Top Concerns of Tweeters during the COVID-19 pandemic

Report created by Larisa-Maria Biriescu

June 2020

1 Introduction

Since the 1980s, human disease outbreaks have become increasingly frequent and diverse due to a plethora of ecological, environmental and socioeconomic factors. The family of coronaviruses was not considered to be highly pathogenic until 2003 and 2012 with the appearance of the severe acute respiratory syndrome in China followed by the Middle East respiratory syndrome in Saudi Arabia.

The recent coronavirus disease (COVID-19) pandemic is taking a toll on the world's health care infrastructure as well as the social, economic, and psychological well-being of humanity. Individuals, organizations, and governments are using social media to communicate with each other on a number of issues relating to the COVID-19 pandemic. Not much is known about the topics being shared on social media platforms relating to COVID-19. Analyzing such information can help policy makers and health care organizations assess the needs of their stakeholders and address them appropriately.

In this direction, a group of researchers conducted an infoveillance study at the beginning of the year in order to identify the main topics posted by Twitter users related to the COVID-19 pandemic and also performing a sentiment analysis on them. This will be an attempt to continue their work, and see how these concerns evolved around the tweeters a few months after the initial results were published by them.

2 Methods

2.1 Data Collection

In my study, I collected coronavirus-related English language tweets between May 17, 2020, and May 30, 2020 using the Twitter standard search application programming interface (API). The hashtags used were: 'corona', '2019-nCov', 'COVID-19', 'COVID19', 'coronavirus', 'lockdown', 'covid-19', 'Pandemic'.

For each tweet I extracted the **time when the tweet was created**, **tweet ID**, **text of the tweet**, **number of retweets**, **number of likes**, **number of followers**, **number of friends**, **user ID**.

The tweets were stored on Google Drive, in .csv files, per day. The tweet ID column has been used as the index. For accesing the Twitter API I used Tweepy Python library, Pandas and Numpy libraries for dealing with the .csv files. The entire code used for this project was written and ran on Google Colab cloud-based Jupyter notebook environment.

	tweet_id	date_created	source	text	retweet_count	favourite_count
0	1262111976634724352	17-05-2020	<a href="https://mobile.twitter.com" rel="nofo...	A little known consequence of attending a #COV...	0	0
1	1262111973954682885	17-05-2020	<a href="http://twitter.com/download/iphone" r...	@MiaWasp @samurai3434 @GovCanHealth I myself k...	0	0
2	1262111973833146368	17-05-2020	<a href="http://twitter.com/#!/download/ipad" ...	Familiarize yourself with these patterns of Tr...	0	0
3	1262111973707309056	17-05-2020	<a href="http://twitter.com/download/iphone" r...	Italian MP Sara Cunial Speaks the Truth About ...	0	0
4	1262111970624495616	17-05-2020	<a href="https://www.fs-poster.com" rel="nofol...	Panic As First Class Gombe Monarch, 'Mai Of Ta...	0	0

Figure 1: Example of tweet data collected

followers_count	friends_count	location	retweeted_status	user_id	full_date_created
1437	1360	Sacramento, CA	NaN	21430553	2020-05-17 20:05:18
39	261	On my horse, Ste-Julie QC CA	NaN	299390457	2020-05-17 20:05:18
316	259	The Beach at Newport, RI	NaN	286967039	2020-05-17 20:05:18
1	10	NaN	NaN	1193526542971932672	2020-05-17 20:05:18
106	88	NaN	NaN	1638968804	2020-05-17 20:05:17

Figure 2: Example of tweet data collected

2.2 Data Preprocessing

The preprocessing step is a very important one and I invested a large amount of time in it because of the concept "garbage in, garbage out". It's futile to invest a lot of time in training and building high-performace models when the data that is being fed to them is not properly cleaned. From the start I gathered only the tweets written in English, without including the retweets, after this step has been completed a few more steps have been added and can be seen in the below

diagram. I have chosen lemmatization instead of stemming due to the fact that even though the first approach is slower than the second one, the result is an actual language word - more than that, it is the dictionary form of it, whereas by using stemming we might end up with a root that is not an actual language word.

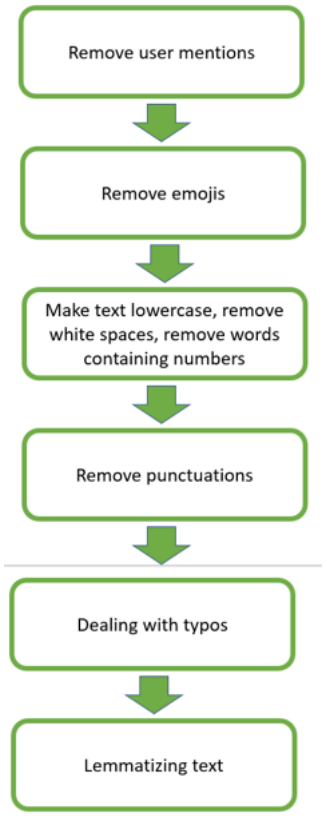

Figure 3: My approach at data preprocessing

2.3 Data Analysis

My approach on data analysis started with building unigrams containing: all words, only nouns, (adjectives, nous) and (adjectives, nouns, verbs). For topic modeling I used three algorithms: LDA(Latent Dirichlet Allocation), LSA(Latent Semantic Analysis), NMF(Non-negative matrix factorization), the latter being known for performing good on smaller pieces of text.

I started with 12 topics but empirically, after running my algorithms and checking the topics distribution I identified 5 more robust topics. Due to the fact

that tweets have a small amount of words, I assumed that only 1 topic/tweet will be present.

I also performed sentiment analysis and extracted the mean number of retweets, likes, and followers for each topic and calculated the interaction rate per topic.

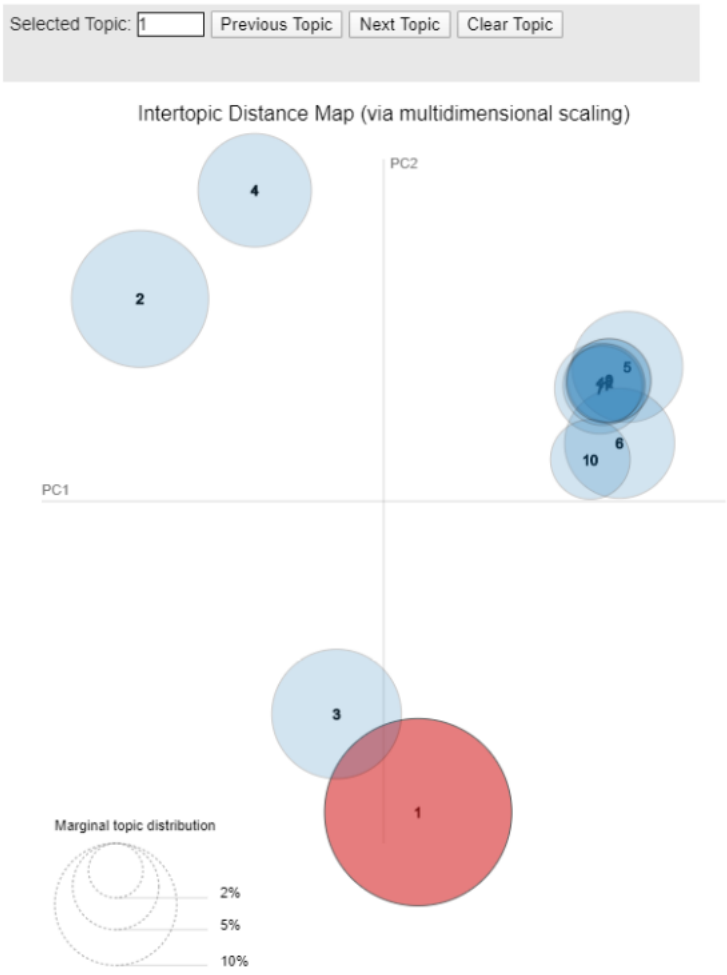

Figure 4: Topic visualization (t12 p25 LDA all)

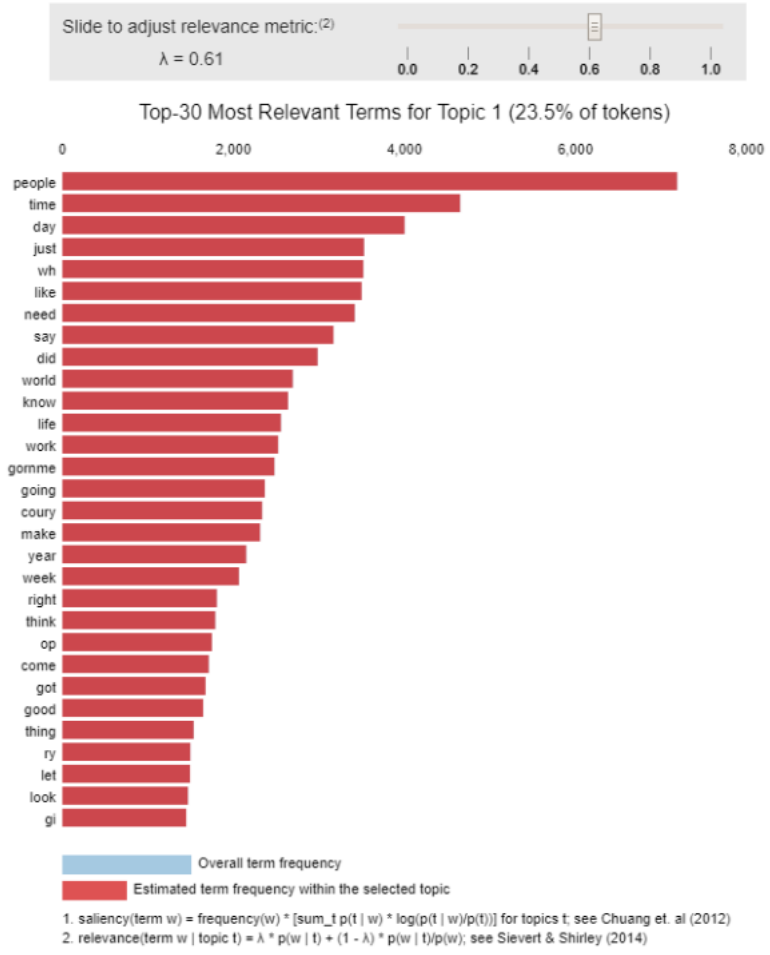

Figure 5: Example of relevant terms for a topic

Figure 6: Topic visualization (t5 p25 LDA all)

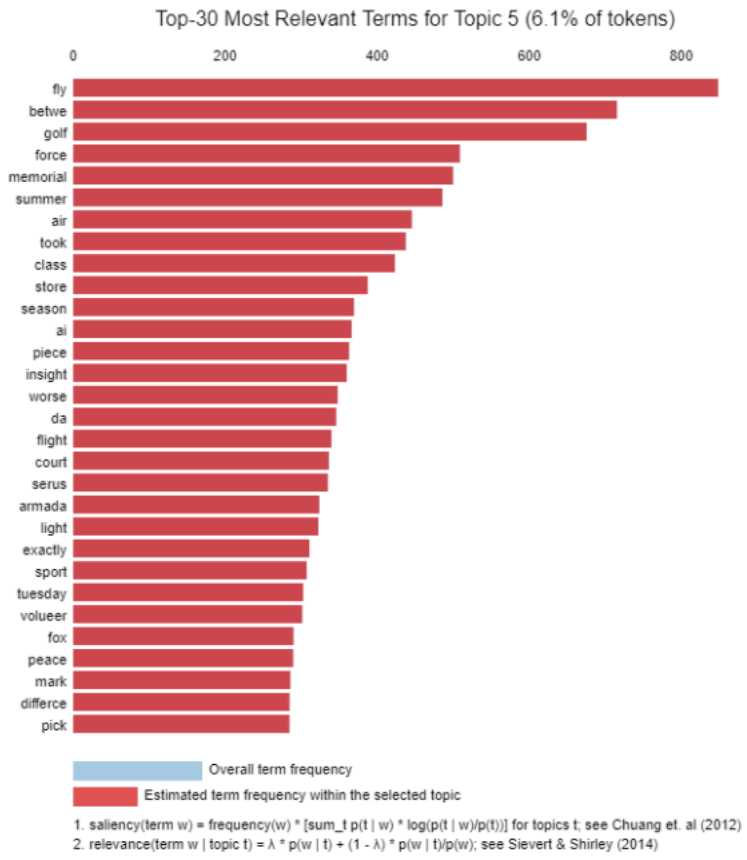

Figure 7: Example of relevant terms for a topic

3 Results

3.1 Results of Tweet Analysis

Between May 17, 2020, and May 30, 2020, I collected **140,000 tweets**, from **102,626 unique users**. Due to Google Colab limitations I ran the models only on 10 days: **100,000 tweets** from **77,929 users**.

The research group identified **12 topics** from the analyzed tweets. The 12 topics were grouped into **four themes**: the origin of COVID-19, the source of a novel coronavirus, the impact of COVID-19 on people and countries, and the methods for decreasing the spread of COVID-19.

Looking at the outputs generated by my models, I identified as well **5 main concerns** of the tweeters: ones regarding the healthcare system, businesses, society, personal life and protection measures/ spread regarding the coronavirus. After I labeled the results assuming that each tweet contains only one major topic, the number of valid tweets(the ones that are labels as containing one of the topics discovered) dropped slightly from **100,000 tweets to 83,832 tweets**.

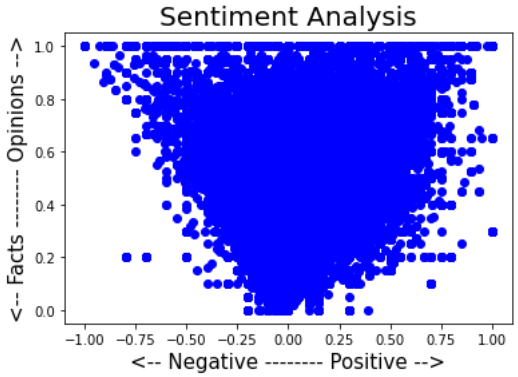

Figure 8: Sentiment Analysis on tweets

As it's shown above, it appears that the tweeters tend to use a more objective tone while writing tweets. Also, the sentiment of the text tends to be slightly positive. The distribution of the topics identified among the gathered tweets can be seen below.

We can see clearly that the **top concerns** were regarding the healthcare system and on how to protect yourself from not being infected.

3.2 Limitations

During the development of this project I encountered many limitations. One of them was at the data gathering step, due to the Twitter requests/day limit.

Topic_number	polarity_mean	polarity_median	subjectivity_mean	subjectivity_median
0	0.103931	0.068182	0.396551	0.427273
1	0.083943	0.000000	0.349420	0.369246
2	0.076440	0.000000	0.345287	0.350000
3	0.133546	0.075000	0.389484	0.424033
4	0.083103	0.000000	0.382259	0.400000

Figure 9: Mean/median on polarity and subjectivity

Topic_number	
0	30991
1	8726
2	7214
3	3488
4	21372

Figure 10: Topics distribution among gathered tweets for 7 days

Topic_number	
0	36050
1	9975
2	8291
3	4177
4	25339

Figure 11: Topics distribution among gathered tweets for 10 days

Due to this fact I decided to gather only 10,000 tweets/day.

Another limitation was regarded the large amount of time needed in the preprocessing step, more precisely in the step involving the correction of typos - due to checking each word, from each tweet. I used the textblob library, more precisley the correct method for this step which is based on Peter Norvig's "How to Write a Spelling Corrector" and it is about 70% accurate. In the future I will parallelize this step, in order to correct more tweets at the same time.

4 Conclusion

Users on Twitter tended to focus on the impact of coronavirus on people, businesses and countries. Besides that, they were also concerned about the healthcare system.

4.1 Further steps

I identified several directions in which this project can be developed. Some of them are:

- Compute interaction rate/topic
- Cleaning the data
- Compute bigrams, trigrams
- Hyperparameter tuning on the already appied models(topic number, number of passes, alpha, beta)
- Models: Testing other algorithms: a deep learning model eg: LSTM combined with LDA
- mixture of topics/tweet
- Gather more data

References

A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah. Top concerns of tweeters during the covid-19 pandemic: Infoveillance study. *J Med Internet Res*, 22(4):e19016, Apr 2020. ISSN 1438-8871. doi: 10.2196/19016. URL `http://www.jmir.org/2020/4/e19016/`.

(Abd-Alrazaq et al., 2020)