweet was sent now indicates the writer's location.

On the other hand, we passed the full data frames through a filter that cleans the text of links and web pages, so that when analysing tweets based on their content, they do not affect the result.

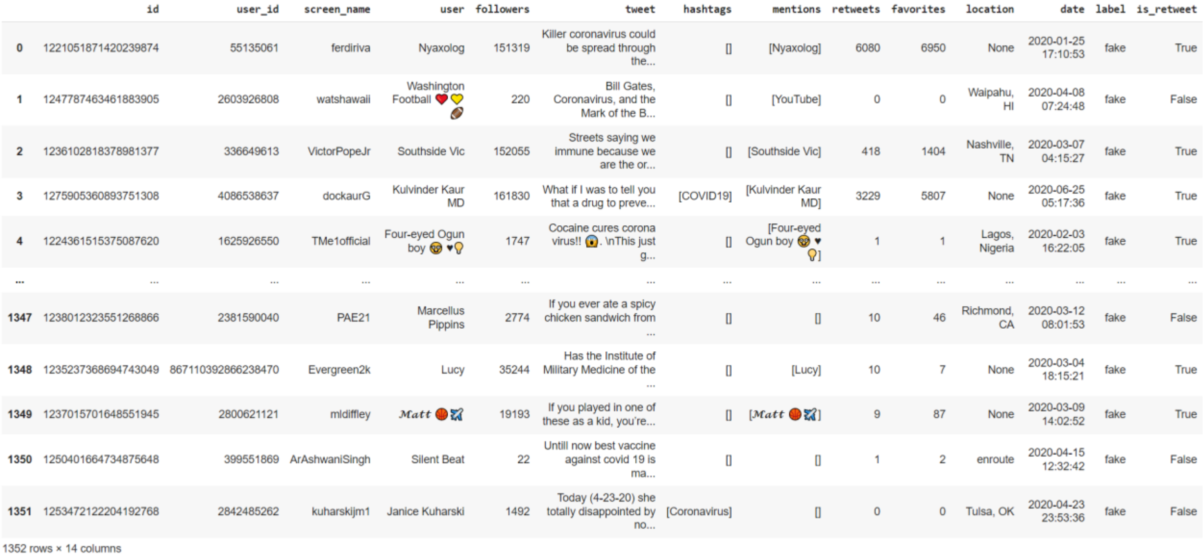
Besides, we updated the location of each tweet which has this field empty to be filled with "None". Here we can see the same data frame as before, but with the data cleaning done:



	id	user_id	screen_name	user	followers	tweet	hashtags	mentions	retweets	favorites	location	date	label	is_retweet
0	1255522020235841538	14499829	WHO	World Health Organization (WHO)	10630215	Media briefing on #COVID19 with @DrTedros.	[COVID19]	[World Health Organization (WHO), Tedros Adhan...	897	2284	Nagpur, India	2020-04-29 16:50:32	real	True
1	1259871554822955008	14499829	WHO	World Health Organization (WHO)	10630215	"In Wuhan, ᴄɴ the 1st cluster of #COVID19 case...	[COVID19]	[Tedros Adhanom Ghebreyesus]	54	124	Geneva, Switzerland	2020-05-11 15:42:40	real	False
2	1264183579652825088	16012783	thedailybeast	The Daily Beast	1349220	NEW: China reported no new confirmed cases of ...	[]	[]	13	30	New York, NY	2020-05-23 13:17:07	real	False
3	1264141525815930880	2392031700	boomlive_in	BOOM Live	74229	.@CDCgov, the apex health agency of USA revise...	[COVID19, CoronaVirusFacts]	[CDC]	1	3	Mumbai, India	2020-05-23 10:30:01	real	False
4	1262095542110339073	23711785	MassDPH	Mass. Public Health	63154	Get the latest updates about #COVID19 on our w...	[COVID19, covid19MA]	[]	1	2	Boston, MA	2020-05-17 19:00:00	real	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1913	1241428174933823490	14499829	WHO	World Health Organization (WHO)	10630215	Women with #COVID19 can breastfeed if they wis...	[COVID19]	[World Health Organization (WHO)]	269	462	Olympia, WA	2020-03-24 19:08:31	real	True
1914	1250310385888002048	37963496	DHSCgovuk	Department of Health and Social Care	711466	If you have either:\n\na high temperature\n\nO...	[StayHomeSaveLives, COVID19]	[]	165	164	England	2020-04-15 06:30:00	real	False
1915	1242471734873174023	8719302	TheRealNews	The Real News	79928	As COVID-19 spreads, experts call for the rele...	[DecarcerateCOVID19, Coronavirus]	[Tyrone Walker, Keith Wallington, Justice Policy]	17	14	Baltimore	2020-03-24 15:22:00	real	False
1916	1240598648620224512	41822696	UKHSA	UK Health Security Agency	522937	We've published a range of #COVID19 guidance:\...	[COVID19]	[]	208	102	United Kingdom	2020-03-19 11:19:01	real	False
1917	1257105163891990535	25928253	WebMD	WebMD	3118383	Could sewage hold the key to bridging the trac...	[]	[]	11	27	USA	2020-05-04 00:30:01	real	False

1918 rows × 14 columns

Figure 6. Complete real news dataset after data cleaning



	id	user_id	screen_name	user	followers	tweet	hashtags	mentions	retweets	favorites	location	date	label	is_retweet
0	1221051871420239874	55135061	ferdiriva	Nyaxolog	151319	Killer coronavirus could be spread through the...	[]	[Nyaxolog]	6080	6950	None	2020-01-25 17:10:53	fake	True
1	1247787463461883905	2603926808	watshawaii	Washington Football ❤️💛🏈	220	Bill Gates, Coronavirus, and the Mark of the B...	[]	[YouTube]	0	0	Waipahu, HI	2020-04-08 07:24:48	fake	False
2	1236102818378981377	336649613	VictorPopeJr	Southside Vic	152055	Streets saying we immune because we are the or...	[]	[Southside Vic]	418	1404	Nashville, TN	2020-03-07 04:15:27	fake	True
3	1275905360893751308	4086538637	dockaurG	Kulvinder Kaur MD	161830	What if I was to tell you that a drug to preve...	[COVID19]	[Kulvinder Kaur MD]	3229	5807	None	2020-06-25 05:17:36	fake	True
4	1224361515375087620	1625926550	TMe1official	Four-eyed Ogun boy 🤓♥💡	1747	Cocaine cures corona virus!! 😱. \nThis just g...	[]	[Four-eyed Ogun boy 🤓♥💡]	1	1	Lagos, Nigeria	2020-02-03 16:22:05	fake	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1347	1238012323551268866	2381590040	PAE21	Marcellus Pippins	2774	If you ever ate a spicy chicken sandwich from ...	[]	[]	10	46	Richmond, CA	2020-03-12 08:01:53	fake	False
1348	1235237368694743049	867110392866238470	Evergreen2k	Lucy	35244	Has the Institute of Military Medicine of the ...	[]	[Lucy]	10	7	None	2020-03-04 18:15:21	fake	True
1349	1237015701648551945	2800621121	mldiffley	Matt 🏀✈️	19193	If you played in one of these as a kid, you're...	[]	[Matt 🏀✈️]	9	87	None	2020-03-09 14:02:52	fake	True
1350	1250401664734875648	399551869	ArAshwaniSingh	Silent Beat	22	Untill now best vaccine against covid 19 is ma...	[]	[]	1	2	enroute	2020-04-15 12:32:42	fake	False
1351	1253472122204192768	2842485262	kuharskijm1	Janice Kuharski	1492	Today (4-23-20) she totally disappointed by no...	[Coronavirus]	[]	0	0	Tulsa, OK	2020-04-23 23:53:36	fake	False

1352 rows × 14 columns

Figure 7. Complete fake news dataset after data cleaning

Up to this point, these data frames were saved in new datasets (complete_fake.csv and complete_genuine.csv), from where the analysis will begin. It is important to note that the generated datasets do not have the same length as the original ones, as some of the tweets that were included at that time are not currently available in the twitter database, either because the source has deleted them or because they have been blocked by twitter due to privacy and authenticity reasons.

- Complete_genuine.csv has **1918** tweets. Almost the full content of the original dataset has been recovered.
- Complete_fake.csv has **1352** tweets.

As we can observe, the real news dataset gathered almost the entire content of the original set, while the fake news dataset has lost more than 30% of its content. It makes sense, since fake news have no credibility and their content is more likely to be blocked or reported in social networks.

3.2. Data Analysis

In this section, our aim is to extract features from the tweets in datasets and to find correlations between the features. For that purpose, we have developed some analysis and created different graphics to show the results.

Polarity Analysis

After applying text cleaning by removing links from tweets we firstly did polarity analysis on both real and fake news. As a result, we noticed that the average positivity of real news is higher than the average positivity of fake news. While the rate of neutral news was higher in fake news than the rate of neutral ones in real news. So, we can say that the real news - in our dataset - tends to be positive and fake ones tend to be neutral.

In order to show the data in more specific sections, we have divided the polarity into 5 levels: very negative, negative, neutral, positive and very positive. From this division we obtain that tweets that have been classified as both negative and positive in both datasets have mostly neutral polarity values (between -0.5 and 0.5) and that a small percentage are evaluated as very negative or very positive.

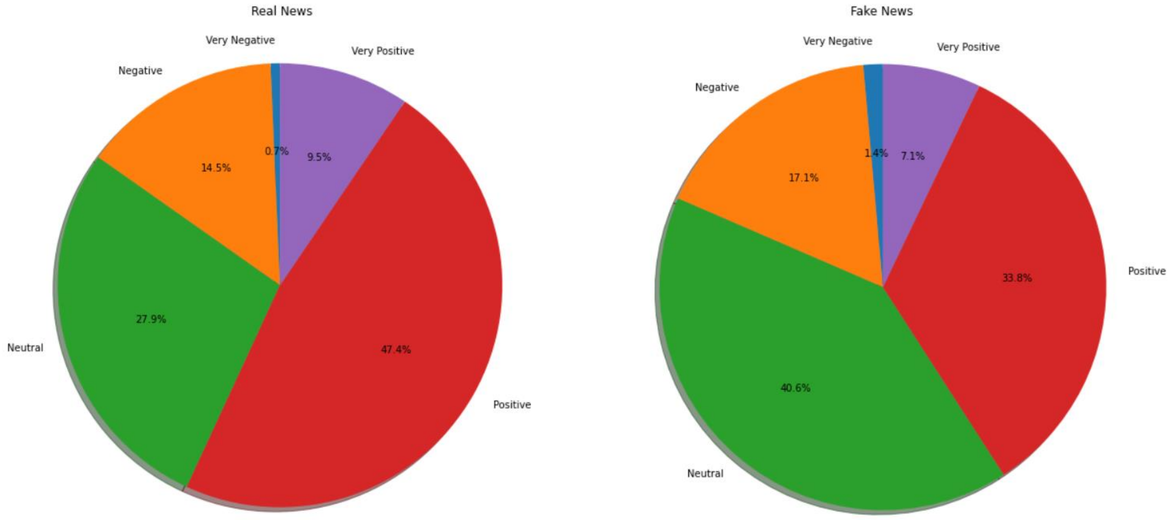


Figure 8. Polarity Analysis

Objectivity Analysis

After applying polarity analysis we did objectivity analysis on both real and fake news. As a consequence , we noticed that there is no apparent difference in objectivity rates between fake and real news.

As in the case of polarity, subjectivity has been divided into five levels, from very objective (values below 0.25) to very subjective (values above 0.75). In this case, it is interesting to note that within both the subjective and objective results, the values are evenly distributed, with some being closer to neutrality and others more extreme.

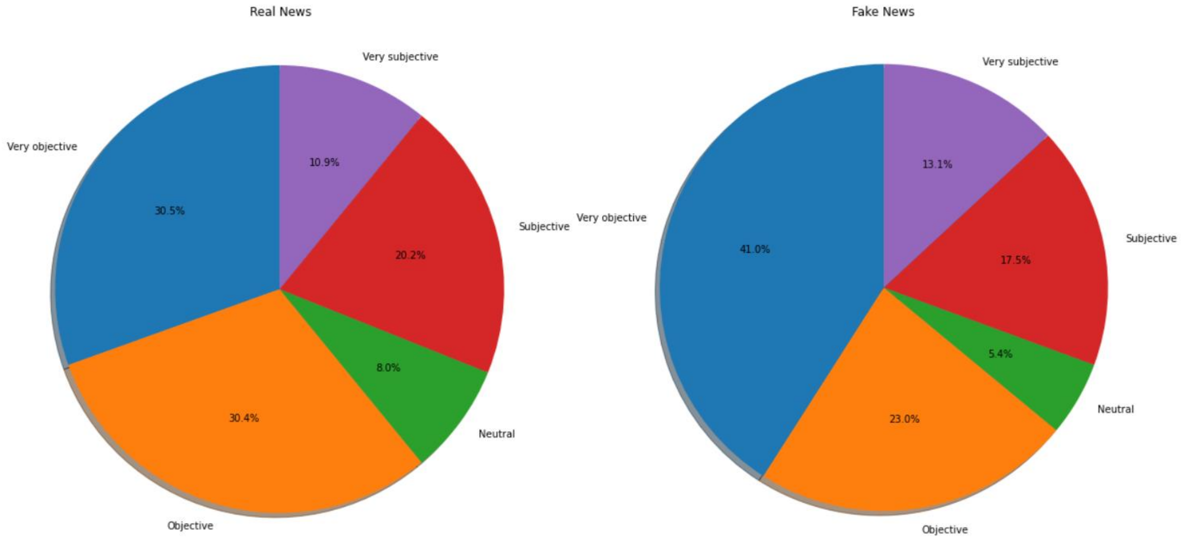


Figure 9. Subjectivity Analysis

Semantic analysis

Word Number Analysis

We started by comparing the average number of words in tweets for both real and fake news. We found that the average number of words in real news is 25.74 word/tweet and the average number of words in fake news is 21.92 word/tweet. So, we interpreted that by saying that the real news tends to have a higher number of words than fake news.

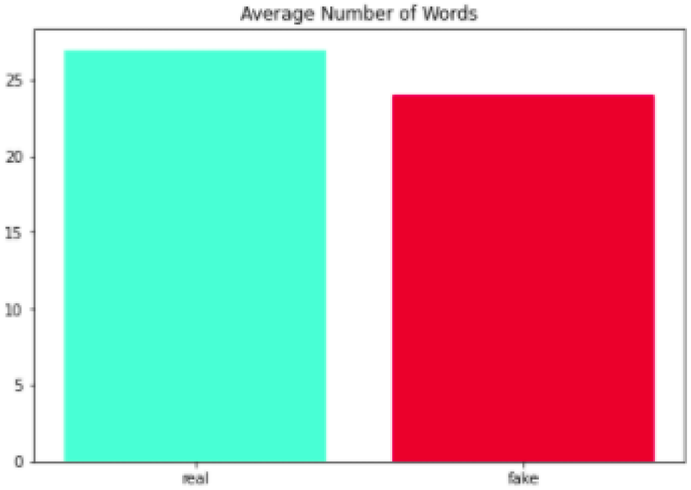


Figure 10. Average number of words per tweet

N-Gram Analysis

Subsequently, we applied lemmatization on texts and then, n-grams analysis on both real and fake news. For n=1 we did not have meaningful results but by increasing n and analyzing phrases we noticed that most frequented phrases in fake news tended to be biased.

Real news results:

1-grams of real news

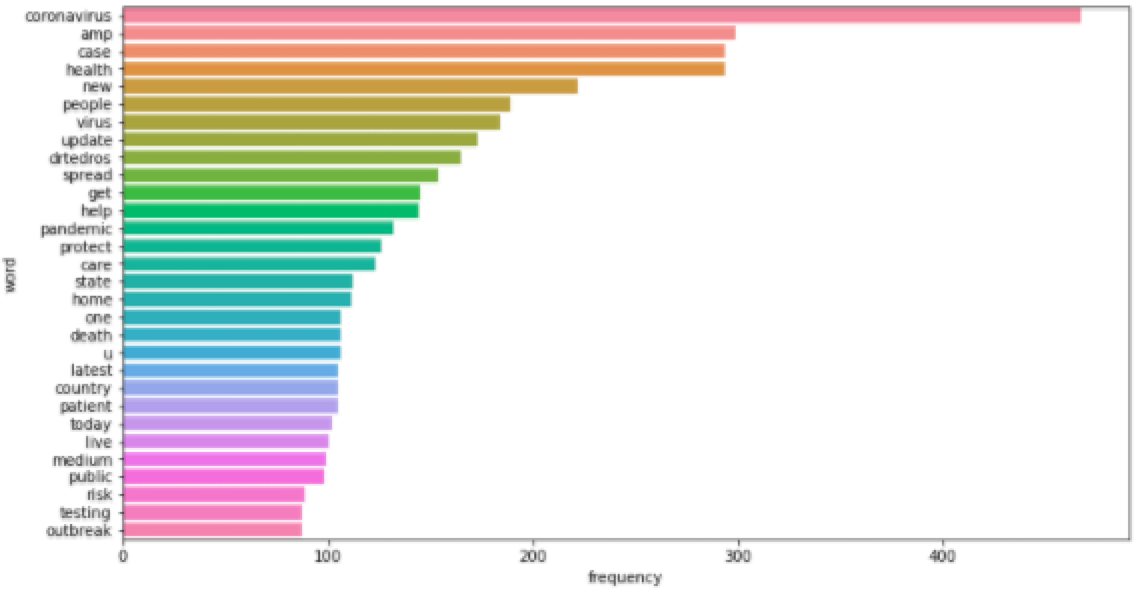


Figure 11. 1-grams visualization of real news

2-gram of real news

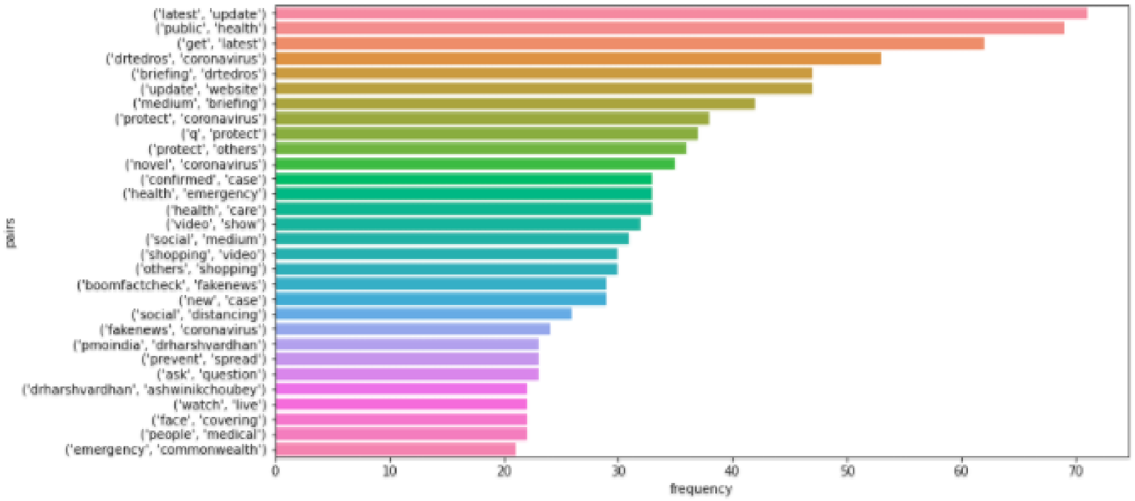


Figure 12. 2-grams visualization of real news

3-gram of real news

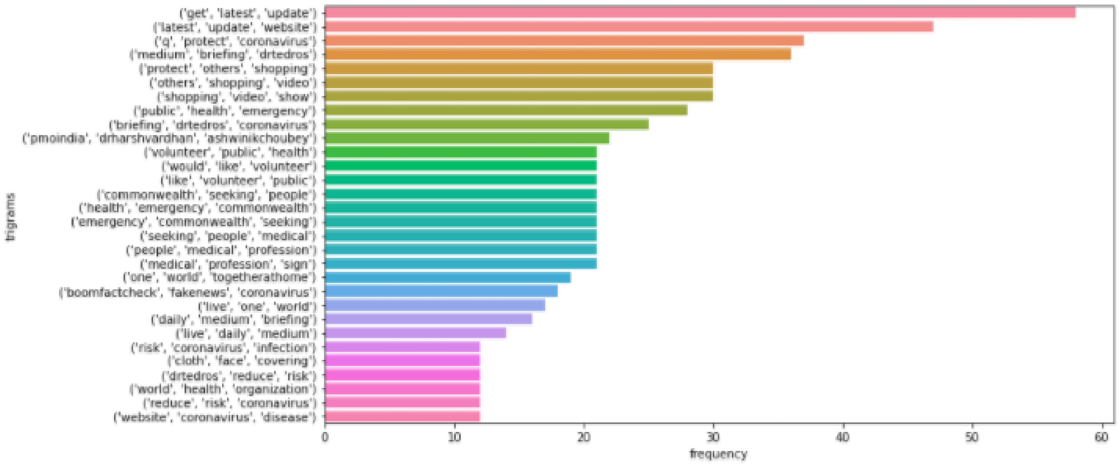


Figure 13. 3-grams visualization of real news

4-gram of real news

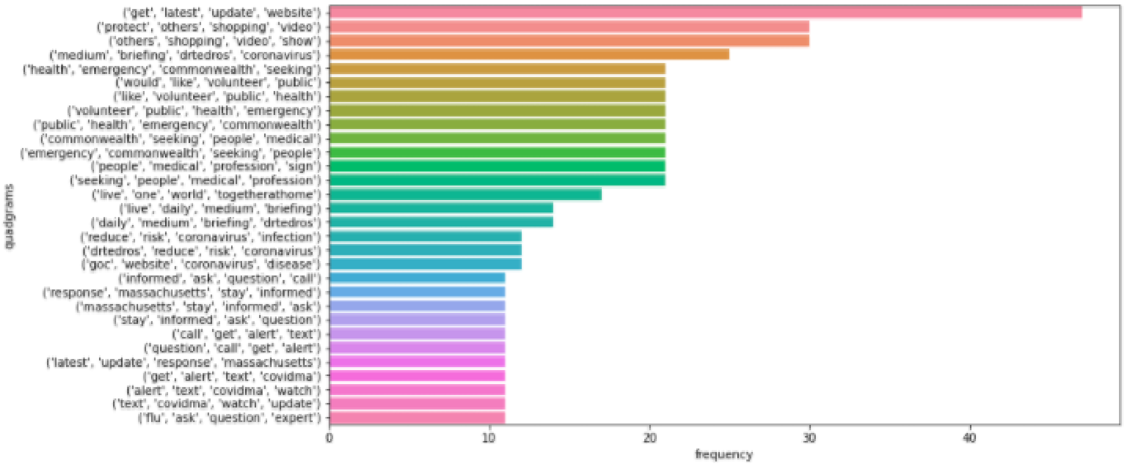


Figure 14. 4-grams visualization of real news

Fake news results:

1-grams of fake news

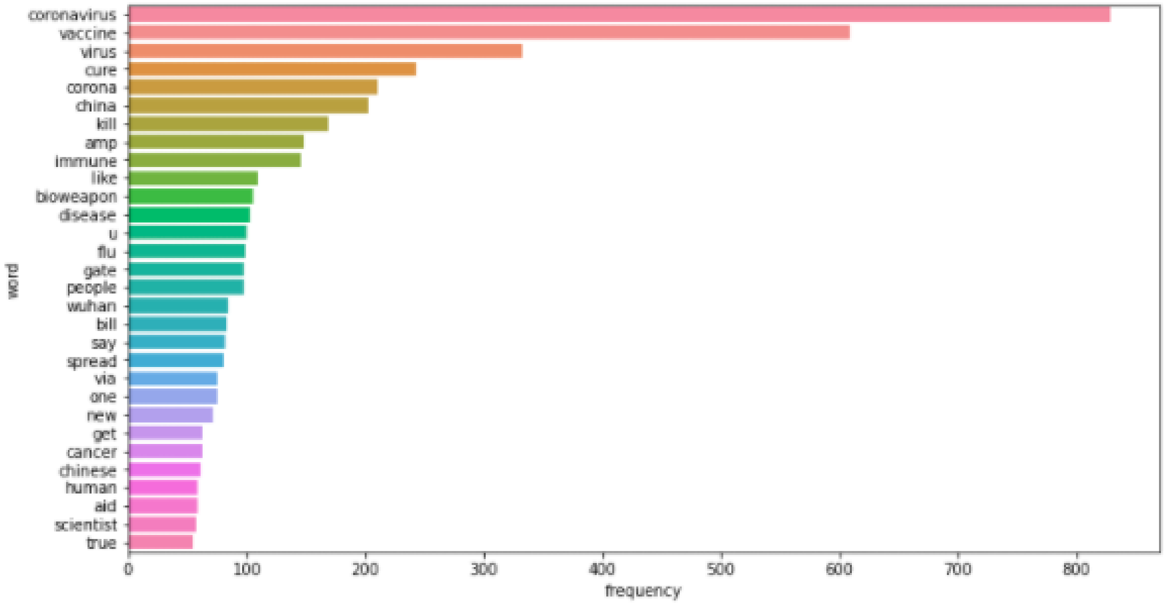


Figure 15. 1-grams visualization of fake news

2-grams of fake news

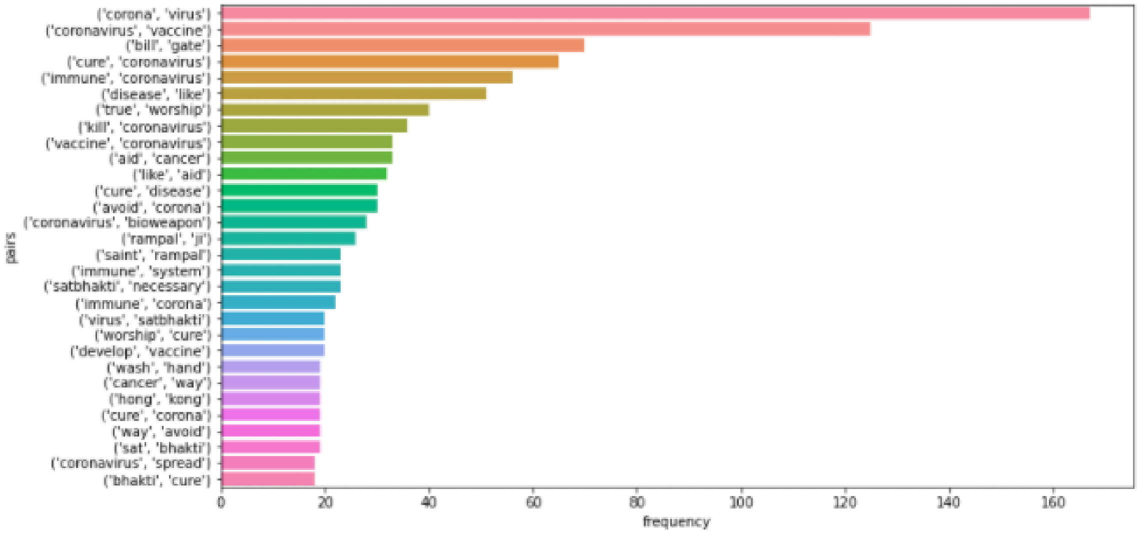


Figure 16. 2-grams visualization of fake news

3-grams of fake news

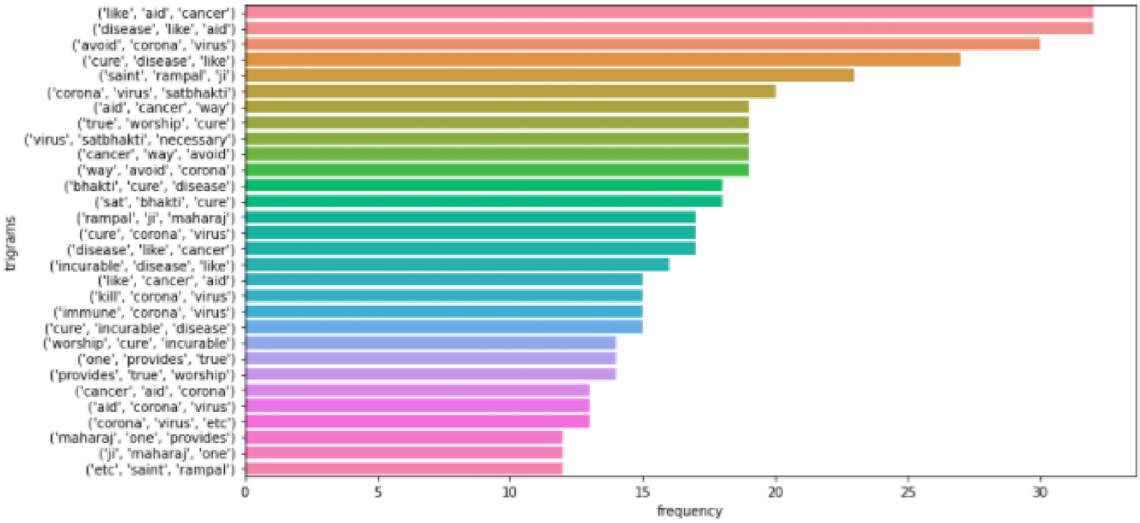


Figure 17. 3-grams visualization of fake news

4-grams of fake news

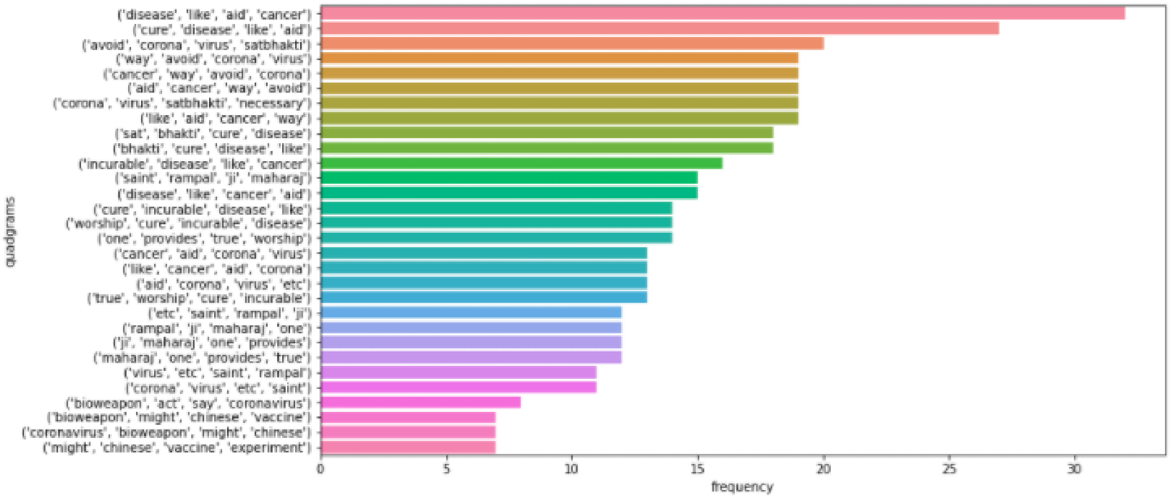
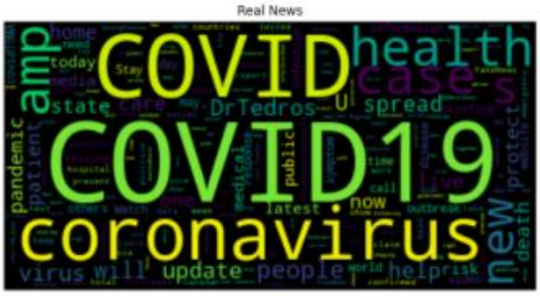


Figure 18. 4-grams visualization of fake news

And for better visualization we feed our data to WordCloud package and got the following results:

1-gram visualization:



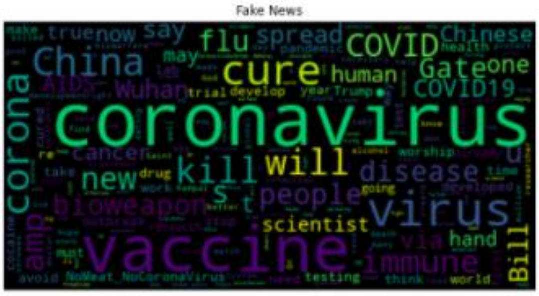


Figure 19. Word cloud visualization of words real and fake news

2-gram visualization:

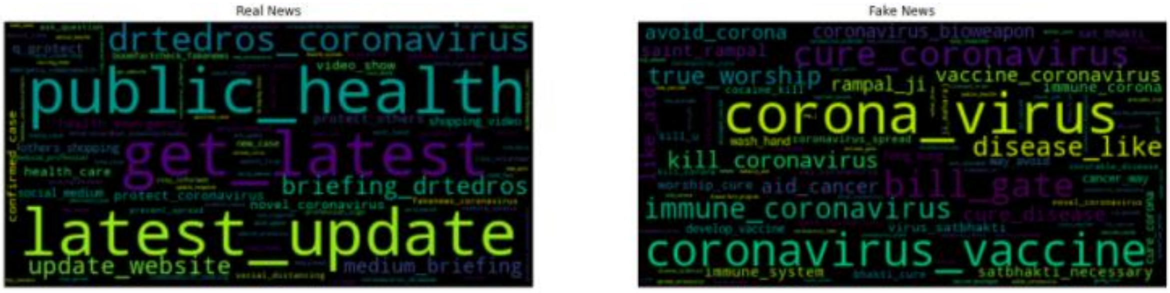


Figure 20. Word cloud visualization of word pairs in real and fake news

Accordingly, we noticed that writers of real news used more scientific and unbiased words like (“health”, “help”, “latest update”, “public health”, “protect others”) and tweets of fake news tended to have biased and slang words like (“kill”, “gate”, “bioweapon”, “chinese”, “bill gates”, “gate foundation”).

Hashtag analysis

As we did over texts, we did on hashtags. In the beginning, we calculated the average number of hashtags in each tweet, then, we calculated the number of tweets that contain hashtags in both real and fake news. Correspondingly, we found that real news tends to have more hashtags than fake news. Subsequently, we did frequency analysis on hashtags too, the results showed us that our observations in text analysis were correct. The hashtags in real news like (“CoronaVirusFacts”, “HealthForAll”, “healthWorkers”, “TogetherAtHome”) were more scientific and unbiased. By contrast with the fake news, they were biased and slang like (“BillGates”, “NoMeat_NoCoronaVirus”, “kill_coronavirus”, “coronaviruschina”, “ChinaVirus”).

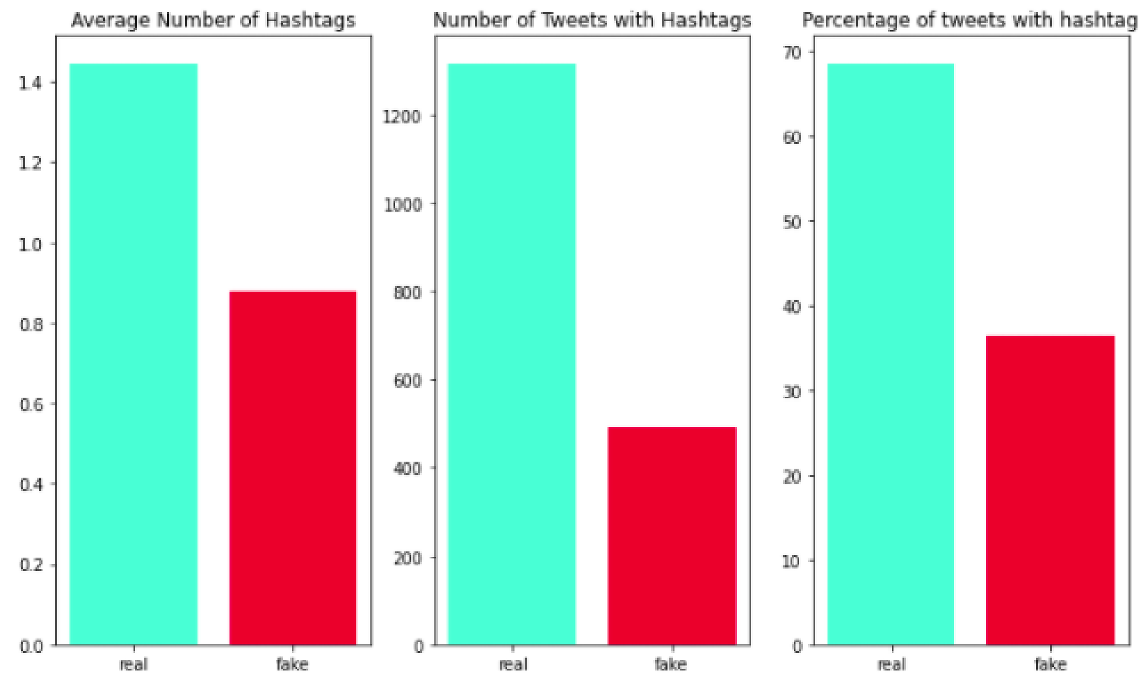


Figure 21. Average number of hashtags in real and fake news

Real news hashtag frequency

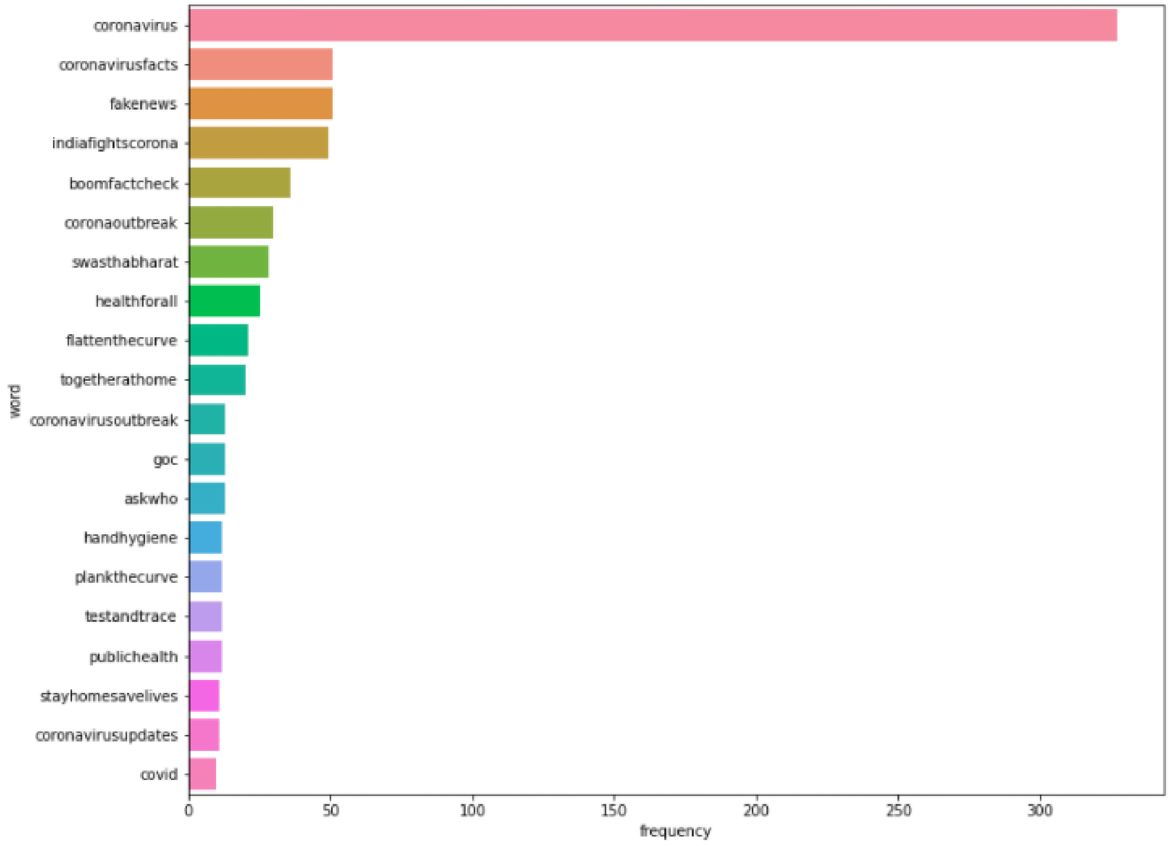


Figure 22. Most repeated hashtags in real news

Fake news hashtag frequency:

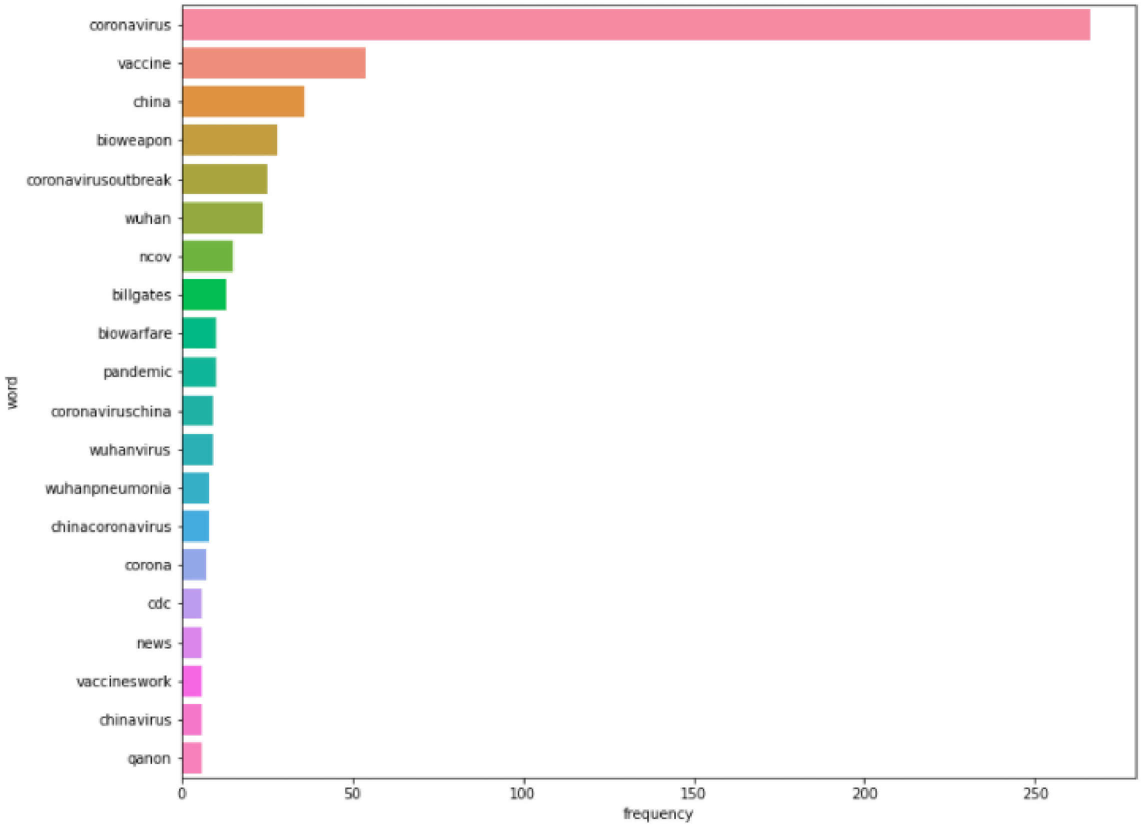


Figure 23. Most repeated hashtags in fake news

WordCloud of hashtags:

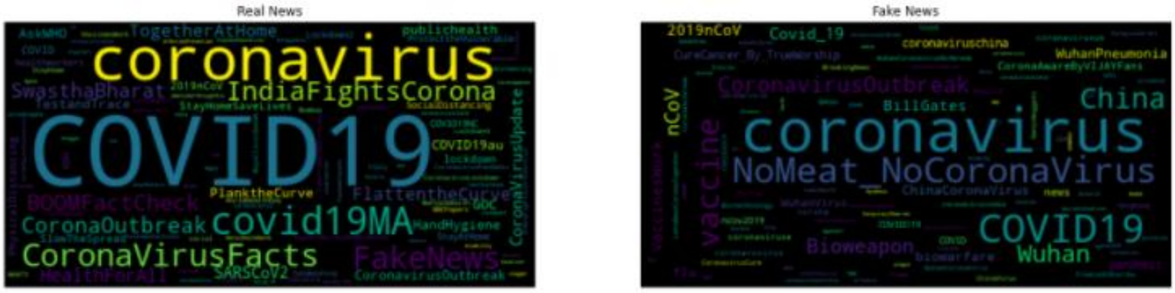


Figure 24. Word cloud of most repeated hashtags in real and fake news

Mention and retweet analysis

Similarly, we calculated the average number of mentions in each tweet, then, we calculated the number of tweets that contain mentions in both real and fake news. Correspondingly, we found that fake news and real news have a very close average of mentions. In addition, we noticed that the mentioned users in real news are reliable resources like Dr. Tedros; the Director-General of the World Health Organization and CDC

Gov; the account of Centers for Disease Control and Prevention. While the most mentioned accounts in fake news were unreliable resources, like youtube, realdonaldtrump. At last, we investigated whether the tweets are retweets or not in both real and fake news. In response, we found that the average number of retweeted fake news is higher than the average number of retweeted real news.

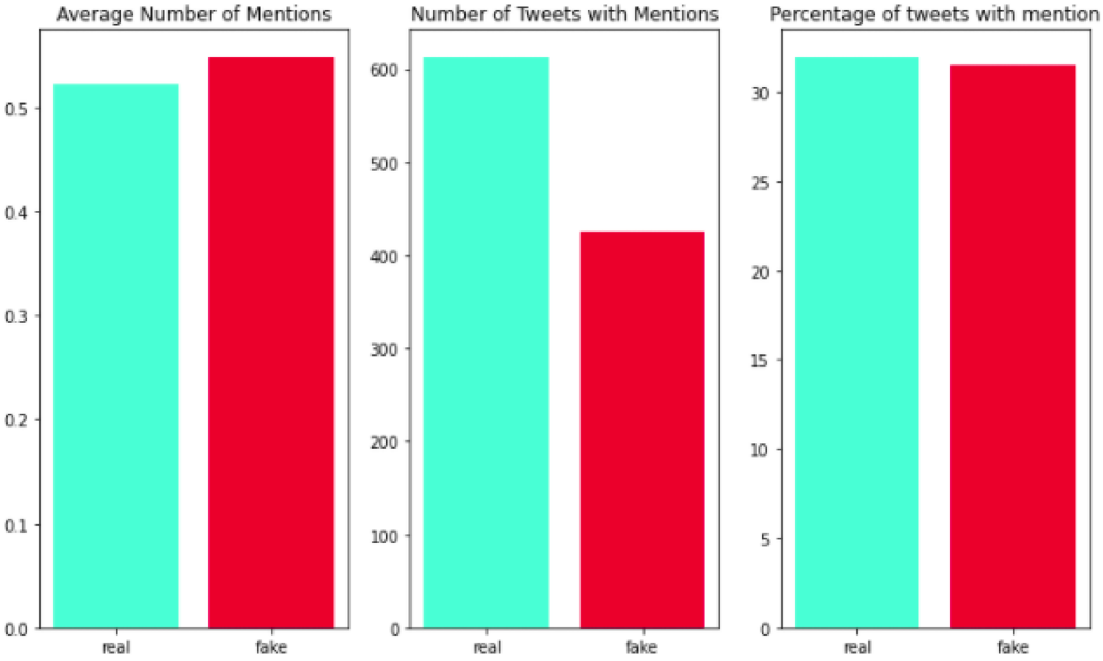


Figure 25. Average number of mentions in real and fake news

Real news mention frequency

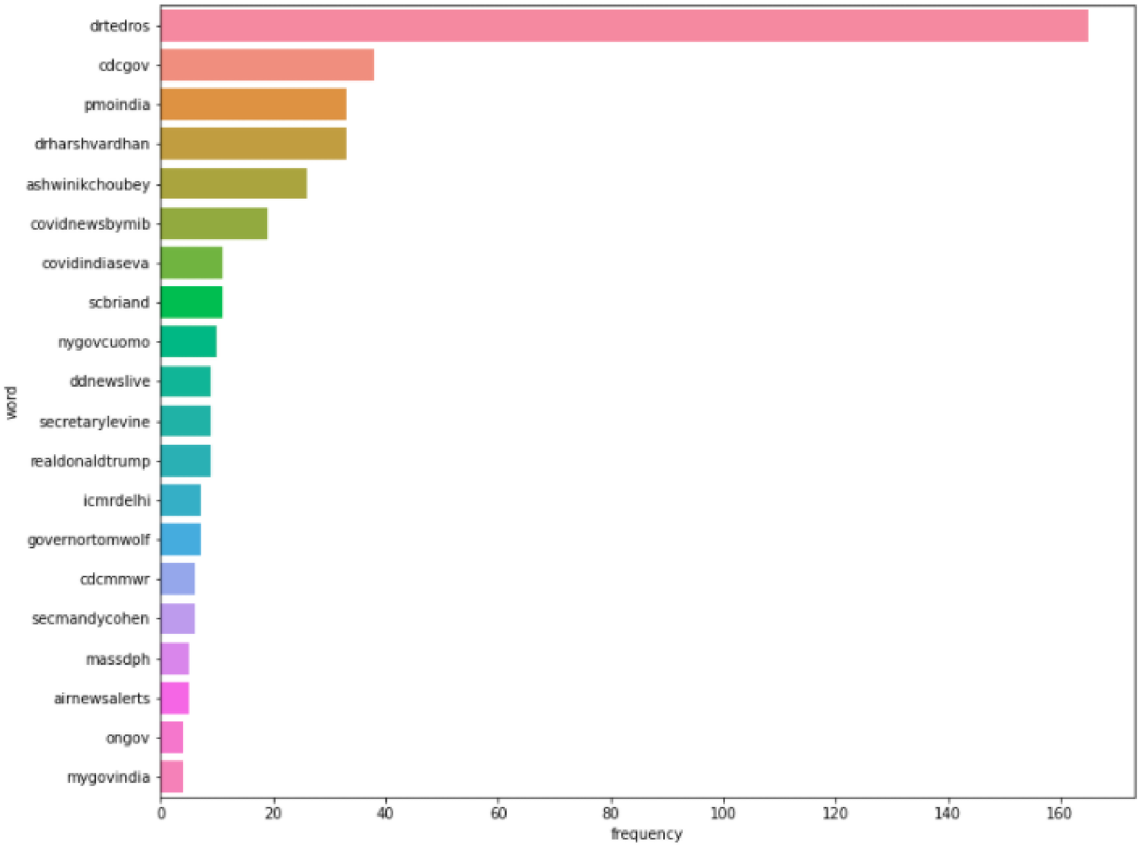


Figure 26. Most repeated mentions in real news

Fake news mention frequency

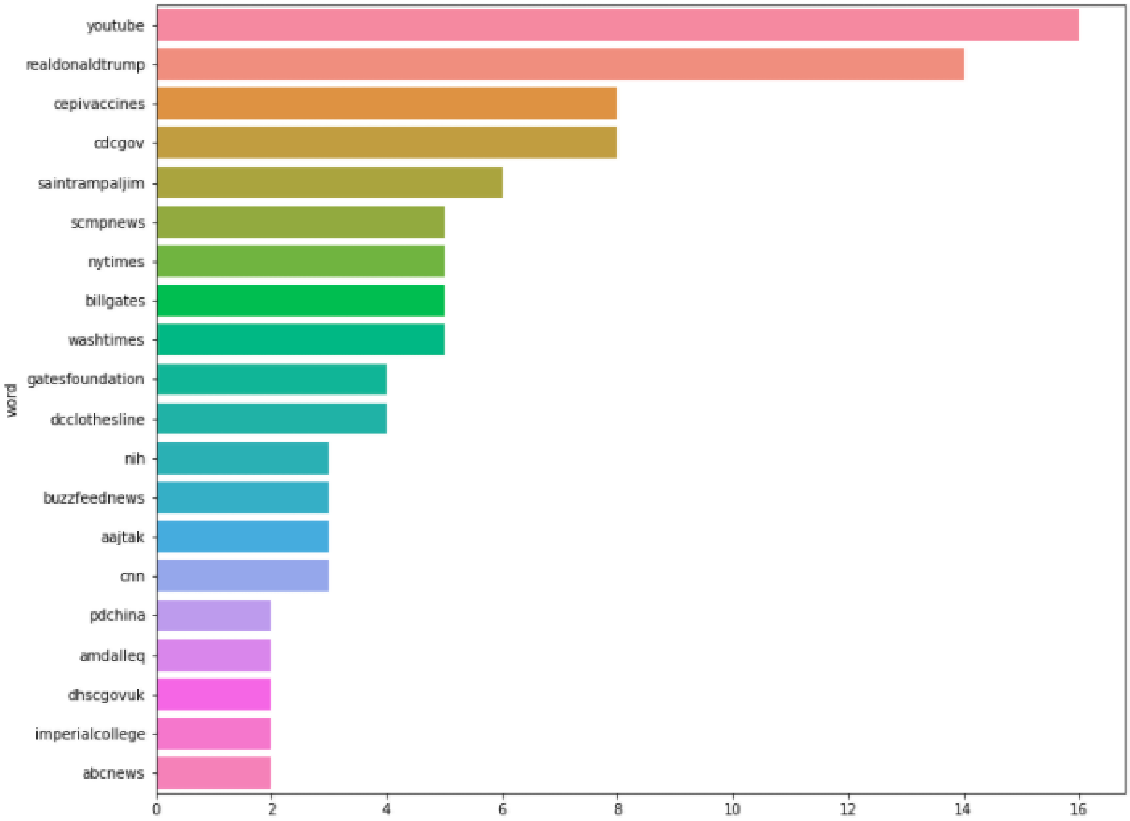


Figure 27. Most repeated mentions in fake news

WordCloud of mentions:

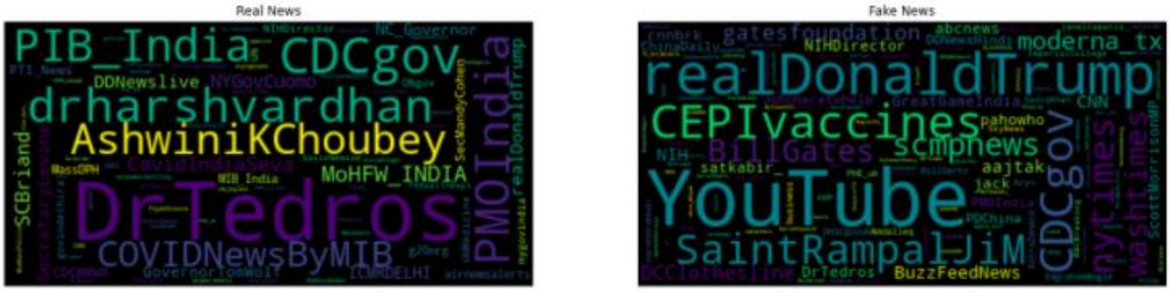


Figure 28. Word cloud of repeated mentions in real and fake news

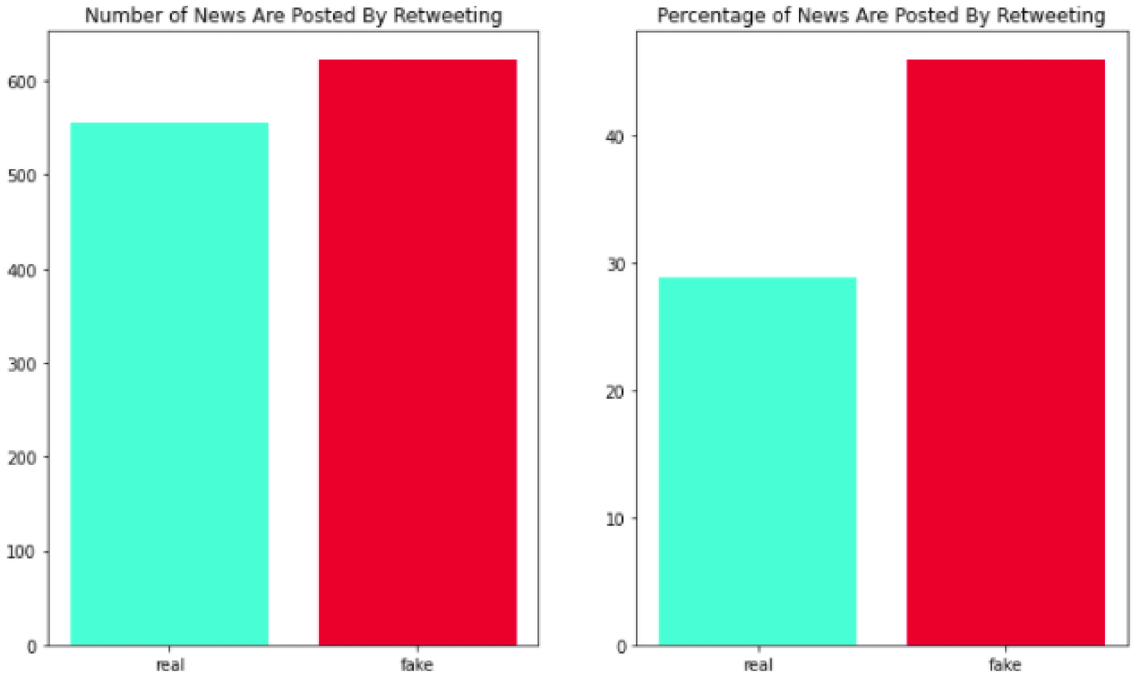


Figure 29. Average numbers of news posted by retweeting

Retweet and favourite analysis

Since popular opinion is very important in the analysis of social networks, we consider it appropriate to analyse the number of RTs and favourites that both real and fake news get.

For this purpose, we have calculated the average RT and favourites values in both data frames, from which the following values are obtained:

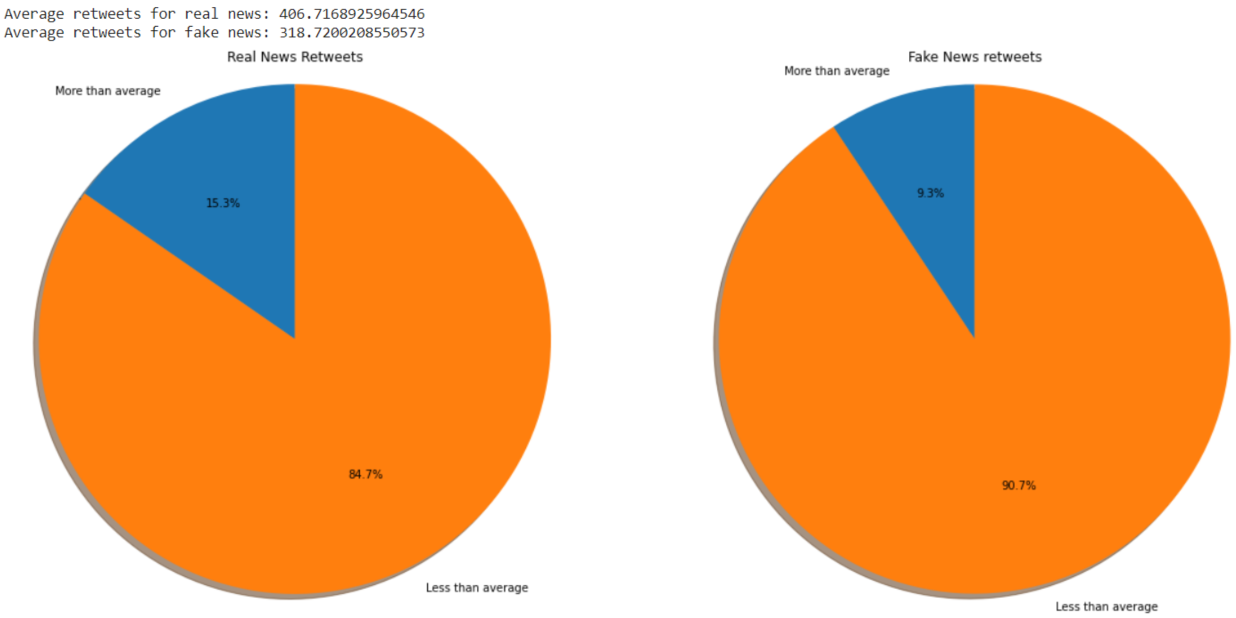
	REAL NEWS	FAKE NEWS
RT AVERAGE	407,19	319,23
FAV AVERAGE	1087,67	1332,31

Table 1: Average retweets and favorites for each tweet of real and fake news

As usual, the average values of favourites are much higher than those of retweets. After calculating these values, the tweets have been divided based on whether they have been supported above average or below it.

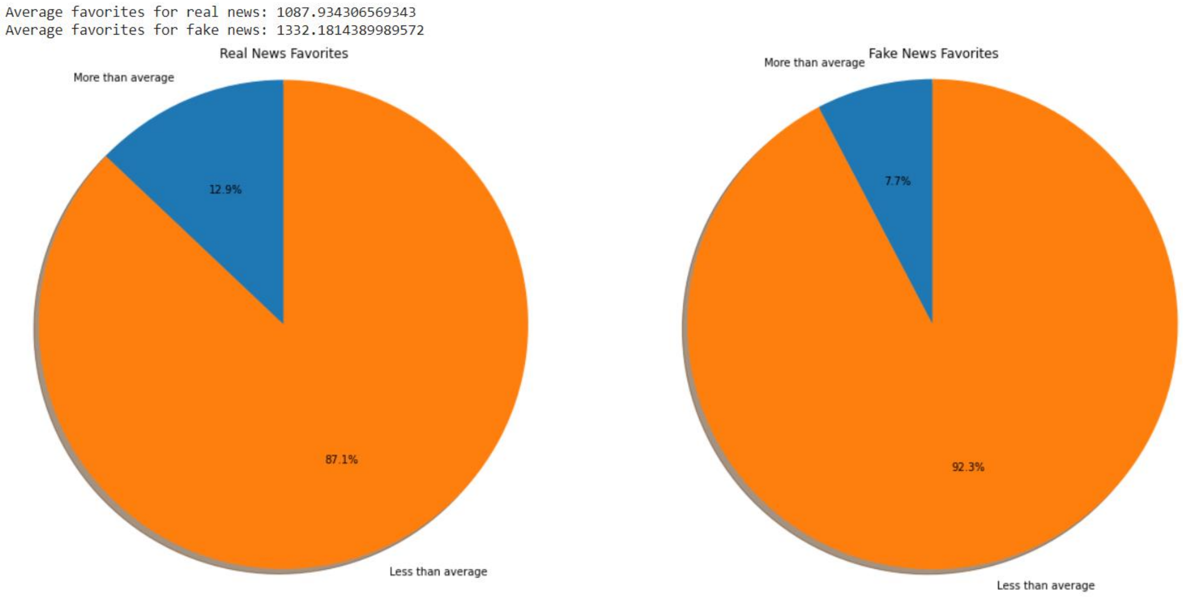
From these graphs we can see that the vast majority of tweets in all cases have values below the average. This could mean that there is a small percentage of tweets that go viral and get a lot of support, which makes the average increase, but that the great majority of the tweets collected in these datasets are not very recognised. We will see this analysis later on, but the main reason for this to happen is that the users of these datasets have few followers, making it difficult for their tweets to have a wide reach.

On the other hand, the average number of favourites of fake news tweets is approximately 30 percent higher than real ones. This is an interesting fact, taking into account that the number of followers of the fake news accounts is quite smaller than the real news accounts.



```
Average retweets for real news: 406.7168925964546
Average retweets for fake news: 318.7200208550573
```

Figure 30. Average Retweets in real and fake news tweets



```
Average favorites for real news: 1087.934306569343
Average favorites for fake news: 1332.1814389989572
```

Figure 31. Average Favourites in real and fake news tweets

Location analysis

From the data gathering process we obtained the location information of all tweets, which indicates the country (and usually the city) from which the tweet was originated. We have analysed the locations from both datasets and we get the results you can observe in the following pictures.

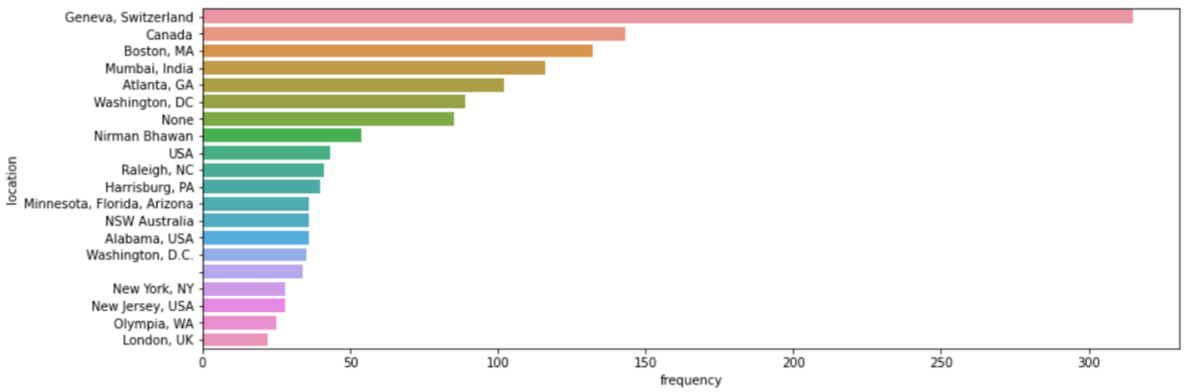


Figure 32. Most frequent locations in real news

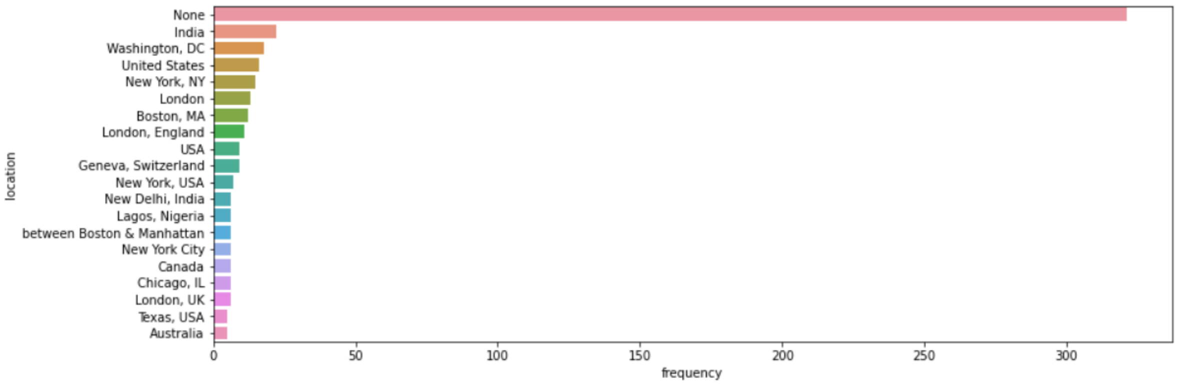


Figure 33. Most frequent locations in fake news

In the case of real news, it is observed that most tweets indicate their location and only a small sector, 6% of the total number of entries, do not specify where the tweet originated. In addition, 315 tweets (more than 16% of the dataset) originate from Geneva, Switzerland, the place from which the WHO (World Health Organization) user sends his tweets, so we conclude who is the prevailing account in the data frame. Moreover, when adding the 20 most frequent locations, the 76% of the full data frame is grouped together, so we concluded that the location information is mainly distributed in some important places, meanwhile others are mostly irrelevant.

When analysing fake news, we get quite different results. First, it is observed that a large percentage of the dataset (almost 24%) has no information about its location. In addition, there is no clear dominance of the dataset, since the frequency with which the locations appear is quite distributed and if we add the 20 most mentioned places, we only collect 37.5% of the total number of tweets.

Popularity analysis

Finally, we thought it would be interesting to analyse the relationship between the popularity of the users in the data frames (frequency with which they appear) and the number of followers each one has. To do this we have obtained the 20 most popular accounts of each dataset and we have extracted two graphs, one for each parameter to be analysed.

In the case of real news, among the 20 most popular accounts are some of the most relevant in the news environment, such as BBC News, NPR or Politico, and accounts such as WHO, which are quite important given the subject of the analysis, Covid-19. The fact that the news sources have so many followers is a good factor in confirming the veracity of the results.

However, it can be seen that the 20 most popular accounts in the fake news dataset are of dubious reliability. The vast majority of them, even the number one on the list, have very few followers, which is strange in the truthful communication media. It is rather curious the case of "The Independent", a British newspaper that has a great media impact, but that in our analysis is classified as one of the biggest fake news accounts in the dataset.

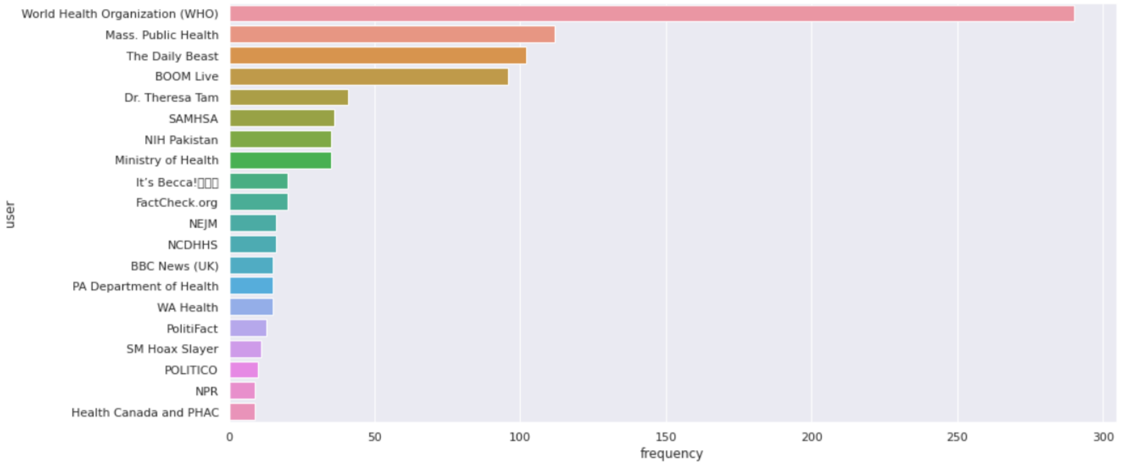


Figure 34. Frequencies with which the most popular 20 real news accounts appear in the dataset

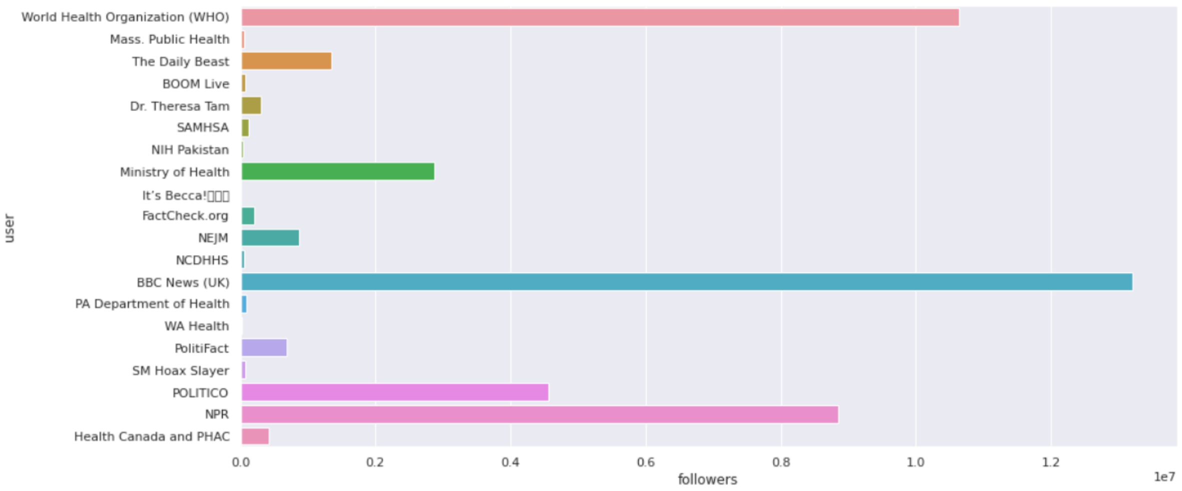


Figure 35. Followers of the most popular 20 real news accounts

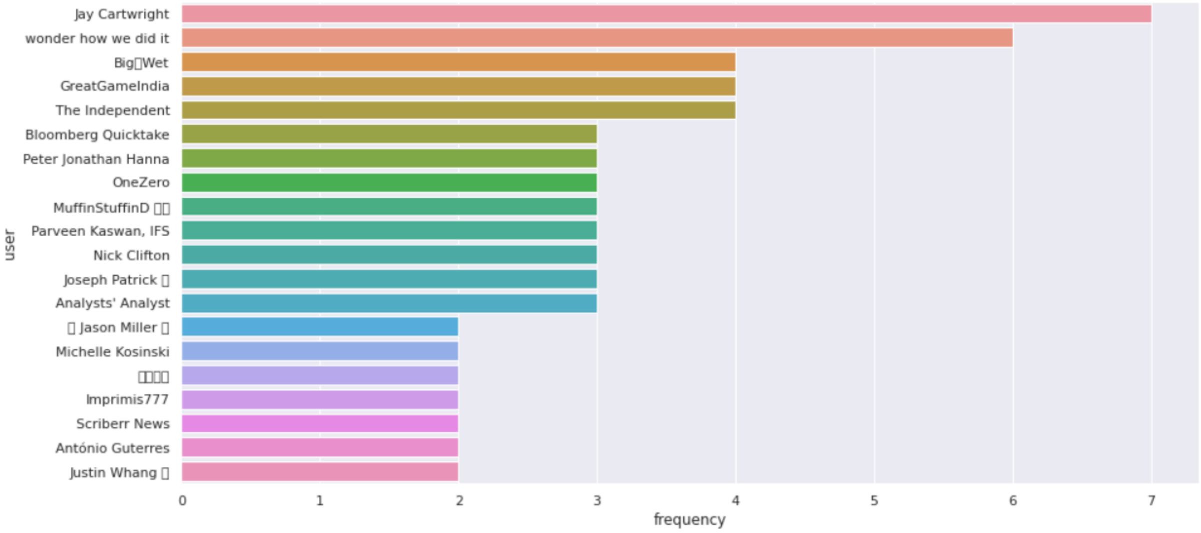


Figure 36. Frequencies with which the most popular 20 fake news accounts appear in the dataset

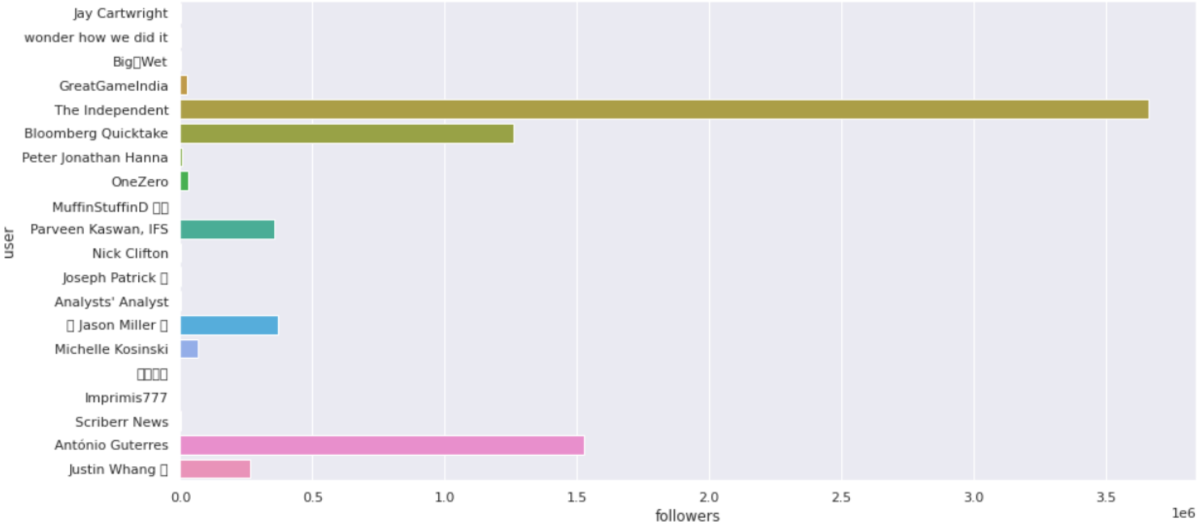


Figure 37. Followers of the most popular 20 fake news accounts

As an additional measure, we have created a scatter plot that allows us to see the distribution of the 20 most popular accounts in each dataset based on their followers and frequency.

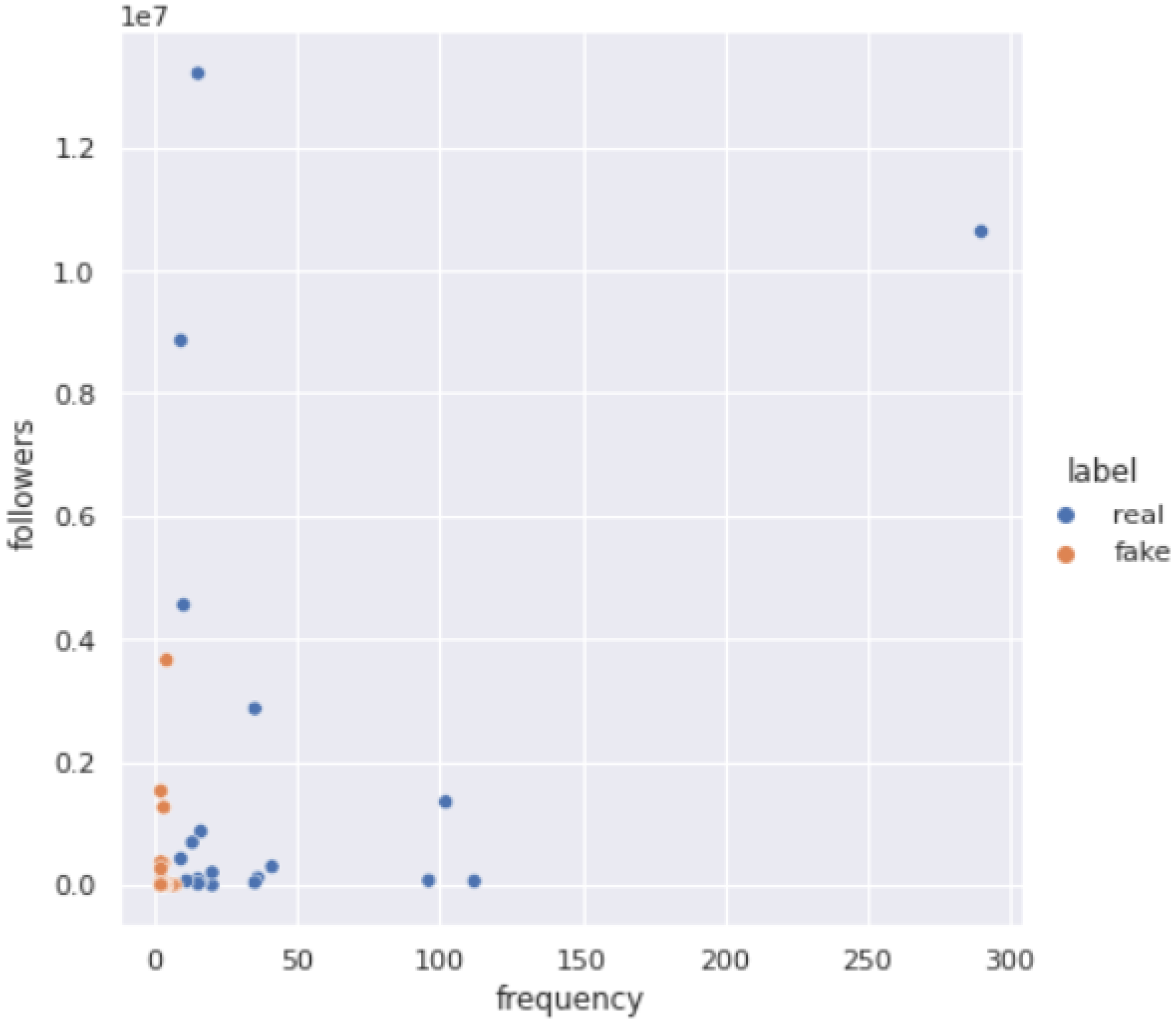


Figure 38. Scatter plot of the frequency/followers relation of the most popular 20 accounts on each news class

As we can see, real news sources are more widely distributed and generally have larger numbers of followers and appear more frequently in the dataset, while accounts classified as fake news sources are generally less popular on twitter and are not mentioned as frequently in our data.

Network Analysis

Accounts Network Analysis

We have designed 2 networks to keep analysing the most used Twitter accounts in real news and fake news tweets.

The first one contains the 15 most used accounts in the real news dataset. The nodes of this network are these 15 twitter accounts and are related according to whether a user follows or is followed by the rest of the nodes.

The second one contains the 15 most used accounts in the fake news dataset and has been created following the same criteria as for the first one.

To obtain the relationships between users we have used the Tweepy library which, as we have already mentioned, allows us to access the Twitter API.

We have used the method show_friendship(source_id,target_id) in which the source_id and target_id are the identifiers of the Twitter users. This method returns an object in which we can see if user A follows user B, user B follows user A, both follow each other or neither follows each other.

MOST USED REAL NEWS ACCOUNTS NETWORK			MOST USED FAKE NEWS ACCOUNTS NETWORK		
Nodes	Edges	Directed	Nodes	Edges	Directed
15	27	True	15	4	True

Table 2: Network parameters for fake and real news accounts

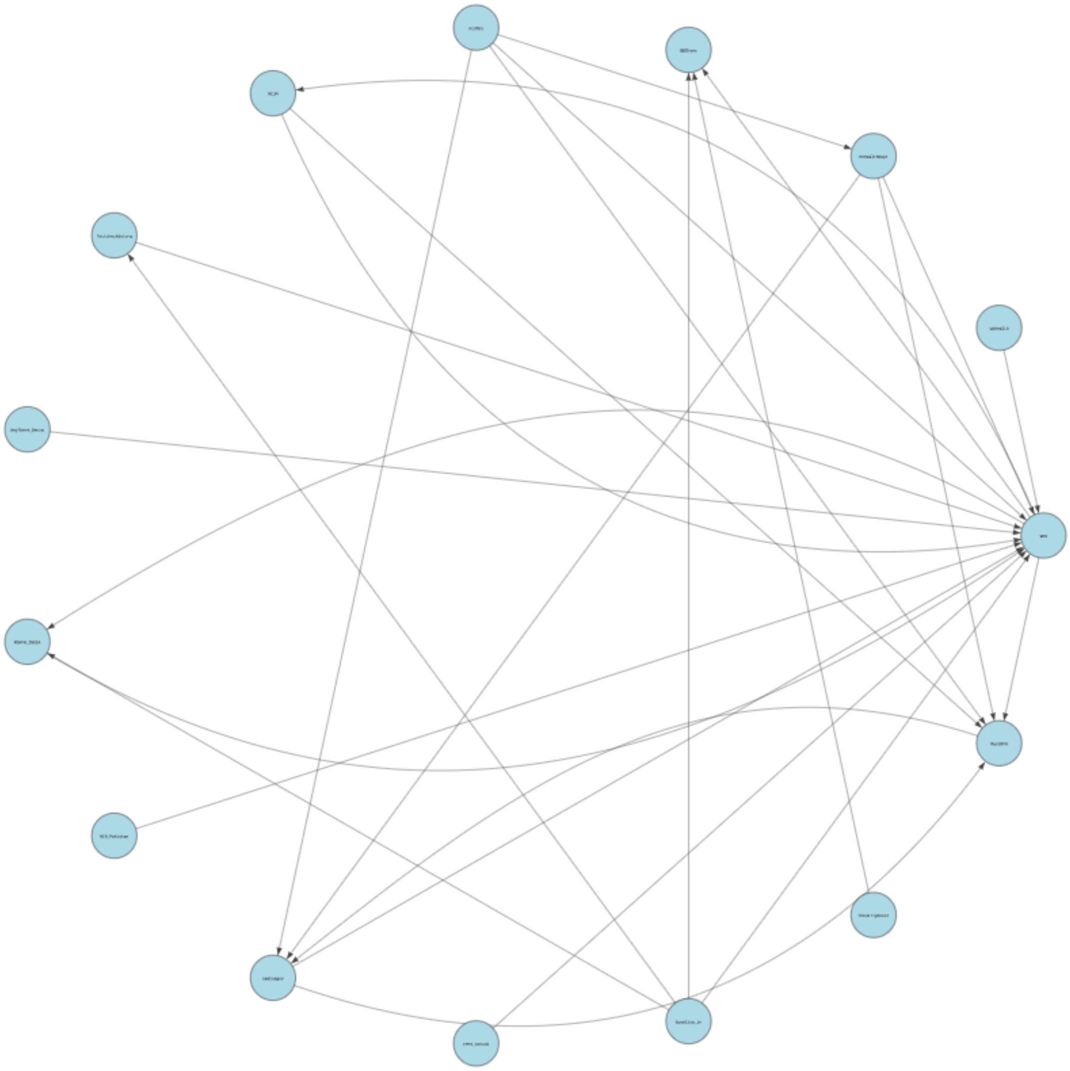


Figure 39. Most used real news accounts Network

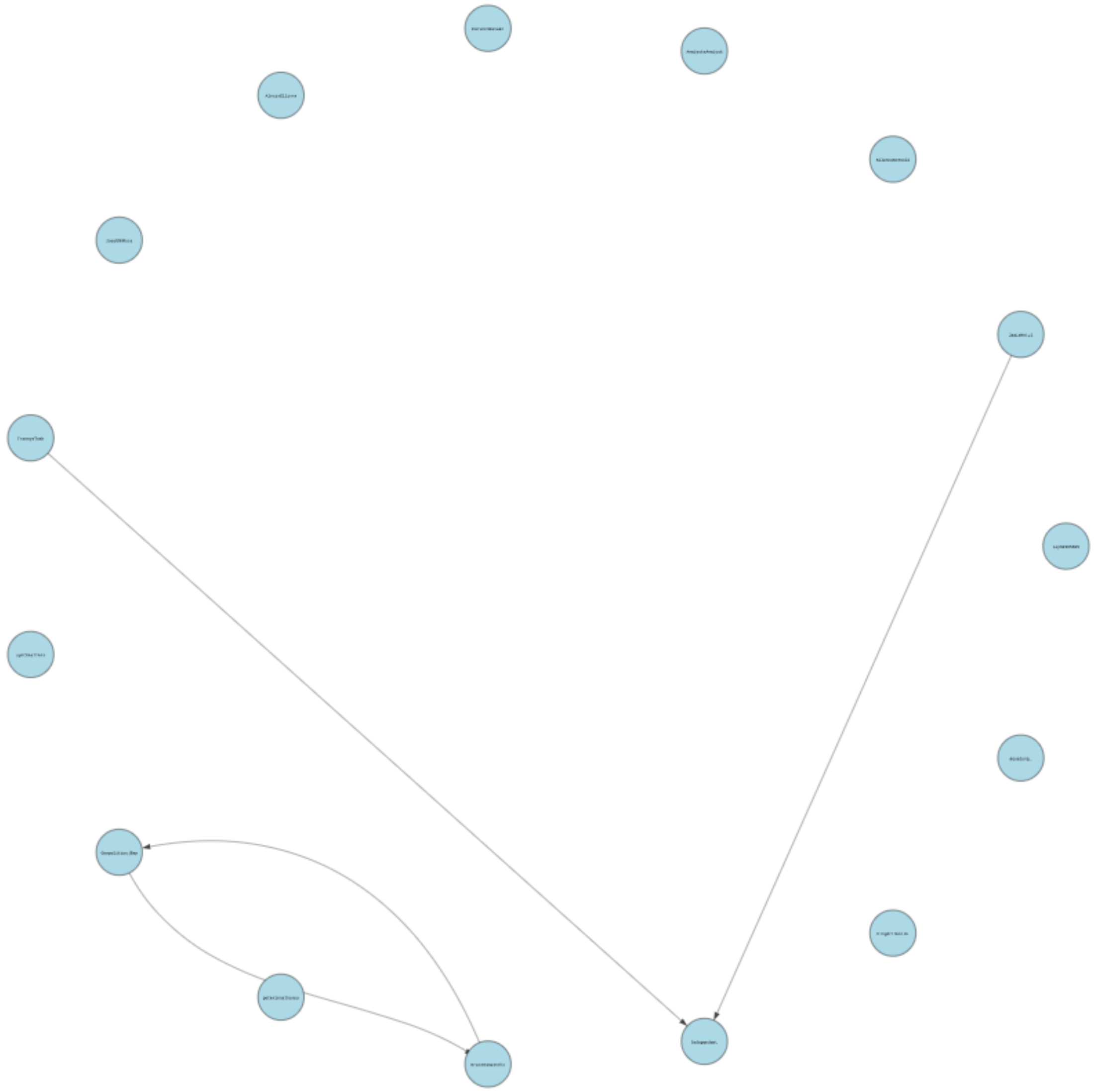


Figure 40. Most used fake news accounts Network

On these two networks we have analyzed which user has the most followers from the network itself and which user follows the most users from the network.

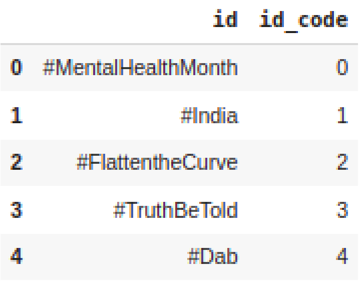
On the first network, the most followed user is World Health Organization (WHO) which is followed by a total of 11 of the 14 users in the network. As we have seen previously, this user is also the most used source of dataset information. The accounts that follow the most users in the network are NCDHHS which is the North Carolina Department of Health and Human Services, World Health Organization (WHO) and Boom Live which is a certified fact-driven journalism. All of them follow 4 nodes within the network.

On the second network, the most followed user is The Independent, which is followed by 2 of the 14 users in the network. Users Jamie Metzl, Nick Clifton, Geopolitics & Empire and GreatGameIndia follow one node within the network.

We can conclude there is much more relation between the network of the accounts that are sources of real news than the one created by the accounts that are sources of fake news.

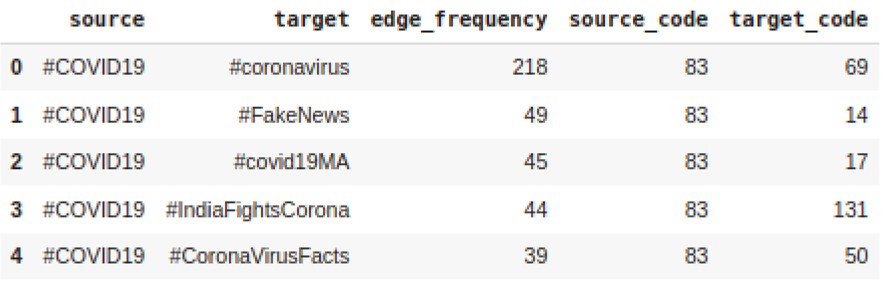
Hashtag Network Analysis

In order to get more detailed features and correlations we decided to use advanced concepts of network science. So, we extracted the hashtags from the tweets and computed the frequency of each hashtag. After that we assumed that hashtags are our nodes and edges are the users that used the both hashtags in their tweets. So, firstly we created our node list with associated IDs and edge list with source, target and edge frequency information.



	id	id_code
0	#MentalHealthMonth	0
1	#India	1
2	#FlattentheCurve	2
3	#TruthBeTold	3
4	#Dab	4

Figure 41. Dataframe of nodes



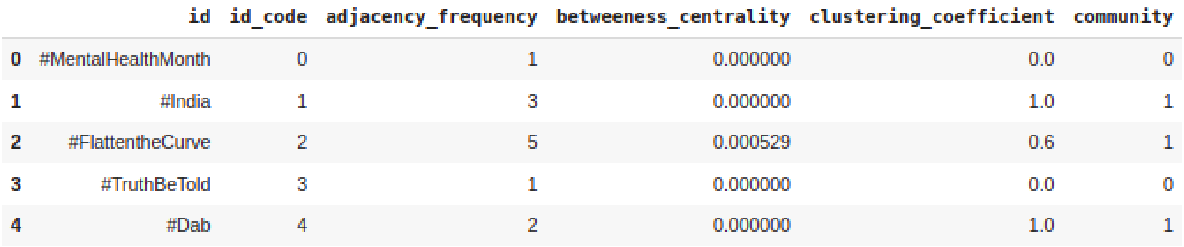
	source	target	edge_frequency	source_code	target_code
0	#COVID19	#coronavirus	218	83	69
1	#COVID19	#FakeNews	49	83	14
2	#COVID19	#covid19MA	45	83	17
3	#COVID19	#IndiaFightsCorona	44	83	131
4	#COVID19	#CoronaVirusFacts	39	83	50

Figure 42. Dataframe of edges

At this point, we created our graph and since it's ready for analysis we did further calculations to enhance the graph. So, we firstly did betweenness centrality analysis [8], which is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

After that, we computed the clustering coefficient [9] for nodes. Which is a measure of the degree to which nodes in a graph tend to cluster together. Finally, we applied community detection using greedy modularity [10] maximization provided by networkx package which begins with each node in its own community and joins the pair of communities that most increases modularity until no such pair exists.

Node list after the enhancement:



	id	id_code	adjacency_frequency	betweeness_centrality	clustering_coefficient	community
0	#MentalHealthMonth	0	1	0.000000	0.0	0
1	#India	1	3	0.000000	1.0	1
2	#FlattentheCurve	2	5	0.000529	0.6	1
3	#TruthBeTold	3	1	0.000000	0.0	0
4	#Dab	4	2	0.000000	1.0	1

Figure 43. Enhanced node dataframe

As the data is ready we plotted the graph, we represented the edge frequency by the size of the node, while their color represents their community

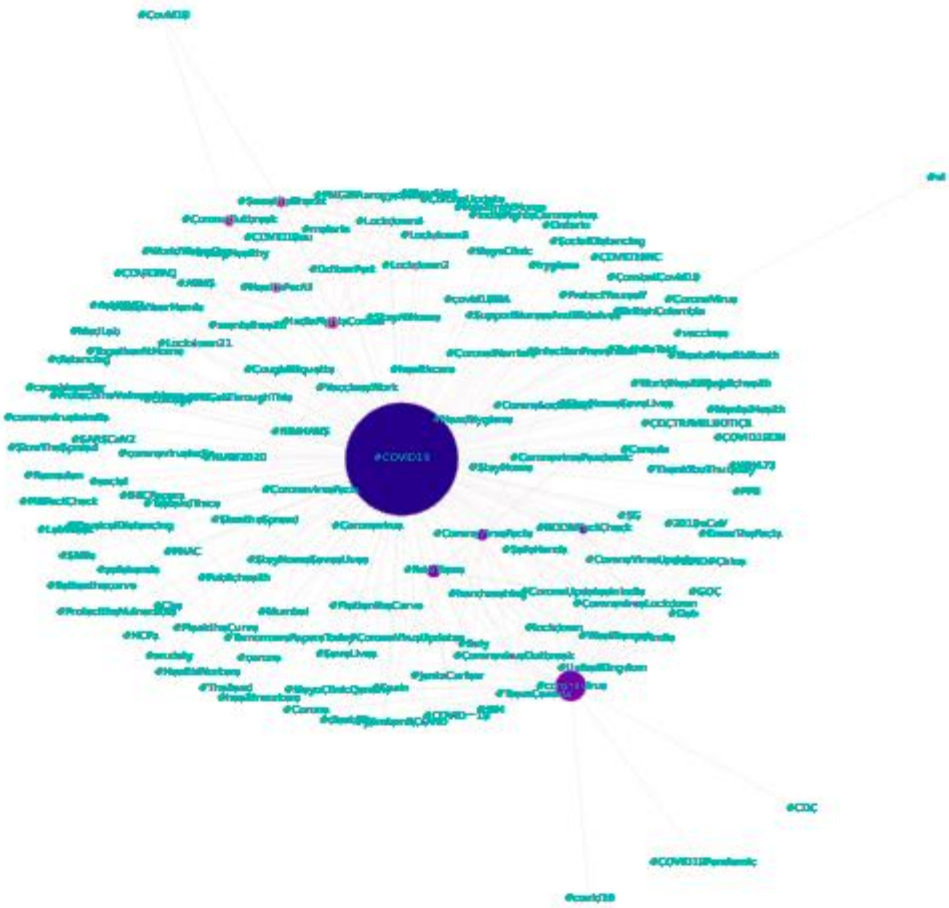


Figure 44. Real news hashtags network

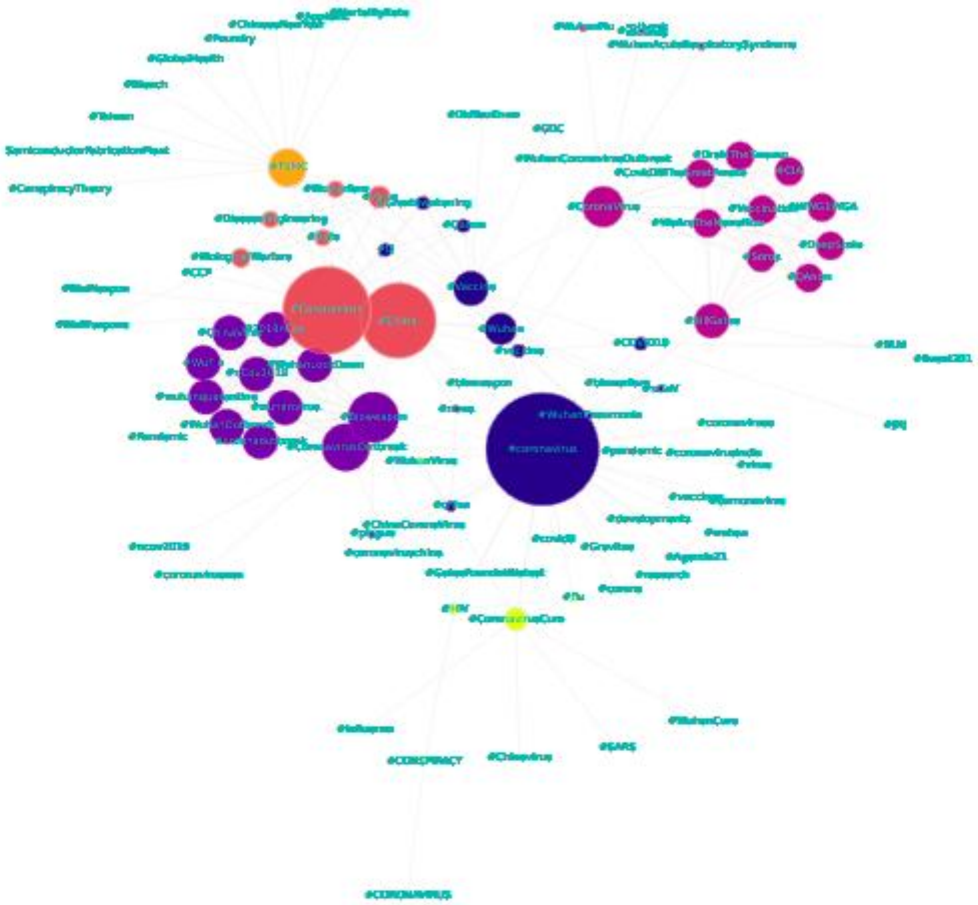


Figure 45. Fake news hashtags network

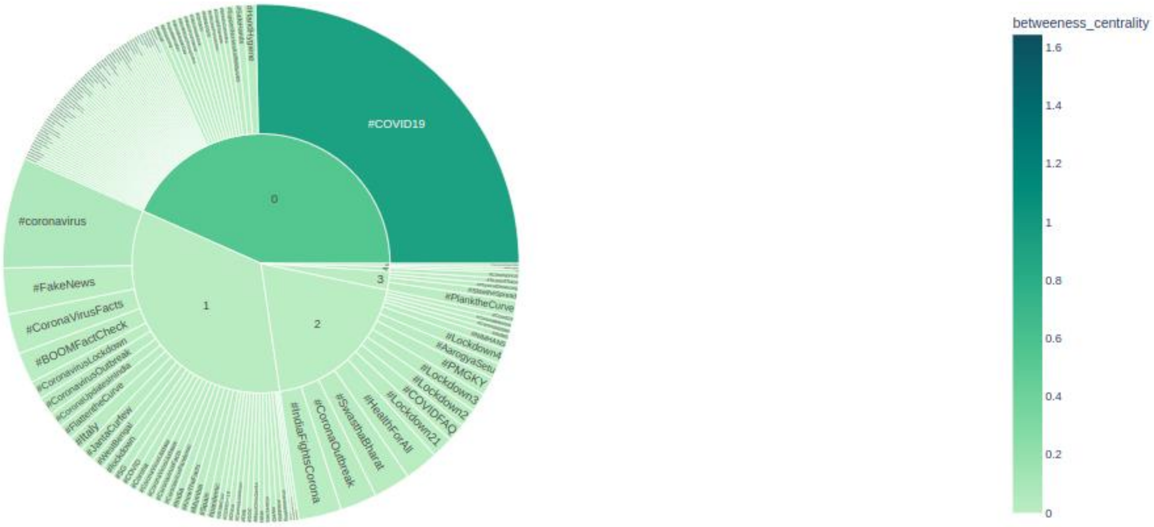


Figure 46. Detected communities in real news

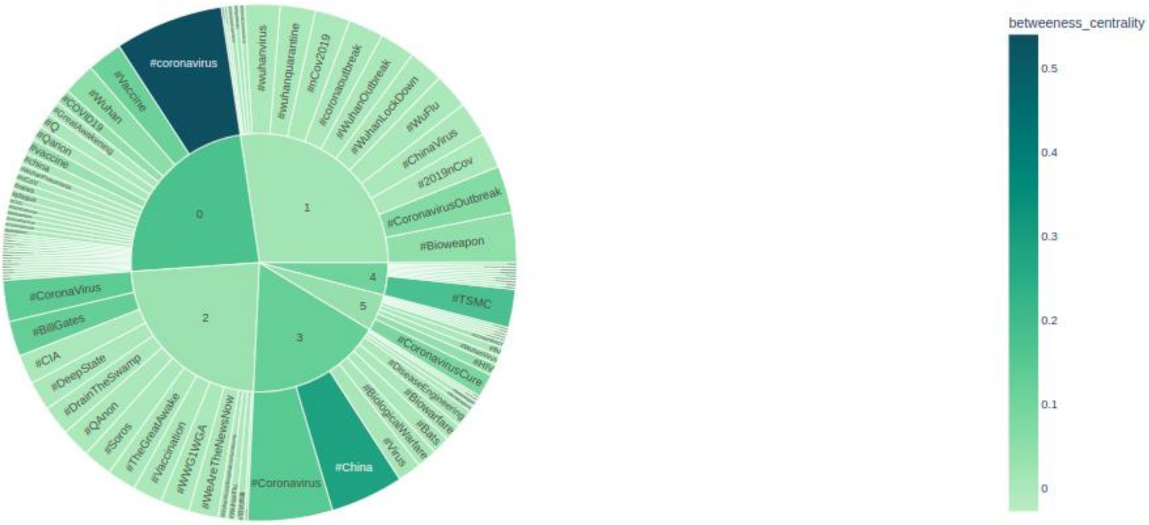


Figure 47. Detected communities in fake news

After that, we realized that that hashtag #COVID19 repeated too many times and it is obviously a basic hashtag in both real and fake news. So we excluded it from analysis and recalculated all values.

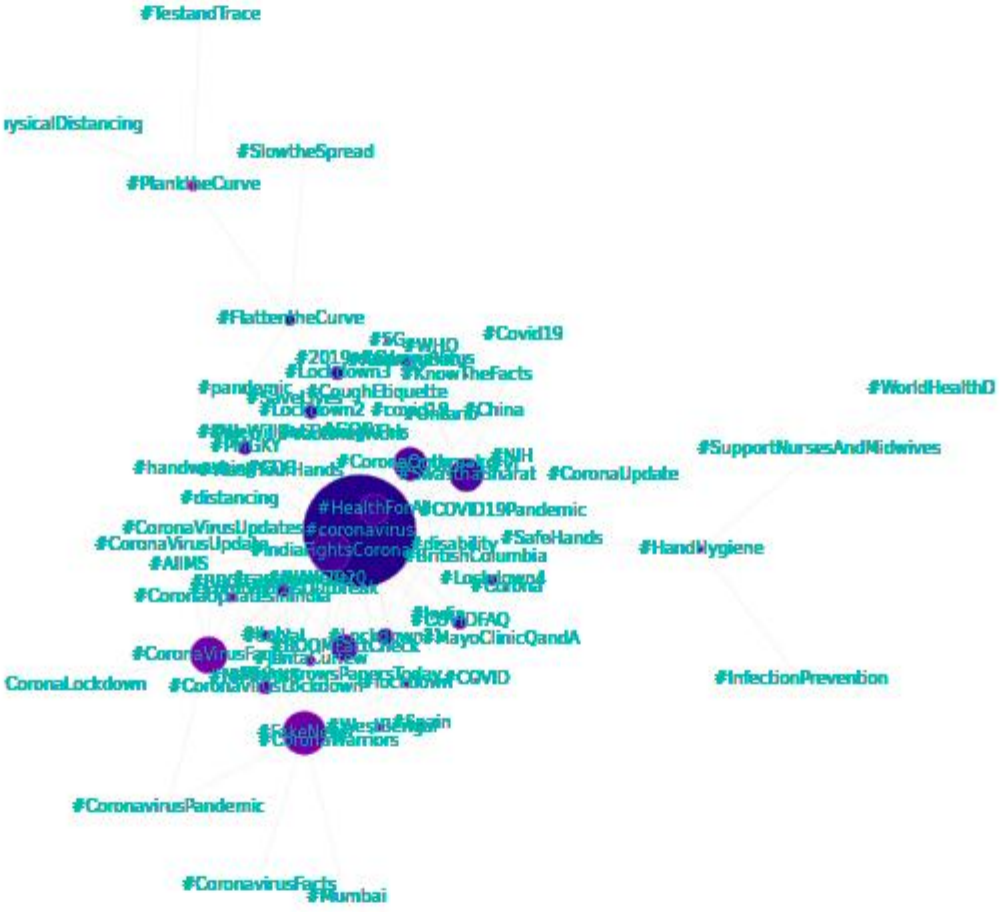


Figure 48. Re-calculated real news hashtags network

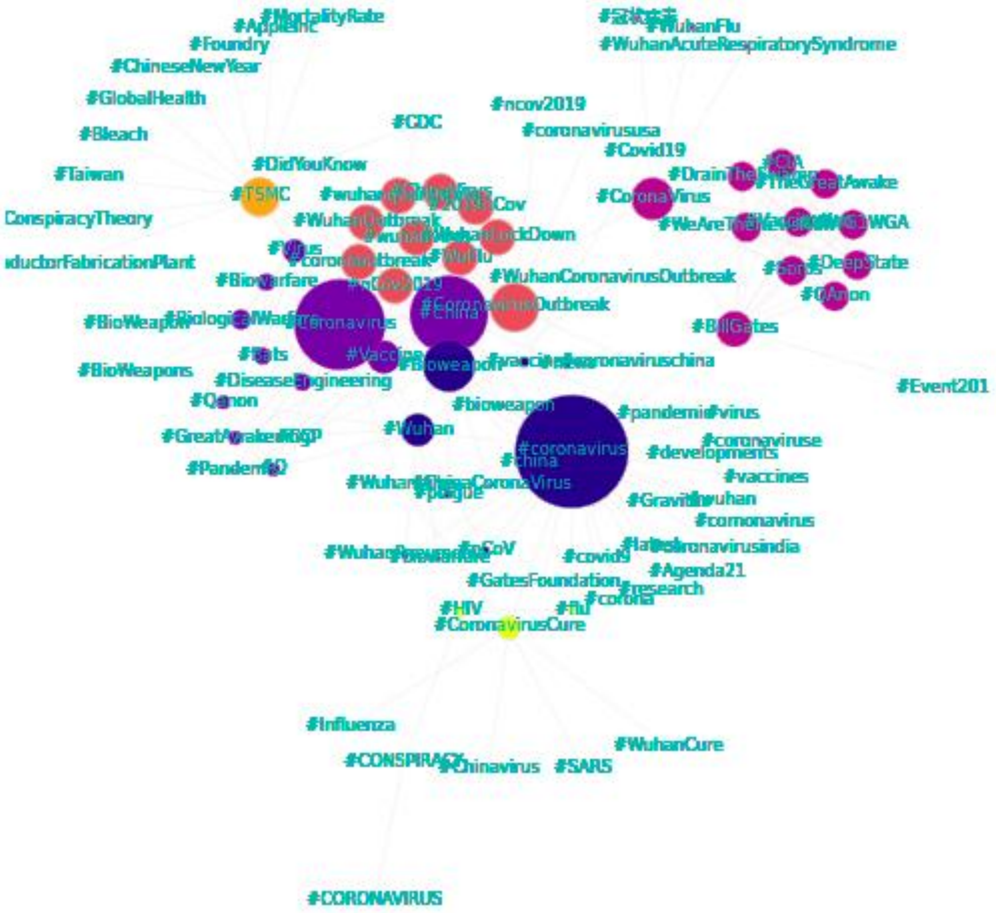


Figure 49. Re-calculated fake news hashtags network

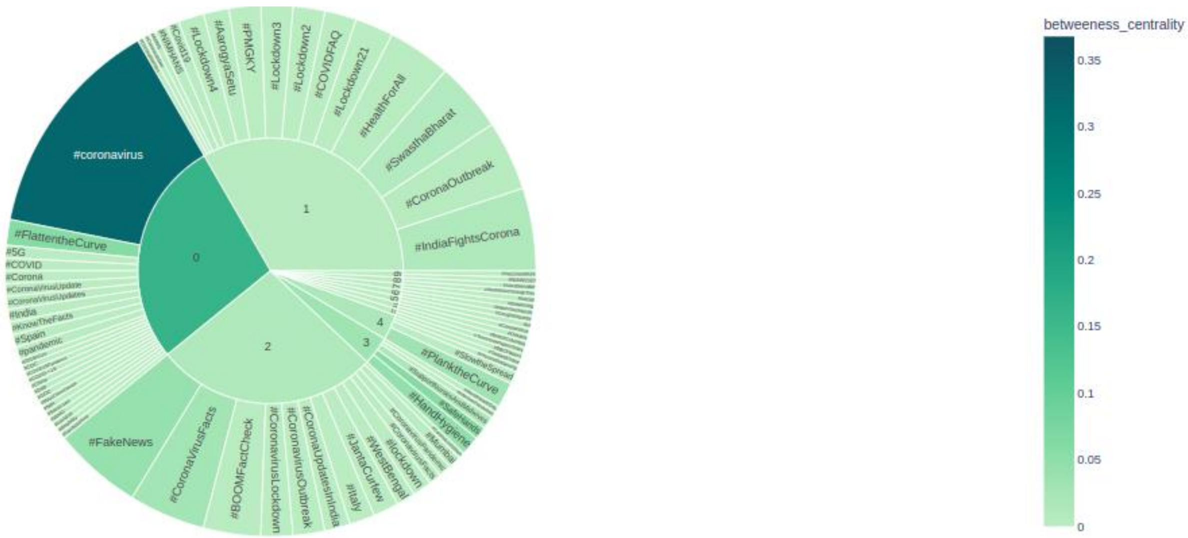


Figure 50. Re-calculated detected communities in real news

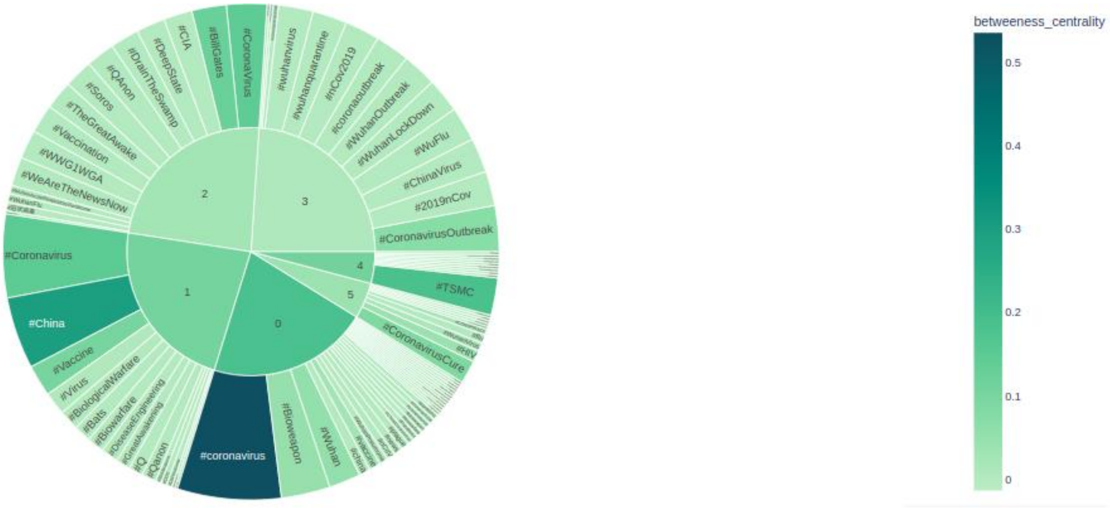


Figure 51. Re-calculated detected communities in fake news

3.3. Classification Models

Machine Learning Based Solution

In the beginning we started by preparing our data for training by stemming them, which means removing morphological affixes from words and leaving only the word stem. After that we vectorized the data by using the TF-IDF concept, which is a technique used to weight terms according to the importance of those terms within the document and corpus. Words that are frequent in a document but not across the corpus tend to have high TF-IDF scores. Afterwards, we splitted our data with fractions of (0.25, 0.75) to training and testing datasets. Lastly, we feeded our data to various algorithms of machine learning like logistic regression [11], K-Nearest Neighbors [12], Decision Tree [13] and Multinomial Naive Bayes [14]. We trained all of these classifiers and compared the results. In response, we found that we got the highest accuracy by training the Decision Tree algorithm.

Algorithm	Logistic Regression	K-Nearest Neighbors	Decision Tree	Multinomial Naive Bayes
Accuracy (%)	83.92	62.05	84.99	70.12

Table 3: Accuracies of tested machine learning models

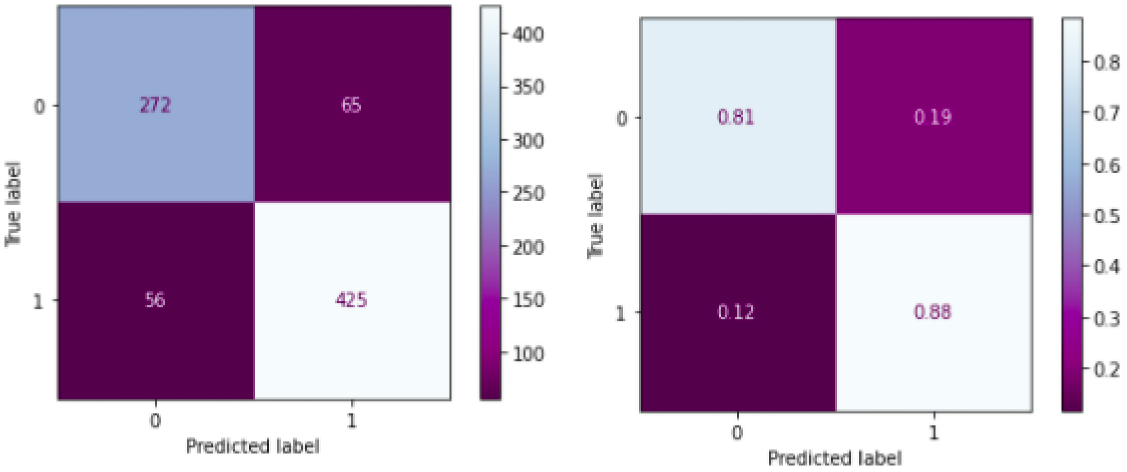
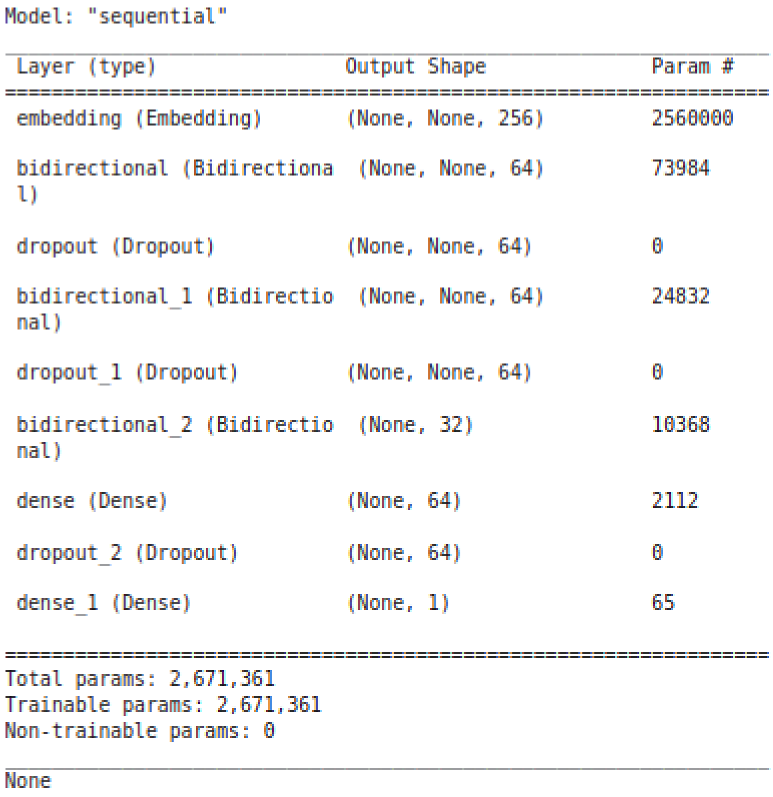


Figure 52. Confusion matrix of best machine learning model

Deep Learning Based Solution

After the results we got from machine learning based classifiers we did not get satisfied. Therefore, we decided to train a deep learning based network so we can get better classifiers with higher accuracy. So, we started by representing our data as a one-hot representation so every word (even symbols) which are part of the given text data are written in the form of vectors, constituting only 1 and 0 . So one hot vector is a vector whose elements are only 1 and 0. Each word is written or encoded as one hot vector, with each one hot vector being unique. This allows the word to be identified uniquely by its one hot vector and vice versa, that is no two words will have the same one hot vector representation.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 256)         2560000

 bidirectional (Bidirectiona  (None, None, 64)         73984
 l)

 dropout (Dropout)           (None, None, 64)          0

 bidirectional_1 (Bidirectio  (None, None, 64)         24832
 nal)

 dropout_1 (Dropout)         (None, None, 64)          0

 bidirectional_2 (Bidirectio  (None, 32)               10368
 nal)

 dense (Dense)               (None, 64)                2112

 dropout_2 (Dropout)         (None, 64)                0

 dense_1 (Dense)             (None, 1)                 65

=================================================================
Total params: 2,671,361
Trainable params: 2,671,361
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 53. Summary of designed network

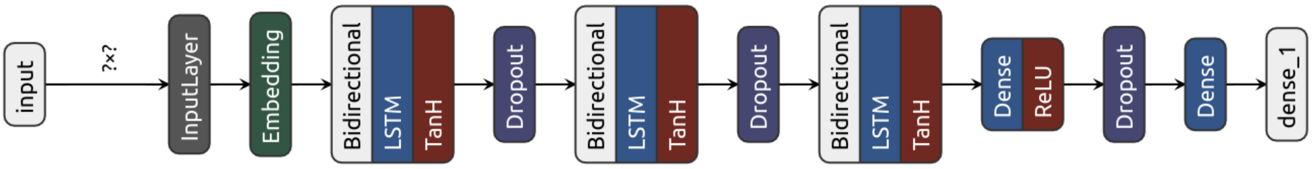


Figure 54. Structure of designed network

Subsequently, we created our network which consists of 9 layers; 1 embedding layer, 3 bidirectional LSTM layers [15], 3 dropout layers and 2 dense layers. We used a batch size of 32, Binary Crossentropy as a loss function and Adam with 1e-4 learning rate as an optimizer. We had 2 stopping criterions, the first is number of epochs, we setted it as 10, therefore, if we exceed the epoch threshold the training is stopped. The other criteria is loss of validation set, if it gets stabilized for subsequent 3 epochs without improving we stop the learning. The best model is protected all over the training procedure, so even if the model gets bad by epochs we can extract the best model, thanks to the EarlyStopping module of Keras. Eventually, our network trained until the 8th epoch and stopped because of the stability of validation loss. We got the best model at 7th epoch with a validation accuracy of 94.5%.

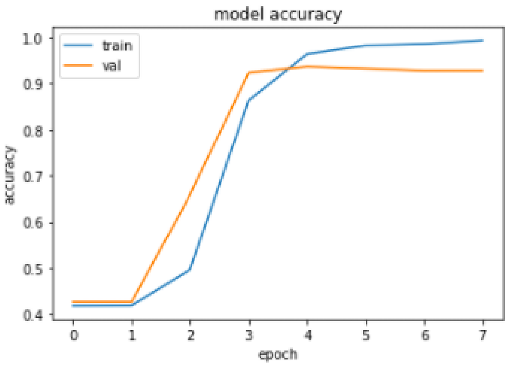


Figure 55. Training and validation accuracy of designed model

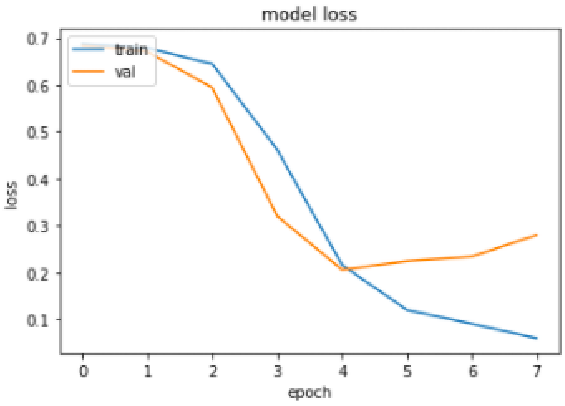


Figure 56. Training and validation loss of designed model

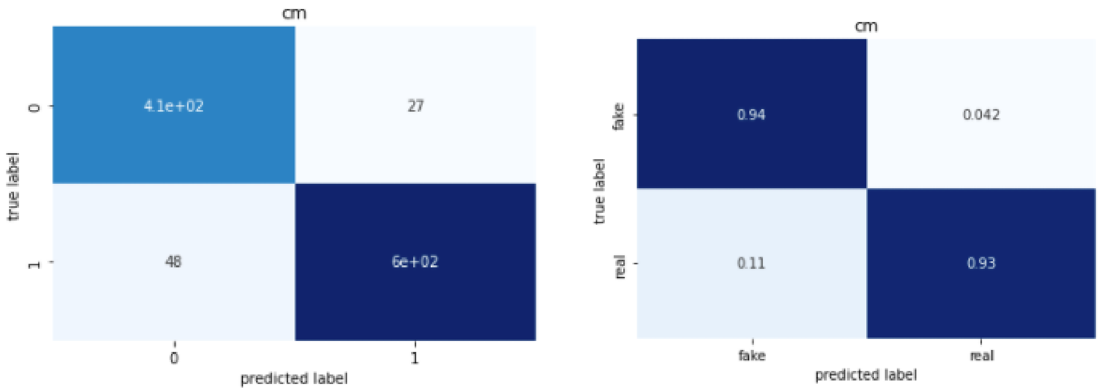
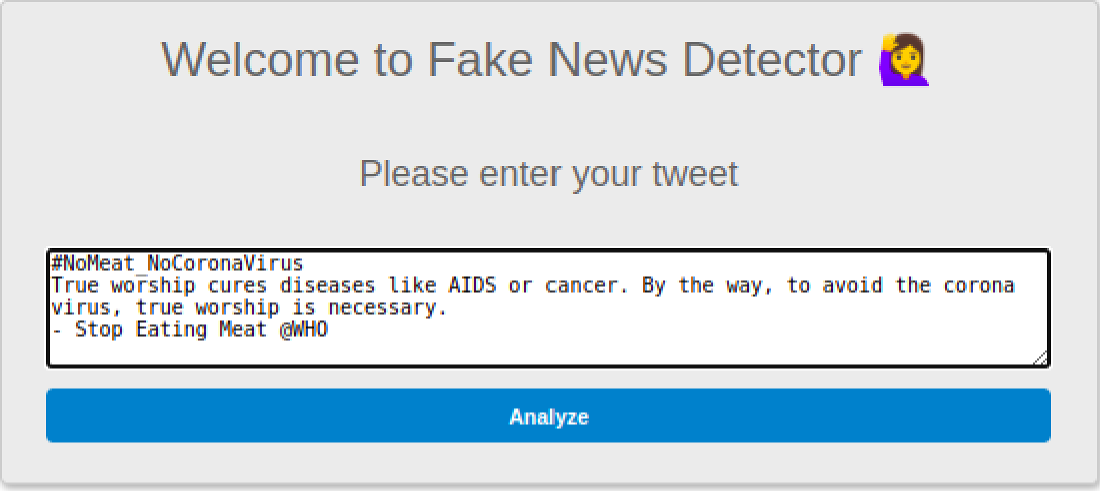


Figure 57. Confusion matrix of designed model

3.4. Deployment

After we got satisfied with our model, we decided to make it easily usable by implementing a web page where we can put news and get the corresponding result from the model. Therefore, we implemented our inference scripts using Flask framework and pushed them to a private GitHub repository. Lately, we created a Heroku project to deploy our project by connecting the Heroku project to the repository on GitHub. So, by opening the link (https://fnd-unipd.herokuapp.com/) we can type texts and monitor the behavior of the model easily.

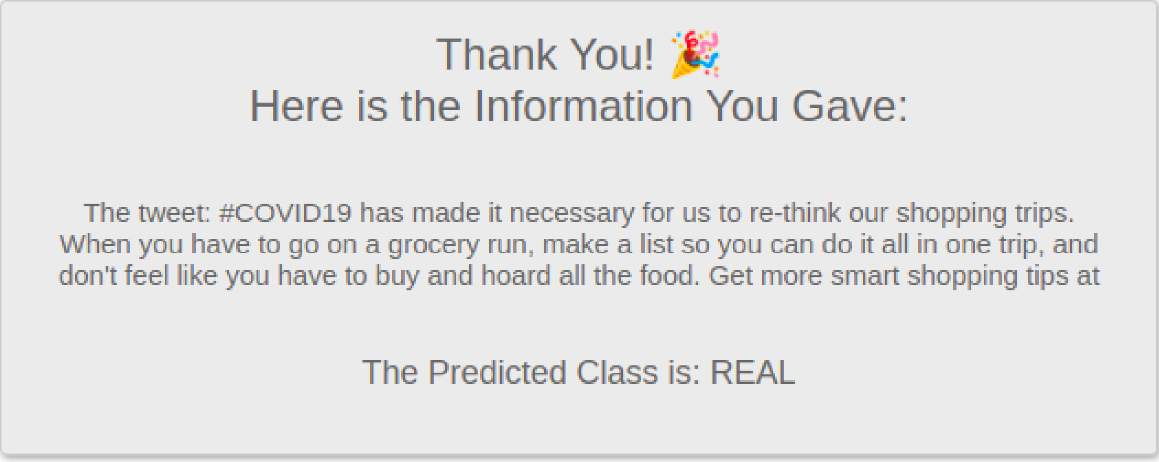


Welcome to Fake News Detector 🙋

Please enter your tweet

#NoMeat_NoCoronaVirus
True worship cures diseases like AIDS or cancer. By the way, to avoid the corona virus, true worship is necessary.
- Stop Eating Meat @WHO

Figure 58. News entering interface

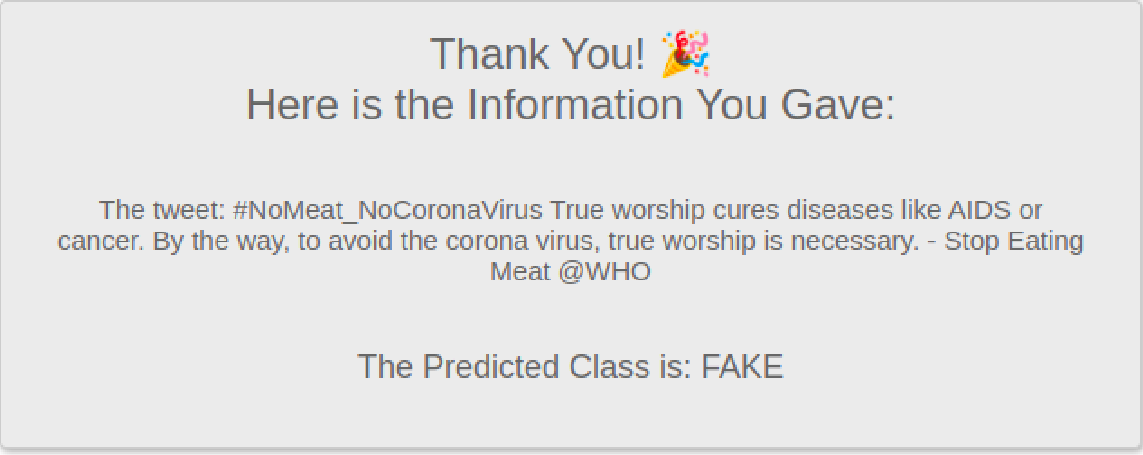


Thank You! 🎉
Here is the Information You Gave:

The tweet: #COVID19 has made it necessary for us to re-think our shopping trips. When you have to go on a grocery run, make a list so you can do it all in one trip, and don't feel like you have to buy and hoard all the food. Get more smart shopping tips at

The Predicted Class is: REAL

Figure 59. Output of case of real news result



Thank You! 🎉
Here is the Information You Gave:

The tweet: #NoMeat_NoCoronaVirus True worship cures diseases like AIDS or cancer. By the way, to avoid the corona virus, true worship is necessary. - Stop Eating Meat @WHO

The Predicted Class is: FAKE

Figure 60. Output of case of fake news result

4. Contributions

In this section of the report we show which member of the team has worked on each part of the project.

Data Gathering	
Accessing data	*Sandra, Martin*
Cleaning Data	*Sandra, Martin*
Analysis	
Sentiment	*Asmaa, Martin, Sandra*
Semantic	*Asmaa*
Hashtag	*Asmaa*
Mentions	*Asmaa*
Retweets	*Martin*
Favourites	*Sandra*
Location	*Sandra*
Popularity	*Martin*
Network Analysis - Account	*Martin*
Network Analysis - Hashtag	*Asmaa*
Classification Models	*Asmaa*
Deployment	*Asmaa*

Table 4: Contributions of each group member to the project

5. Conclusions

After carrying out this work we can conclude that, although fake news is very popular, it is quite difficult to get enough data to make a dataset of it, as no source presents itself as a creator of fake news. This has made obtaining data for this project very challenging.

On the other hand, accessing real news has been a fairly simple task, since nowadays all the media and reliable sources of information can be found on any social network, as Twitter is.

In the analysis we carried out in this research, we discovered that both words and hashtags of real tweets tended to be unbiased and more scientific, while words and hashtags in fake news tended to be biased and slang. In addition, we noticed that users that shared real news mentioned reliable pages and accounts, while users that shared fake news mentioned unreliable pages and accounts.

As far as we analyzed in this research, we found that there are possible relations between hashtags used in one tweet that can give us insights about real and fake news. On the other hand, we found that accounts that are sources of real news are more popular than the ones sharing fake news in terms of followers.

As already pointed out in the models part, we found that it is possible to create a model with an acceptable performance that can automatically distinguish fake news using machine learning based methods. Besides we realized that deep learning based models provide quite better and more satisfactory performance in the context that we investigated.

References

[1] Bovet, A., Makse, H.A. Influence of fake news in Twitter during the 2016 US presidential election. Nat Commun 10, 7 (2019). https://doi.org/10.1038/s41467-018-07761-2

[2] Al-Ahmad, Bilal & Al-Zoubi, Ala & Abu Khurma, Ruba & Aljarah, Ibrahim. (2021). An Evolutionary Fake News Detection Method for COVID-19 Pandemic Information. Symmetry. 13. 1091. 10.3390/sym13061091.

[3] Okoro, Efeosasere & Abara, Benjamin & Alex, Umagba & Ajonye, Anyalewa & Isa, Zayyad. (2018). A hybrid approach to fake news detection on social media. Nigerian Journal of Technology. 37. 10.4314/njt.v37i2.22.

[4] Ngada, Okuhle & Haskins, Bertram. (2020). Fake News Detection Using Content-Based Features and Machine Learning. 1-6. 10.1109/CSDE50874.2020.9411638.

[5] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), 2018, pp. 272-279, doi: 10.23919/FRUCT.2018.8468301

[6] Laboratory for Computational Social Systems (LCS2): Home. (2022). Retrieved 1 February 2022, from https://lcs2.iiitd.edu.in/

[7] GitHub - LCS2-IIITD/ENDEMIC. (2022). Retrieved 1 February 2022, from https://github.com/LCS2-IIITD/ENDEMIC

[8] Bader, David & David, & Kintali, & Shiva, & Madduri, Kamesh & Kamesh, & Mihail, Milena & Milena,. (2007). Approximating Betweenness Centrality. 10.1007/978-3-540-77004-6_10

[9] Li, Yusheng & Shang, Yilun & Yang, Yiting. (2017). Clustering coefficients of large networks. Information Sciences. 382-383. 350-358. 10.1016/j.ins.2016.12.027.

[10] Almukhtar, Ahmed & Al-Shamery, Eman. (2018). Greedy Modularity Graph Clustering for Community Detection of Large Co-Authorship Network. International Journal of Engineering & Technology. 7. 857. 10.14419/ijet.v7i4.19.28058.

[11] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[12] Guo, Gongde & Wang, Hui & Bell, David & Bi, Yaxin. (2004). KNN Model-Based Approach in Classification.

[13] Jijo, Bahzad & Mohsin Abdulazeez, Adnan. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2. 20-28.

[14] Xu, Shuo & Li, Yan & Zheng, Wang. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. 347-352. 10.1007/978-981-10-5041-1_57.

[15] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.