Exploring Tweets Related to the COVID-19 Disease

Malandrakis Eftychios-Angelos
student-id: 35363431

Abstract

The aim of this project is to explore and analyze data related to the 2019-2020 coronavirus pandemic. We have collected all of the twitter's daily top-tweets, related with coronavirus most common hashtags (313591 tweets) until the last week of April. We analyze the most commonly used words and hashtags per week and we implement a sentiment analysis to approximate the percentage of optimistic and pessimistic tweets per week.

1 Introduction

The coronavirus pandemic (or COVID-19) is an on-going pandemic of coronavirus disease, which was first identified in Wuhan, China in December 2019. As of May 2020, more than 4.5 million cases have been reported in more than 188, changing the everyday life for billions of people world-wide. In this study, we collect Twitter's top daily tweets and we analyze them per week. First, we explore the data for the most common words and hashtags per week, and how it changes through the time of our analysis. Then, we will run a sentiment analysis in our collected data, to explore the pessimism and optimism of the tweets. We will define optimists as the people who believe that future events are going to work out for the best. On the other side pessimists expect the worst [1]. As this is a difficult period for many people in the world, we will try to identify how people are feeling in terms of optimism and how this situation is changing, through the different stages of the disease.

2 Methodology

2.1 Data Collection

For the collection of our data, we have used python's library "GetOldTweets3". Twitter's official API, has two main limitations for the free accounts. It can return up to 300 tweets and the search is limited to the last 7 days. "GetOldTweets3", is a python implementation of twitters web application API. It can search for tweets with no time limitation and it can return up to 14000 tweets per request. For each tweet we get the variables of: user's username, to whom it replies, text, retweets, favorites, replies, id, permalink, author_id, date, hashtags, mentions, geo location, and urls.

For the purpose of this project, we collected all of the daily tweets that twitter has marked as "top tweets" from the first week of 2020 (30/12/2019-05/01/2020), until the last week of April (27/04/2020-03/05/2020). Twitter, defines "top tweets", as the most relevant tweets to a search, based on the popularity of a tweet (e.g., when a lot of people are interacting with or sharing via Retweets and replies), the keywords it contains, and other factors. The data were collected on May 10th, so that tweets made at the end of April, would still have some time to be considered as "top-tweets", from Twitter's algorithm. Twitter return's an error if we try to request more than 14000 tweets during a short period of time. To overcome this issue, the collecting algorithm waits for one minute each time it receives an error, and then it tries again. Then, tweets' texts and hashtags were stored in csv files per week, without implementing any pre-processing. In total, the number of the collected tweets is 313591.

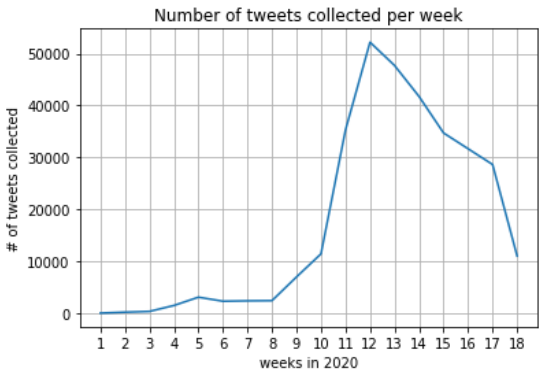

Figure 1: Number of tweets collected per week, from the first up to the 18th week of 2020.

2.2 Data Pre-processing

For the exploration of tweets' texts, the data were imported per week, then they were cleaned from mentions, hashtags, links and emojis. After that, each text has been processed through a sequence of fixes, using the ftfy python library. Then, each word in the text is recognized as token, it transformed to lowercase and stored in python "Counter" objects. Finally, the weekly "Counter" objects, containing tweets' text tokens of the week, were stored in a list.

Tweets' hashtags, were collected as a separate variable from each tweet, so we didn't implement any preprocessing, other than lowercasing the characters.

2.3 Data Exploration

In the data exploration process, first, we explore the most common words per week, after removing the stop-words and the most common hashtags per week. We also explore the most common words and hashtags for the whole period and how their appearances are changing during the weeks.

2.4 Finding Optimism and Pessimism

We run a sentiment analysis, to classify each tweet as optimistic or pessimistic. For this part of the process, we use the dataset published from Xianzhi et al., 2016 [2]. This dataset contains 7,500 tweets, annotated through amazon turk. Providing the annotators specific definitions of optimism and pessimist, each tweet was labeled by 5 annotators on a scale from -3 (very pessimistic) to 3 (very optimistic). Tweets with average annotation scores between -1 and 1 are considered "neutral". Tweets with scores greater than 1 or smaller than -1 are considered "pessimistic" and "optimistic" respectively [2].

We will preprocess the data by importing them as corpus objects. Each corpus object has 5 attributes. Object text, tokens frequency, part-of-speech tags list, part-of-speech tags frequency and the number of tokens. We will use the same pre-processing and tokenization process as with the data exploration.

We will train two different classification models. The first one will be a three-way classification model, where we will distinguish the data between pessimistic, neutral and optimistic. The second one will be a two-way classification model, where the data will be distinguished between optimistic and pessimistic. For both of them, we will try to classify them using three different classifiers, different numbers of tokens (from 50 to 5000) and where using also as a feature the function words frequency or the part-of-speech tags frequency [3].

Then we will predict the class of the collected tweets related to the coronavirus pandemic for both of these models, we will calculate the percentage of optimistic and pessimistic tweets per week and we will discuss the results.

3 Results

3.1 Data Exploration

The most common words used during the whole period, are: covid-19, people, cases, health. The table below, shows the 10 most common words and the times used.

	words	***# of times used***
1	covid-19	250211
2	people	35223
3	cases	29825
4	health	26729
5	now	25382
6	new	25016
7	coronavirus	22177
8	pandemic	22046
9	just	18006
10	help	17587

Table 1: Most common words used for the whole period.

The most common hashtags used, excluding the common coronavirus hashtags that we used, to collect the data (e.g. #coronavirus, #covid19, etc.), are #stayhome, #china, #stayathome, #breaking, #wuhan.

	hashtags	# of times used
1	#stayhome	2982
2	#china	2131
3	#stayathome	1814
4	#breaking	1711
5	#wuhan	1411
6	#coronaoutbreak	1323
7	#indiafightscorona	1255
8	#coronavirusupdate	1031
9	#lockdown	885
10	#corona	791

Table 2: Most common hashtags used during the whole period, excluding the coronavirus hashtags that we used, to collect the data.

Figure 2, represents the number of times that some interesting words used per week. Considering that we have collected the majority of the weekly top-tweets related to the coronavirus disease, we can see that there are some interesting insights.

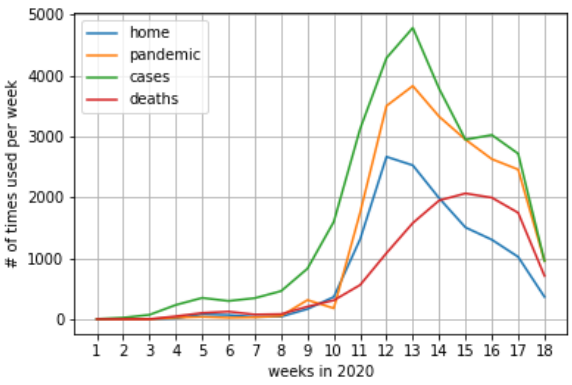

Figure 2: Some interesting words from the tweets' most commonly used words, with the number of times used per week.

The world health organization, officially recognized the spread of covid-19 as pandemic on 11 March 2020, week 11 [4]. Most of the European countries reached the highest percentage growth of new cases during week 13. Most of the countries announced the first social distancing measures during the first week of March (week 10).

In the next figure, we also present the number of times per week that some interesting hashtags appeared, during this period.

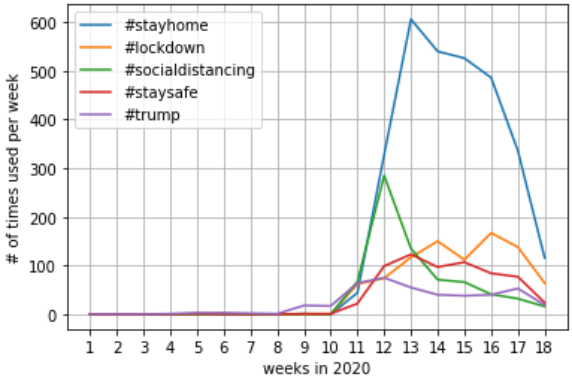

Figure 3: Some interesting hashtags, collected from the whole period most used hashtags, with the number of times used, per week.

It's also interesting to examine the number that some hashtags have been used per week, divided by the number of tweets that we collected during this week. From this graph, we can see that during the first two weeks of the year, almost one out of five tweets, contained the hashtag "#china". From the other side, the "#stayhome" and "#stayathome# hashtags (that we have treated as one) can be found in almost one out of forty during the weeks 13 to 16. Of course, we need to take into consideration that during these weeks we collected more than 150,000 tweets, while during the first weeks, we collected less than 100.

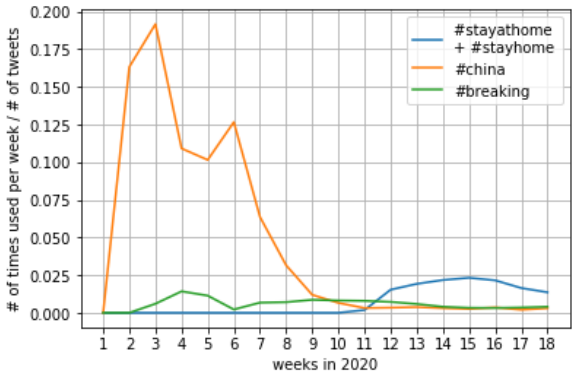

Figure 4: Number of times that the most common hashtags used per week, divided by the number of tweets collected during these weeks.

Weeks	*Most common words per week*
1	new, twitter, respiratory, viruses, ruled
2	new, china, novel, outbreak, chinese
3	china, new, novel, outbreak, cases
4	china, wuhan, cases, new, people
5	china, wuhan, people, outbreak, health
6	china, people, chinese, cases, hospital
7	covid-19, coronavirus, china, cases, virus
8	covid-19, cases, coronavirus, china, outbreak
9	covid-19, coronavirus, health, cases, people
10	covid-19, coronavirus, cases, people, health
11	covid-19, people, coronavirus, health, cases
12	covid-19, people, now, cases, health
13	covid-19, people, cases, now, health
14	covid-19, people, cases, now, pandemic
15	covid-19, people, pandemic, cases, new
16	covid-19, people, cases, new, pandemic
17	covid-19, people, new, cases, pandemic
18	covid-19, people, new, pandemic, cases

Table 3: Top tweets' most common words per week.

3.2 Finding Optimism and Pessimism

To classify our collected data as optimistic or pessimistic, we use two different classification models. In the first one, we build a three-way classification model, where we classify the data as optimistic, neutral, or pessimistic. In the second case we build a two-way classifier, where it classifies the data as optimistic or pessimistic. To build a good classification model for each case, we test the three commonly used classifiers, Naïve Bayes, Gradient Boosting and Logistic Regression. We also test the results for seven different numbers of features: [50, 100, 150, 500, 1000, 2500, 5000] and whether it's better or not, to use function-words or part-of-speech tags frequency as one more parameter.

For the three-way classification model, based on the F1 weighted score, the best classifier is Naïve Bayes, using 150 features and the part-of-speech tags. Using these parameters, it has 64% accuracy. The figure bellow represents model's confusion matrix. Even though, the accuracy is not very good, it's interesting that most of the false predicted data, were predicted as neutral. From the tweets annotated as pessimistic, only 11 out of the 197 were predicted as optimistic and from the tweets annotated as optimistic, only 8 out of the 395 were predicted as pessimistic. However, there also 295 tweets out of 903 annotated as neutral, that they were false predicted as pessimistic or optimistic.

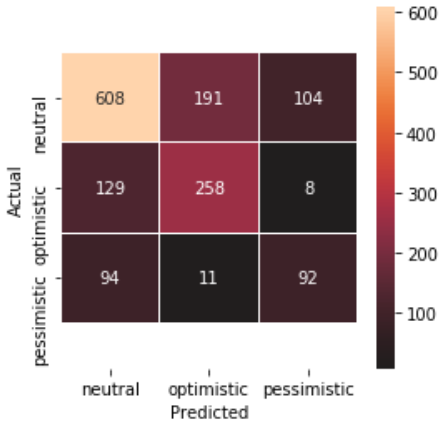

Figure 5: Confusion matrix of the three-way classification model.

Regarding the two-way classification model, where it classifies the data between optimistic and pessimistic, based on the F1 weighted score, the best classifier is the logistic regression, with 1000 features and the part-of-speech tags frequency. The overall accuracy of the classifier is 73%. From the confusion matrix of the model presented in the figure bellow, we can see that 156 tweets from the 837 annotated as optimistic, were predicted as pessimistic and 243 out of the 658 annotated as pessimistic, were predicted as optimistic.

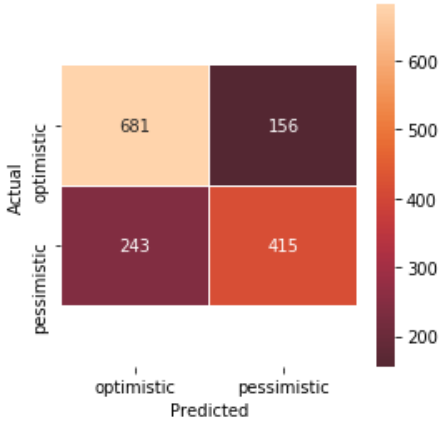

Figure 6: Two-way classification model confusion matrix.

Bellow we are presenting the results of the two models for the tweets collected. As the main purpose of this study is to look for either optimistic or pessimistic tweets, we will remove the number of tweets per week classified as neutral. Also, as we have collected different number of tweets for each week, it seems more sensible in this case to present our results as percentage of the total number of tweets collected per week.

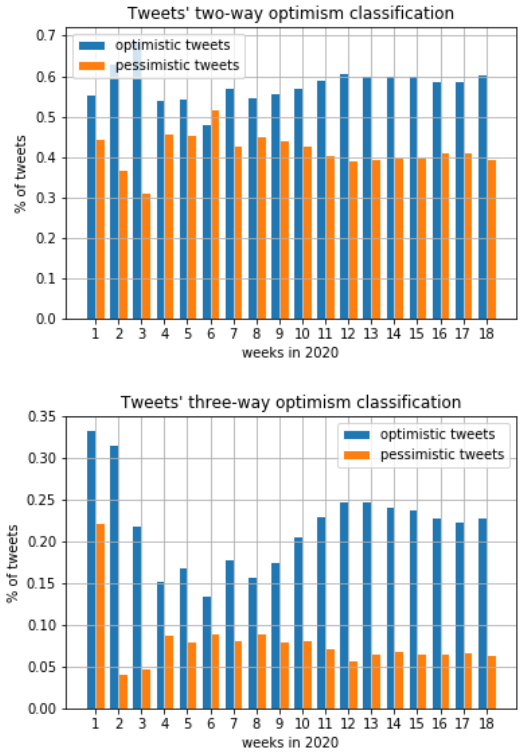

Figure 7: Two- and three-way optimism classification for the collected tweets.

About the two-way classification model, considering from the confusion matrix that there are way more data misclassified as optimistic, compared to the data misclassified as pessimistic, we would expect the percentage of the optimistic tweets to be greater than the actual one. On the other side, regarding the three-way classification model, we would expect the predicted results to be closer to the actual results, as the percentage of the tweets classified as optimistic and pessimistic, doesn't diverge very much from the actual value. It's Interesting however, that even though in the first case, the percentage of the classified data as pessimistic are much closer to the percentage of the data classified as pessimistic, in both cases, the data seem to follow a similar pattern. The main reason that in the second case the percentage of the pessimistic data are much smaller than the optimistic data percentage, is most probably because many data that previously considered as pessimistic, are now considered as neutral.

4 Conclusions

Regarding the data exploration, it's very interesting that the data match with specific events during the covid-19 timeline. The number of the "top-tweets" per week can be correlated with the spread of the virus, while they reach their peak value during the weeks that most of the countries implemented lockdown measures.

The sentiment analysis shows that both models have a similar pattern. During the first weeks, the percentage of the optimistic and pessimistic tweets is similar. After week 9, the percentage of optimistic tweets increases, until week 12. Then it slightly declines until week 17. Considering the confusion matrices for the two models, the three-way model seems to be more trustworthy for predicting percentage of optimism and pessimism, as the percentage of the tweets predicted in each class, is close to the actual number.

5 References

[1] Carver, "Optimism," 2010.

[2] Xianzhi R. et. al., "Finding Optimists and Pessimists on Twitter," in *54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016.

[3] Bengfort B. et. al., Applied Text Analysis with Python, O'REILLY, 2018.

[4] "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," 11 March 2020. [Online]. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.