Twitter sentiment and text analysis with stock market prediction applications during COVID-19 pandemic

Aysina Maria
Institutional ID: 486051
Email:
aysina.maria01@universitadipavia.it

Ferrari Luca
Institutional ID: 488039
Email:
luca.ferrari11@universitadipavia.it

Lazzarin Paolo
Institutional ID: 475081
Email:
paolo.lazzarin01@universitadipavia.it

Introduction

Twitter is one of the most popular social networking sites. Everyone can post a contribution, give feedback, express their own opinion and much more. For this reason, Twitter can be a good source for performing an analysis of individual cases/individual voices. The tweets are of course by no means a representation of the entire society, but they nevertheless provide a good opportunity to extract a subset of the voices in society. Also, social networks provide useful insights during crises. Especially during the Coronavirus pandemic social networks were one of the first platforms to share people's moods and opinions. In our project we analyse tweets from January, the 1st until August the 31th 2020.

Our analysis will take into account three different datasets: TripAdvisor Twitter data, Zoom Twitter data and COVID-19 Twitter data. We also integrated our main dataset with a financial one to verify one of the core hypotheses of our analysis. We decided to target TripAdivsor and Zoom because both of the companies experienced some major changes (both positive and negative) during the COVID-19 pandemic.

We look at three topics to have different insights and to understand how tweets changed with the progression of Covid. Our project has two main steps for getting better results of Twitter data analysis. First we analysed the textual part of tweets and then Twitter activities to discover some user insights. We divided users in two groups - public user and private user. By public users are meant accounts of the known persons/companies/media etc. and normal people by private users. In the second part we performed sentiment analysis on our Twitter data in order to build sentiment time series and to assess if any correlation existed between the target companies' stock market features and them.

I. DATA

A. Dataset 1: "Twitter data"

The main data for this project was data from Twitter. The time range of the data was 01.01.20 – 31.08.20. Three datasets with the following search terms were used: TripAdvisor, Zoom, Covid, each one containing the top 300 daily tweets in English. The following information was extracted for every tweet:

- The user who posted the tweet (column "User").
- The content of the tweet (column "Text").
- The date the tweet was posted at (column "Date").
- The amount of times the tweet was liked (column "Favorites").
- The amount of times the tweet was retweeted (column "Retweets").
- The mentions made in the tweet (column "Mentions").
- The hashtags assigned in the tweet (column "Hashtags").

B. Dataset 2: "Financial data"

For the stock market / sentiment analysis we retrieved freely available financial data from the popular and reliable Yahoo Finance website using the Yfinance open source Python library.

The tool gives us the possibility to create large datasets of financial data containing all the commonly used features for analysis by simply specifying the stock symbol and the time period we are interested in.

Data Science and Big Data Analytics Course – 2019/2020. Report of group WNAN. Submitted on 20/09/2020
Supervisor: Prof. Antonino Nocera

II. Exploratory Analysis

A. *"Twitter activities analysis"*

Hypothesis 1: "*Twitter data represents topics meaningfully*"

The most common words and word bigrams extracted from tweets are a representation of the meaning of the datasets topic and therefore can explain the main idea of the topic.

Hypothesis 2: *"The Covid data set has the most negative tweets"*

Covid pandemic has a lot of negative consequences, therefore we are expecting that Covid dataset will have the most negative words.

Hypothesis 3: *"Covid hashtags are the top hashtags during the all time period for all data"*

Covid – related hashtags are popular for all data during a wide range of the evaluated time period (especially beginning of March).

Hypothesis 4: *"Public users are top users in the data"*

The most active users are public users and have top score activity (favorites, retweets, mentions).

Hypothesis 5: *"Top data is from reliable sources"*

Tweets with most retweets and favorites are written by reliable sources and contain relevant information

B. *"Sentiment and stock analysis"*

Hypothesis 6 : "Twitter is a reliable data source for sentiment analysis"

Twitter is used every day by millions of people. Because of the great number of active users expressing their thoughts and the heterogeneity of them, we believe that Twitter can represent a robust and reliable source to catch trends in people's opinion through sentiment analysis.

Hypothesis 7 : "Correlation exists between Companies sentiment polarity TS and Companies stock market features"

People express opinions on companies every day. Comments or reviews written by influencers or accounts with big follower fan bases can have a strong impact on people's beliefs and actions. As we recently saw, financial markets can be strongly affected by social events. Assuming that Twitter data can be a reliable data source (HYP6) we think Twitter users' opinions can somewhat reflect the average sentiment on a company and therefore some correlation (or at least some kind of relationship) exists between companies stock market features and that.

Hypothesis 8: "Correlation exists between COVID-19 sentiment polarity TS and Companies stock market features"

COVID-19 has been the biggest event in 2020 and affected the world we live in many ways. According to what we stated in HYP7, we think some correlation exists between people's perception on COVID-19 and companies stock market features.

III. Methodology

A. *"Twitter activities analysis"*

Metric 1: "Text Analysis"

To reach precise results we first cleaned the data. First, we dropped all tweets with duplicate text from the data sets. Social network data has some noise caused by bots, advertisement, spam etc. Now search for suspicious hashtags in order to identify tweets that are not quite duplicate (i.e. where there are minor changes in the text, e.g. due to changes in a posted link), but are still similar enough, that their inclusion would distort our analysis. We therefore checked some of the most common hashtags, that looked suspicious like #python, #data and #bulk. As expected, some spam accounts were detected and therefore deleted. The by far most contaminated data set was the TripAdvisor data set (with lots of tweets that had to be removed). The Zoom and Covid data sets had far less data that had to be removed.

Next we cleaned the contents of the tweets. This part consisted of lowercasing the tweets and using regular expressions to remove links, numbers, punctuation, non – ASCII characters and other characters that would distort a further analysis. For the verifying of analysis we also used advanced text cleaning methods like lemmatization, bigramms extraction, counting of the most common words and sentiments (using TextBlob).

Metric 2: "Hashtags and mentions detection"
For the hashtags and mentions detection we used Python data wrangling techniques.

Metric 3: "Retweets and favorites detection"
For the retweets and favorites detection we used PyMongo data wrangling techniques.

B. *"Sentiment analysis"*

Metric 4: "Sentiment Features"

Verifying that Twitter is a reliable data source for sentiment analysis is quite a difficult task from the technical point of view. We need to define a strategy and a metric that help us to accomplish that. How can we measure the goodness of a tweet? The idea we had was to evaluate them by using a combination of the two main Sentiment features: Sentiment polarity and Sentiment subjectivity.
By extracting these sentiment scores for every single tweet in the dataset we were able to analyze their distributions and see how they relate to each other. In first approximation, we can consider tweets as good tweets if their subjectivity distributions aren't topic-dependent (similar shape and mean value).
The sentiment scores were extracted using TextBlob (which is an open source Python library designed for natural language processing) by running the tool on all the tweets of each dataset.

Metric 5: "Sentiment features time series "

After computing the two sentiment scores for each tweet, tweets were grouped by day and averaged (200 tweets per day). By doing this we could build a time series of estimated average daily sentiments that we could use to verify and analyze HYP6, HYP7 and HYP8.

C. "Financial metrics"

Metric 6: "Financial data features"

To verify the possible relationship between sentiment and companies stock market features we used different financial data features. The first one we took into account was the adjusted closing stock price which is one of the main values we need to analyze to evaluate a stock market performance. The second one we considered is the stock market return which is computed by using the adjusted closing price and they represent the returns that the investors generate out of the stock market. By predicting that it is possible to make wise investments.

IV. Analysis Design

A. Database

We decided to use MongoDB because it is an open source project database and fitted to our needs of a storage solution in a Big Data context. Mongo DB has also been considered because its queries' syntax can be easily implemented in Python using the PyMongo library.
The first thing we did was creating a MongoDB database and defining different collections, one for each dataset considered. The used datasets stored data in a text format so to make our job easier we used MongoDB casting functionalities to transform data into their correct data type. This choice was dictated by the ease of the instrument and because of the limited number of features.

B. Dimensioning the Problem

As a first approximation, we tried to conduct the analysis by following the standard "Pandas" pattern with no data structures dedicated to the Big Data scenario.
After testing the water, we imagined what could be the critical points that needed to be considered when designing our solution as scalable and suitable in a Big Data scenario:

- Big amount of Twitter data
- Possible Twitter data stream
- Big amount of financial data

To cover all these points, we think a possible hardware architecture we can implement would be composed by multiple nodes. The number of nodes really depend on the daily amount of data we are going to retrieve and the replication factor we decide to use, so making a good estimation of that will be very difficult.
Any cloud platform can be used for this purpose, but obviously if our hypothetical company already invested in physical servers and hardware we can implement a solution based on OpenStack by creating multiple virtual machines and by using the multiple networking and storage management services provided by the platform. For instance, a private network of VMs can be set up to create a distributed filesystem to use in combination with Hadoop and underneath PySpark.
Every single node will have a specific function (storing data, computing, orchestrating other nodes etc.) to deal with the various tasks, but we can also create duplicates to ensure reliability (if the resources let us do it) or, if more computing power is needed, add new hardware and therefore create new nodes in the virtual environment we previously set up.

C. Hadoop/Hive

We decided to use Hadoop to perform a MapReduce word count job. The text used is the result of the above-mentioned cleaning and lemmatization procedure.

- the mapper takes the file.txt as an input and removes the white spaces from the leading and the trailing. Then it splits the line into words to which is assigned a counter mark (1). The output of this program is a list of disordered and not grouped words that are fed to the reducer.
- the reducer uses the output of the mapper as an input and exploits a counter function to increment its values if a match between two words is present. After the counts, the function displays the top 50 words that appeared the most in descending order.

Even if our datasets cannot be considered "big", the solution presented with this simple word count could be scaled in a Big Data context using a multi-node approach, retaining efficiency and speed of calculation.

D. Spark

Spark is a very powerful tool we used in our project to deal with the entire pipeline. The libraries we exploited the most were the PySpark standard module and the PySpark SQL module. We used them to perform every single data transformation (tweets cleaning, data filtering etc.) and also for the actual analysis (calculating correlation etc.).
The major drawback of using PySpark for analyzing data is the fact that no good visualization module exists yet and that makes the process inconsistent and sometimes also computationally inefficient.
The only feasible solution to visualize PySpark data in-house right now is converting PySpark dataframes to Pandas dataframes and then use the usual visualization libraries with them (Matplotlib, Seaborn …). However, the PySpark to Pandas conversion is extremely time consuming so in a real world scenario this solution can't be adopted. Something that can be done, but we didn't adopt is using the DataBricks platform as a normal Python notebook.
The website offers a lot of services in terms of storage and cloud computing but most importantly let us run PySpark with a fast and custom visualization library making the whole analysis process smooth and consistent.

V. Analysis description and Results

Text Analysis

For precise text analysis results we also used some advanced techniques for text cleaning. First, we deleted all English stop words, which are defined in NLTK library together with some extra words related to Twitter slang (e.g. RT) or topic data. After that we lemmatized the text to simplify analysis of various lemmas related to the same base form. Then we started with text analysis. As expected, the most common words in each topic represent the meaning of the datasets topic *(Figure 1, 2, 3)*.

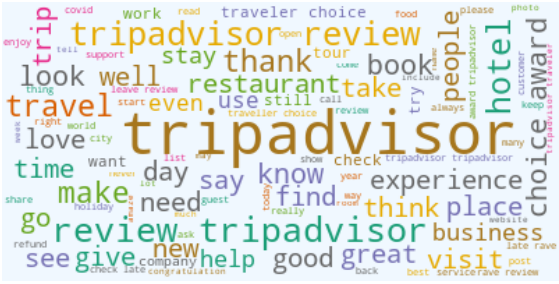


Figure 1

Figure 2

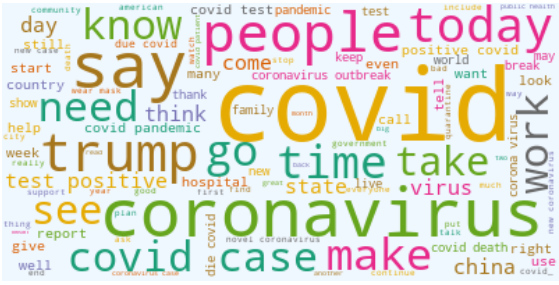


Figure 3

To gain more insights into the data we counted bigrams (i.e. co-occurring words) in the tweets. Identification of word co-occurrence also showed meaningful phrases, which are strictly related to each topic. For example, for the TripAdvisor data set the most common bigrams were ('review', 'tripadvisor'), ('choice', 'award'), ('traveler', 'choice'). The most common bigrams for the Zoom data set are ('zoom', 'call'), ('zoom', 'meeting'), ('use', 'zoom') and the most common Covid bigrams were ('covid', 'case'), ('test', 'positive'), ('covid', 'pandemic'). The frequencies provide information about the characteristic values of the examined text. Therefore, here appears a contextual embedding of a word in one sentence, as meaningful. This confirms the second part of the first hypothesis, that the most common bigrams are a representation of the meaning of the datasets topic. In this case the analysis of the most common words gives us a correct presentation of the topic.

Next we evaluated some basic statistics for words sentiments, such as the relative amount of positive tweets (0.84 for TripAdvisor data, 0.85 for Zoom data, 0.74 for Covid data). This evaluation confirmed our hypothesis that the Covid data set had the most negative tweets (relatively speaking). It is interesting to note that some words (like TripAdvisor, Zoom, Covid or also Trump) were heavily present both in negative tweets and in positive tweets, which can be explained by the fact that these words are used in different contexts. However, we can also detect clearly positive words like ”love”, “like”, “best”, “great” and clearly negative words like “bad” and ”death”.

Using appropriate hashtags is one of best strategies to increase the reach per tweet and increase the number of interactions. For the data collection phase we used the hashtags #tripadvisor, #zoom and #covid. As already mentioned, TripAdvisor data had a lot of issues due to the relatively high number of spam. The reason could be that #zoom and #covid were the most discussed topics during this time period and therefore spam data did not occur in the window of 200 tweets per day while collecting. We also extracted the most common hashtags for each dataset and analysed the change in popularity of the respective hashtags over time. Interestingly, we found that Covid related hashtags decreased starting in April in the TripAdvisor data and in Zoom data even in March. Hashtag #Travelerchoice from TripAdvisor increased since July while the policies in most of the countries were relaxed and borders were opened. Also the usage of #airbnb increased since June and #travel had no notable changes. Interestingly, all travel related hashtags decreased from approximately February/March till May/June, exactly for the time of Covid related restrictions in multiple countries *(Figure 4)*.

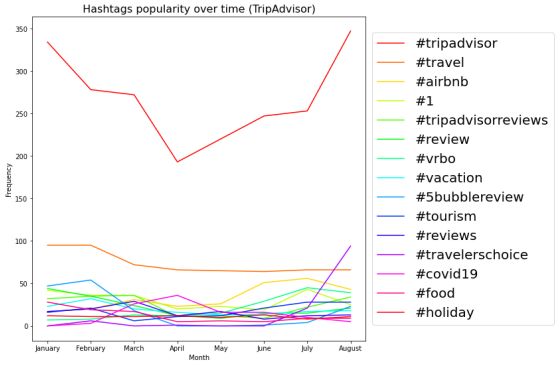


Figure 4

The highest rate of Covid related hashtags from Zoom dataset was in March and then it started to decrease. On the other side #zoom and education related hashtags increased in April, probably because of the start of E-Learning in this month. But after that time the popularity of these hashtags decreased. The highest rate of #remotework was in March,

although Covid was not yet a fully-fledged pandemic at that time and this hashtag was used before E-Learning hashtags (*Figure 5*).

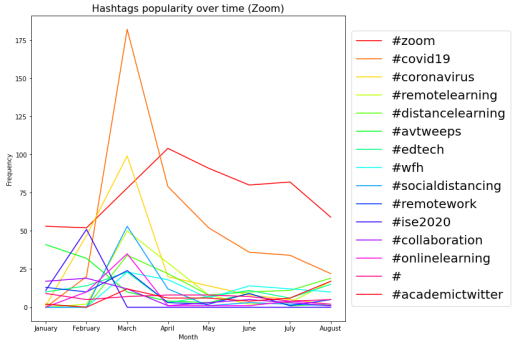


Figure 5

In the Covid hashtags picture, we can clearly see the change of the name of the pandemic. First, at the beginning of the virus only #coronavirus was used, then #covid19 started increasing property and later both hashtags strongly decreased in usage. On that point we can only confirm half of the hypothesis - Covid hashtag was extremely popular in March, but then the activity decreased in all datasets (*Figure 6*).

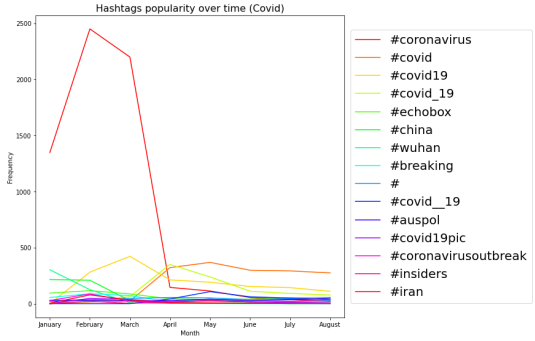


Figure 6

For verification of the next hypothesis we first analyzed mentions activities. The analysis of mentions shows us an overview of the accounts that stand in a relationship. The goal of the analysis was to detect most common mentions in order to understand which interaction between them could be discovered and how important are the public users (companies/ media). For better understanding we plotted most common Mentions as a network for each dataset (*Attachment*). Network is based on the users with the most retweets. We observed that TripAdvisor mentions are highly concentrated on companies. But it is possible to identify small groups of private users and the groups that have no relation to the companies accounts. The reason could be that Covid time had a negative influence on travel and TripAdvisor is part of that area. Therefore twitter activity was very low. In comparison, Zoom data is more partitioned in groups and there is no clear split in companies and private users. In this situation zoom popularity has increased significantly during the Covid time, therefore there is more communication through mentions between public and private users. Covid mentions have even more groups which are more mixed. Therefore according to Zoom and Covid data, there is an interaction between public Twitter accounts and private users. We can observe that a mention of interaction on Twitter depends on the context. This part is part of the hypothesis: "The most active users are public users and have top score activity (favorites, retweets, mentions)". At this point it can be verified that the Mentions activity depends on the context and can't be applied to every dataset. To verify the full hypothesis we did analysis of Retweets and Favorites with PyMongo which will be explained in the next part of analysis.

In order to gain more insight on the data and to verify furthermore our hypothesis, we performed queries on the databases with PyMongo. The analysis was performed in the same way on all three datasets and in order to improve the efficiency of the code we decided to put all the queries in a function. Because the data was collected on a period that ranges from January 2020 to August 2020, the analysis is performed both for the whole period and then for each single month. Firstly, we wanted to investigate if tweets with most retweets and favorites were written by reliable sources like newspapers, news TV channels and by famous people like politicians, celebrities and scientists. To achieve this goal, we performed a query that selected in each period the top three tweets sorted by Retweets and Favorites variables in descending order. Surprisingly, for all datasets and for each period almost all tweets with multiple Retweets and Favorites come from accounts of non-famous people or reliable sources (*Figure 7*).

```
# Covid Tweets' statistics on whole period
fun(corona,jan,sep)
```

```
Top 3 tweets by Favorites and Retweets
```

	User	Text	Date	Favorites	Retweets
0	HeavenSentMel	Due to Covid-19 I will not shake hands or hug ...	2020-07-08 23:51:38	155674	597335
1	the_real_bnell	The Office: Coronavirus Michael ignores the "w...	2020-03-09 23:55:59	99321	458562
2	Katlyn_Marie06	Meredith Grey & Cristina Yang would already ha...	2020-07-09 23:37:35	90502	322625

Figure 7 (Top three tweets from Covid19 dataset – whole period)

Another interesting aspect is inherent to the relevance of tweets. The top three tweets from the Covid dataset are much more ironic and have less serious content than the Zoom and Tripadvisor datasets. This trend has also occurred for every moment in time considered. This proves that Covid tweets are about a wider range of topics compared with the ones from TripAdvisor and Zoom datasets.
Then we proceeded to calculate other basic statistics like average number of Retweets and Favorites, the maximum and minimum in the considered period. Interestingly, we found that not necessarily the top three tweets are the ones that have the most Retweets or Favorites. In fact, the query on the top three is based on the joint count of the two variables and not taken individually.

```
Average of Favorites and Retweets
{'_id': 'Favorites', 'avg': 103.07878902389396}
{'_id': 'Retweets', 'avg': 359.48636707168185}


Maximum of Favorites and Retweets
{'_id': 'Favorites', 'max': 155674}
{'_id': 'Retweets', 'max': 1140270}


Minimum of Favorites and Retweets
{'_id': 'Favorites', 'min': 0}
{'_id': 'Retweets', 'min': 0}
```

Figure 8 (statistics on Covid19 dataset – whole period)

```
Average of Favorites and Retweets
{'_id': 'Favorites', 'avg': 1.997840838988279}
{'_id': 'Retweets', 'avg': 8.488102582180312}

Maximum of Favorites and Retweets
{'_id': 'Favorites', 'max': 2558}
{'_id': 'Retweets', 'max': 12374}

Minimum of Favorites and Retweets
{'_id': 'Favorites', 'min': 0}
{'_id': 'Retweets', 'min': 0}
```

Figure 9 (statistics on TripAdvisor dataset – whole period)

```
Average of Favorites and Retweets
{'_id': 'Favorites', 'avg': 46.380901879268656}
{'_id': 'Retweets', 'avg': 308.4507487379629}

Maximum of Favorites and Retweets
{'_id': 'Favorites', 'max': 199196}
{'_id': 'Retweets', 'max': 792962}

Minimum of Favorites and Retweets
{'_id': 'Favorites', 'min': 0}
{'_id': 'Retweets', 'min': 0}
```

Figure 10 (statistics on Zoom dataset – whole period)

From these statistics we can also see that Covid and Zoom top tweets have a higher number of Favorites and Retweets considering both variables and individually, confirming that during the considered period they were hot topics in contrast to TripAdvisor tweets. To seek further confirmation of this statement, we decided to calculate the total number of tweets that possessed Favorites, Retweets or both. As can be shown in the below figures, only half of TripAdvisor tweets have both features active while Covid and Zoom have both more than 90%.

```
Total number of Tweets, Retweets and Favorites
{'total_Tweets': 68804}
{'total_Retweets': 68568}
{'total_Favorites': 67774}

Tweets with at least both Retweets and Favorites
{'tweets_Ret_Fav': 67735}
```

Figure 11 (number of tweets with Retweets, Favorites or both in Covid dataset- whole period)

```
Total number of Tweets, Retweets and Favorites
{'total_Tweets': 22694}
{'total_Retweets': 13279}
{'total_Favorites': 6529}

Tweets with at least both Retweets and Favorites
{'tweets_Ret_Fav': 6016}
```

Figure 12 (number of tweets with Retweets, Favorites or both in TripAdvisor dataset- whole period)

```
Total number of Tweets, Retweets and Favorites
{'total_Tweets': 47146}
{'total_Retweets': 46007}
{'total_Favorites': 35448}

Tweets with at least both Retweets and Favorites
{'tweets_Ret_Fav': 35340}
```

Figure 13 (number of tweets with Retweets, Favorites or both in Zoom dataset- whole period)

Lastly, we wanted to check if top scorers are also the most active users in the considered period. The results of the analysis showed that most active users are not also top scorers. Moreover, the most active users of Covid dataset for each considered period are famous newspapers or TV channels. We can then assess that Covid topic is more frequently talked about by reliable sources while TripAdvisor and Zoom have a broader audience. Moreover, this can be also stated by most active user's counts.

Most active users

	_id	count
0	business	219
1	thehill	212
2	CNN	197
3	guardian	125
4	cnnphilippines	125

Figure 14 (most active users on Covid dataset – whole period)

Most active users

	_id	count
0	porjotonlipi	341
1	uvhr4u	135
2	StartupJobs	98
3	the_spherical	69
4	TheWanchHK	58

Figure 15 (most active users on TripAdvisor dataset – whole period)

Most active users

	_id	count
0	zoom_us	68
1	SneakerNews	66
2	SOLELINKS	55
3	KicksDeals	47
4	PolyCompany	43

Figure 16 (most active users on Zoom dataset – whole period)

Sentiment Analysis (introduction)

After being properly processed and filtered, our Twitter data is ready for further analysis. In this section we will investigate two different hypotheses: verify that Twitter can be considered and used as a reliable data source (in particular for Sentiment Analysis) and assess if correlation exists between sentiment polarity and companies stock market features.

To do that, we need to define a metric that accounts for two different aspects of an opinion: the subjectivity of the statement and the polarity of it.

The process of extracting these two pieces of information from text is called Sentiment Analysis. The standard way of approaching this problem would be gathering a big amount of labeled text (positive/negative, subjective/not subjective) and then train a classification model (usually Naïve Bayes is used) which is then able to provide the two scores we're interested in.

Since the goal in this section is getting a time series of sentiment values and not verifying what kind of words make a tweet considered positive or subjective, we will follow a different approach.

After researching on this topic, we decided a good tool to complete this task would have been TextBlob, which is a powerful natural language processing Python library that provides several useful features. The main one we're interested in is obviously the Sentiment Analysis one which is computationally fast and very effective for our purpose. We give as an input the tweet text to the sentiment extraction function and that returns a polarity value [$-1 \leq x \leq 1$] and a subjectivity value [$0 \leq x \leq 1$]; scores are self-explanatory.

Sentiment Analysis

As explained in previous sections (Metrics) we ran the tool on all the tweets of each available dataset. To make TextBlob suitable for BigData processing and to ensure scalability, we wrapped the sentiment extraction function in a PySpark object that let us use user-defined functions on PySpark DataFrames.

After doing that we computed the average daily sentiment time series for each dataset and plotted them (*Figure 17*).

The first hypothesis we wanted to investigate is about the goodness of Twitter as a data source for sentiment. The hypothesis can be easily verified by simply looking at the graphs we made. The plots showed us exactly what we expected:

- (Zoom) Since Zoom got a lot of attention during the first months of quarantine and made communicating possible with friends, co-workers and family members easier, we expected a growth of the sentiment polarity during this time (march/april).
- (Tripadvisor) During quarantine no travelling was possible in almost any country so what we expected was a negative or at least steady sentiment polarity value during this period and a growth during summertime, when moving to other places was possible and people where probably engaging more with Tripadvisor's platforms and talking more about vacations.
- (COVID-19) For COVID-19 data the only thing we expected was to observe a lower average sentiment polarity value compared to the other two datasets.

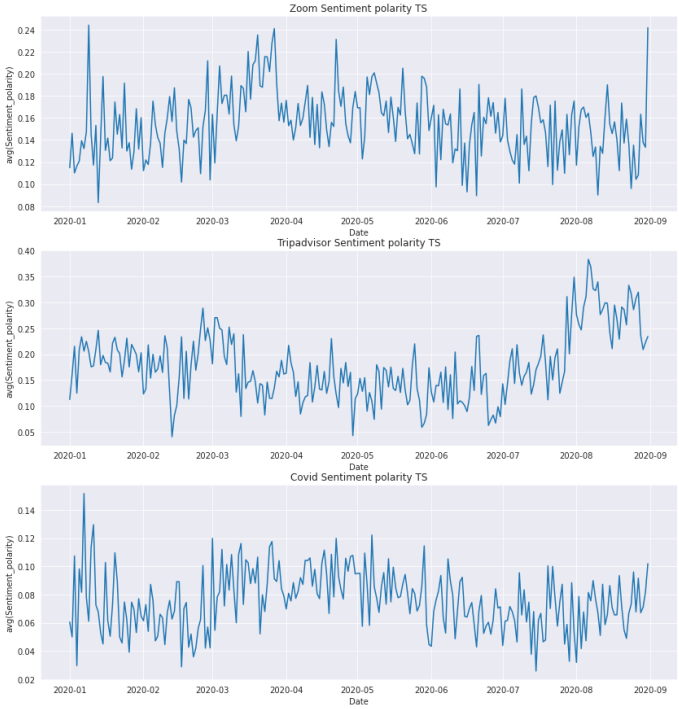


Figure 17

Even if these time series already provide good evidence for the hypothesis verification, we want to further demonstrate that providing also another way to visualize this data. By plotting the sentiment polarity distributions, we can see how the COVID-19 mean is lower than the other two and has a way lower variance (*Figure 18*).

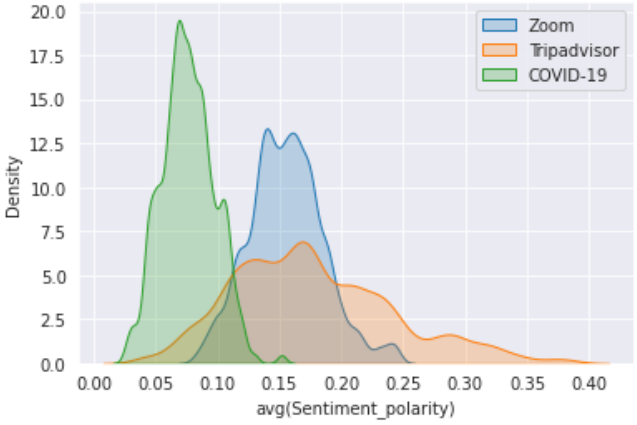


Figure 18

Moreover, the tweets' sentiment subjectivity values of all the datasets are mostly distributed around the same mean (*Figure 19*). That further strengthens the truthfulness of our hypothesis because a different mean in subjectivity would reflect into a topic-dependent sentiment bias in tweets, but luckily that doesn't happen. So, we can finally conclude that Twitter can be considered a good sentiment data source.

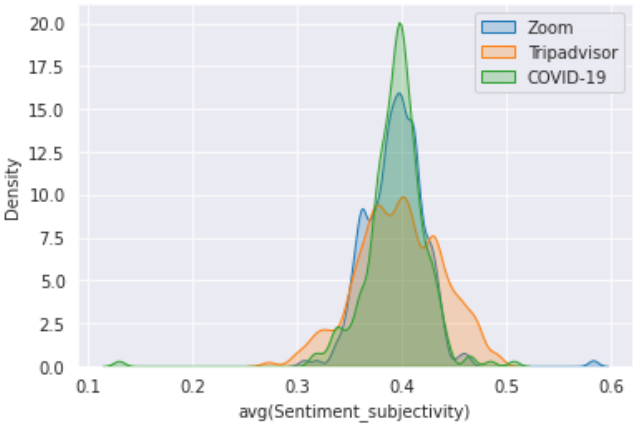


Figure 19

After verifying that hypothesis, we decided to continue our investigation on that path by looking if any correlation existed between the two sentiment features. Nothing interesting could be found by plotting the joint distribution, but then by plotting the data in a scatterplot and by computing the actual correlation for each dataset, we observed a peculiar behavior.

Medium strength correlation values $[\,0.4 \leq corr \leq 0.7\,]$ with a very small p-value were found for both the Zoom and the Tripadvisor datasets (*Zoom_corr*: 0.54; *Tripadvisor_corr*: 0.5). A weaker correlation was found for the COVID-19 dataset. (*Figure 20*)

Figure 20

After thinking about the results we got, we concluded that these values are probably the consequence of the fact that people tend to overstate when they are talking about personal experiences (maybe by using words that are considered particularly positive by our model) resulting also in higher scores for sentiment subjectivity. Obviously, we can't observe this behavior with COVID-19 data because as we saw in previous sections, most of the tweets are informative content that can't be subjective.

Stock Market / Sentiment Analysis

In this section of the analysis we are going to investigate stock market prices of our target companies and to assess if it is possible to observe some kind of relationship between that and the different time series we built previously.

Yfinance is an open source Python library that gives us the possibility to access financial data available for free on Yahoo Finance which is a very reliable platform for this kind of data. With Yfinance it is possible to retrieve data for every single company listed on the stock exchange just by providing the company symbol and the time period we are interested in. Yfinance then returns a dataset with a lot of different financial features. The one we are really interested in is the "Adjusted close price" which represents a more statistically stable version of the Closing price value (Cash value of the last traded price before the market closes). Without getting too technical, the Adjusted closing price is one of the features used to evaluate stock performance.

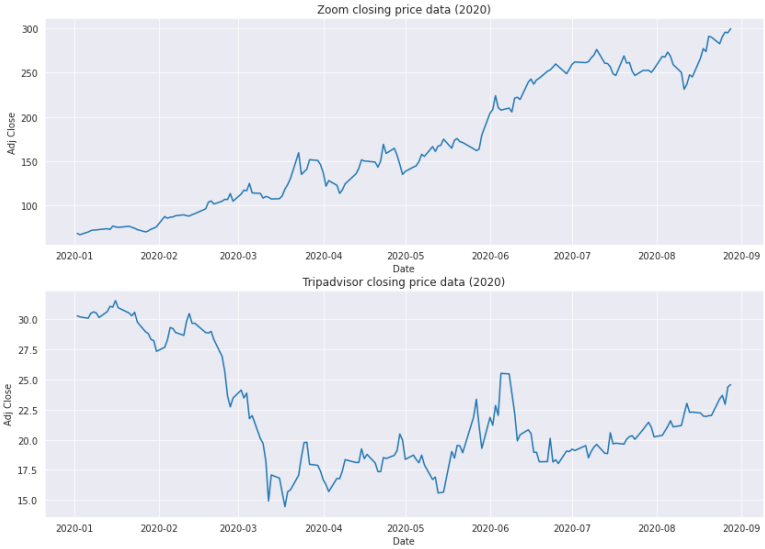


Figure 21

Figure 21 shows the stock market data (Adjusted closing price) for our target companies. As we could expect, Zoom stocks grew for the whole selected time period while Tripadvisor stocks collapsed after the big COVID-19 economic turbulence in the end of february. After downloading and visualizing our targets' stock data to verify that our initial beliefs were matching reality, we started the actual analysis.

As already explained, the final goal of this section is to check if there's a correlation or at least some kind of relationship between sentiment data and stock markets data. In practical terms, the ultimate goal would be to

build a model that by taking into account sentiment data and other data sources can predict stock market prices. Even if predicting the closing prices of a certain stock would be extremely useful for an investor, what he/she would be really interested in is predicting the return value of that stock in a specific day. By having a good estimate of the daily return on the investment, an investor can follow a data-driven approach and spend his/her money wisely.

To verify our hypothesis the first thing we did was computing the returns on our data. Since we wanted our solution to be completely scalable (maybe taking into account multiple years-worth of data in the future) we used the PySpark facilities also in this case. Once we had the returns time series for every company, we needed to compare them to the previously created average daily sentiment polarity time series.

Both a visual inspection of the lineplots and the scatterplots didn't suggest anything interesting or usable. A formal correlation computation confirmed that no relationship at all existed between sentiment polarity time series (both COVID-19 and target companies related) and the companies returns (*corr ~ 0.01*).
(*Figures 22, 23*)

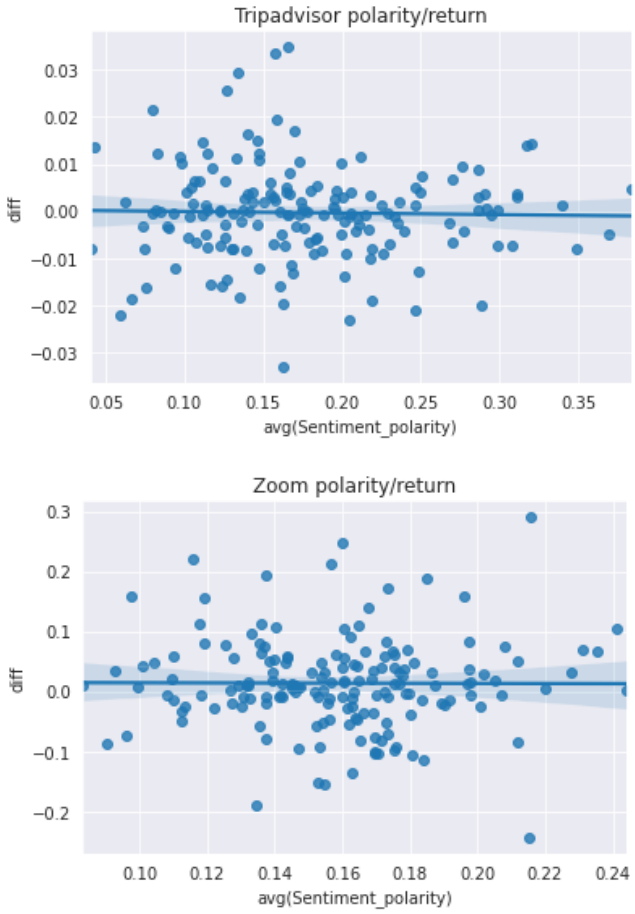


Figure 22

Since this approach didn't provide any useful insights on the analyzed data, we decided to adopt a new strategy. Instead of focusing on the almost impossible task of predicting the stock returns, we decided to see if any relationship existed between the sentiment polarity time series and the Adjusted closing price time series.

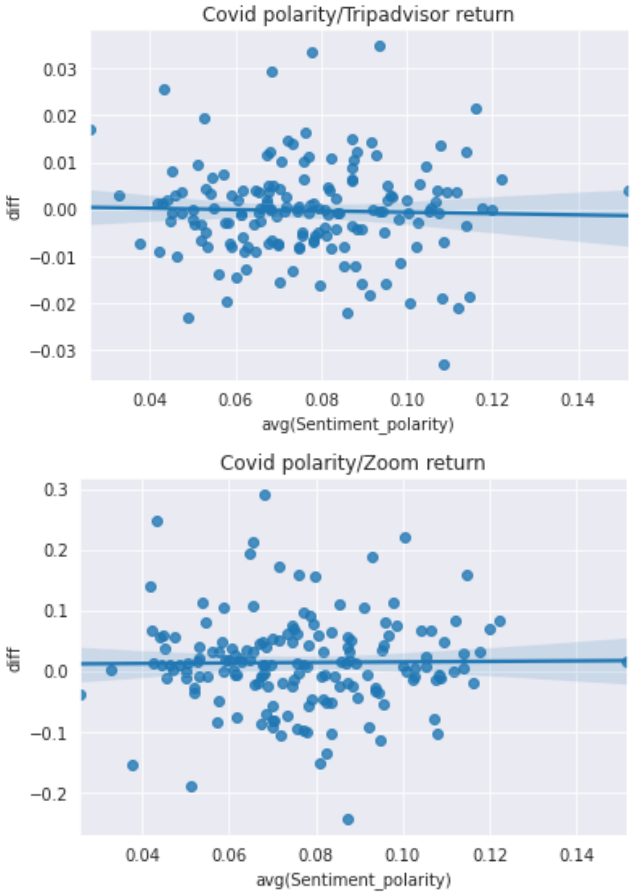
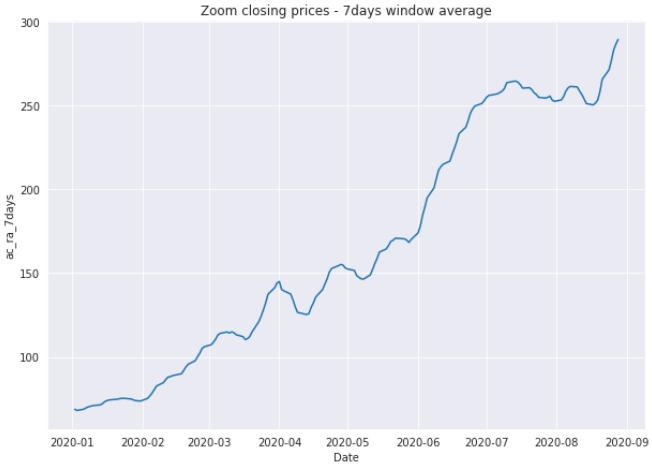


Figure 23

Both the time series can have a lot of uninformative variability inside of them, so to remove that we adopted a common signal processing technique which is called low-pass filtering or rolling average. By applying this mathematical tool, we could remove a lot of noise (unwanted variability) preserving just the signal part (informative variability). To do that, we used the Window object from the PySpark SQL library that gives us the possibility to select a data window (time window in this case) where a certain function is applied (*avg()* in this case). The time window we chose is a seven day time window (seven days is a good compromise in removing noise and preserving information trade-off). *Figure 24* shows an example of filtered data through rolling average.

Figure 24



After filtering the time series from all the datasets we could finally see if sentiment polarity and closing prices are somewhat correlated or linked by some kind of relationship.

First we computed the correlations scores between on all the possible filtered sentiment time series and filtered closing prices and run a chi-squared Pearson test from the Scipy Stats library (No chi-squared Pearson test available in PySpark) to verify that the correlation value found wasn't generated by random fluctuations in the data.
(The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets.)

Figure 25

Since the correlation scores we found were pretty interesting we plotted the data in multiple scatterplots to visually verify that the linear relationship the correlation was promising was real. (*Figure 25*)

Unfortunately the correlation we found wasn't representative of the relationship between the two features we took into account (hypothesis not verified), however something extremely interesting can still be observed.
By carefully inspecting the scatterplots we can notice a peculiar pattern recurring in all the datasets: a region (left side of the plot) where points are clustered in two different groups and another region (right side of the plot) where points are clustered in a single group.

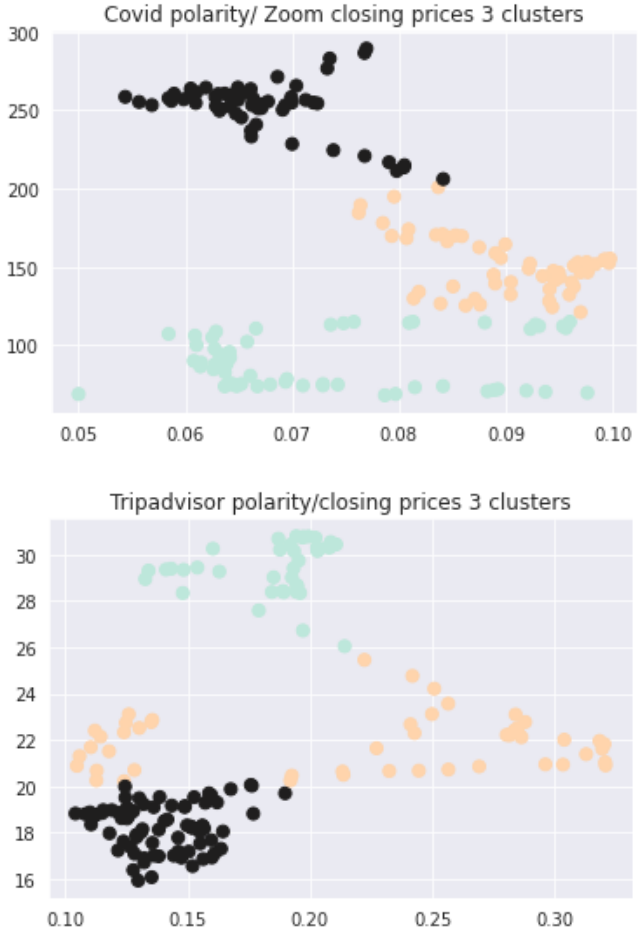


Figure 26

To make it more visible and to further investigate our intuition, we ran a K-means clustering algorithm from the SciKit-Learn Cluster library on the datasets. The results showed us the clustering we expected to find (*Figure 26*).
Unfortunately we can't use this model to make good predictions, but it can help us to acknowledge that some non-linear relationship exists between sentiment polarities and closing stock prices. The next step would be extracting the underlying relationship by training some kind of non-linear model that can represent that and then use it to exploit the new information. After trying some kernel-based transformations and some other feature engineering techniques, no solution could be found (apart using a Gaussian Kernel that basically projects features into a infinite-dimensional space producing a obvious overfitting) so we concluded that the only feasible solution would be training some kind of artificial neural

network for regression which is out of the scope of the analysis.

VI. CONCLUSIONS

- Based on most common words analysis and extracted bigrams, Twitter discussions represent topics meaningfully.
- As expected, Covid data sentiment is more negative with respect to the one of the other datasets.
- Usage of #Corona hashtag was extremely popular only in the beginning of pandemic and then the activity of this hashtag decreased.
- Activity of users depends on the topic. The more popular the topic is, the more private users participate in the discussion and show the activity through likes, retweets and mentions (Zoom, Covid). The more inactive the topic is, the more only public users have a twitter activity. An example is TripAdvisor, which was inactive during the Covid restrictions in most of the countries.
- Public users are not the top active users (except for Covid dataset in which newspapers are top scorers, because of the posting news on twitter). On the other hand public users don't have top score activities (favorites, retweets and mentions).
- Twitter is a good and reliable source of data for sentiment analysis.
- A non-linear relationship exists between sentiment time series and the companies' stock market features but some further analysis and complex modelling (building a regression artificial neural network) are required.

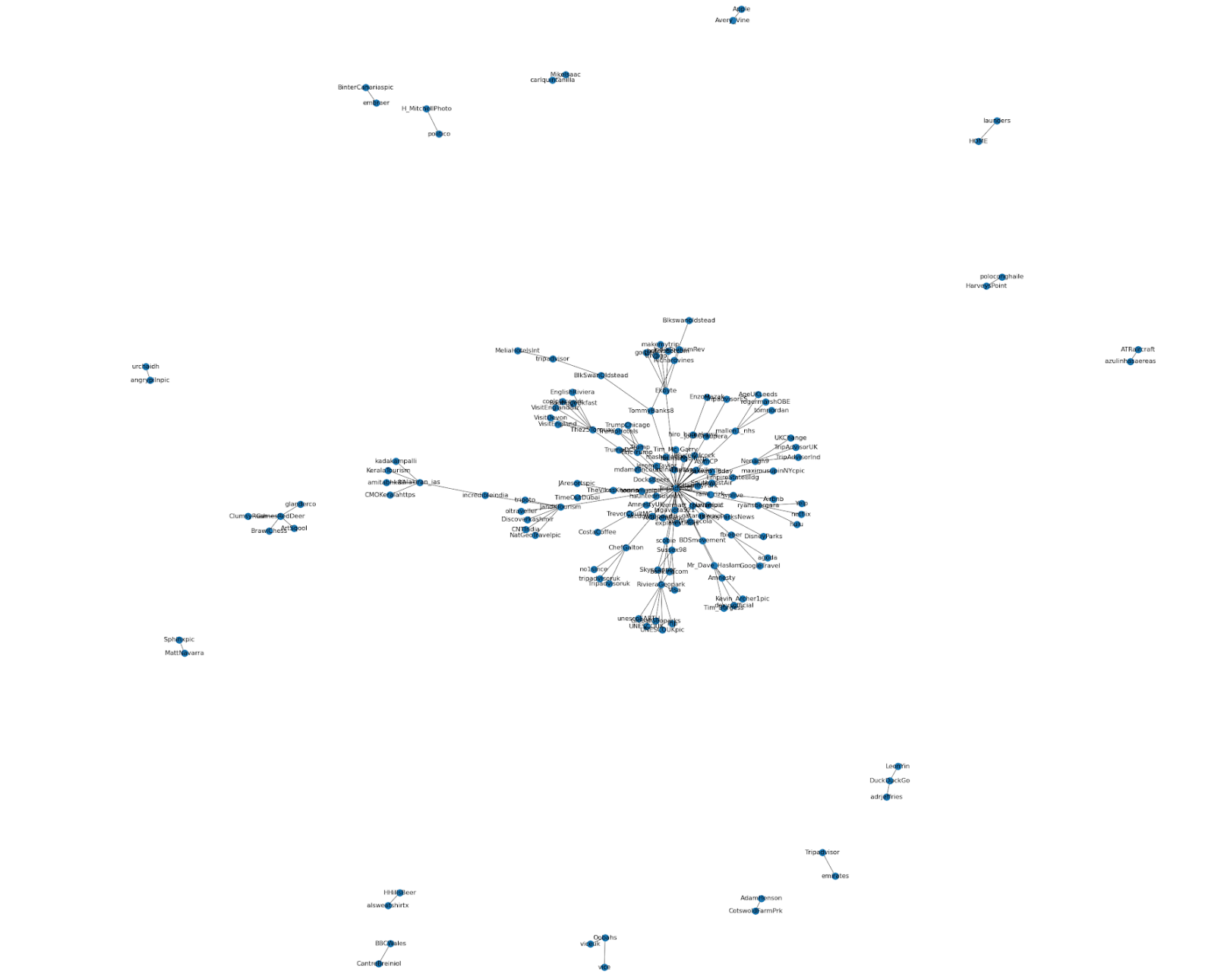
Open points:

- Do a real time analysis on real-time Twitter stream.
- Building and training a regression artificial neural network to extract the underlying non-linear relationship between sentiment time series and companies' stock market features.

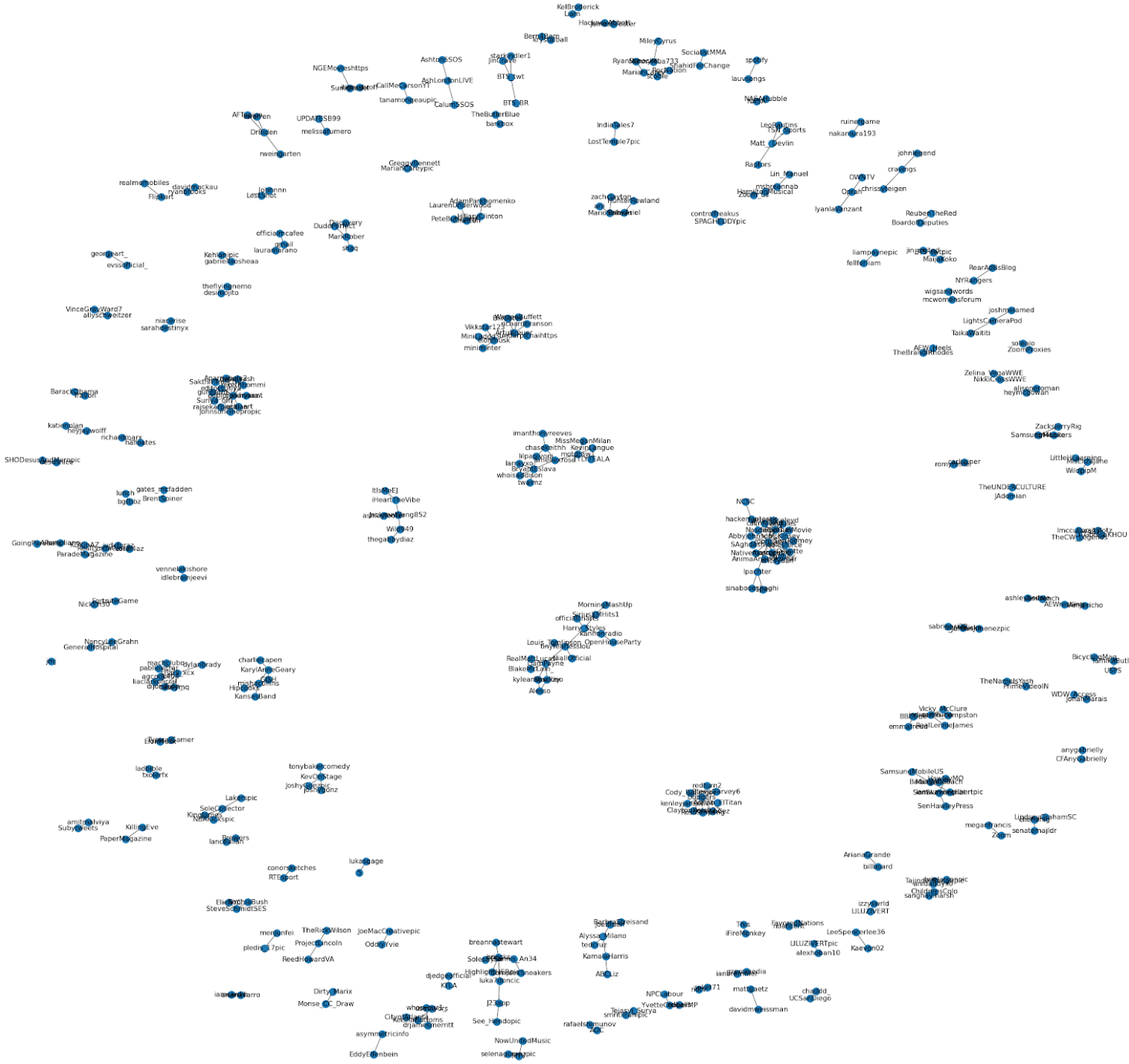
VII. REFERENCES

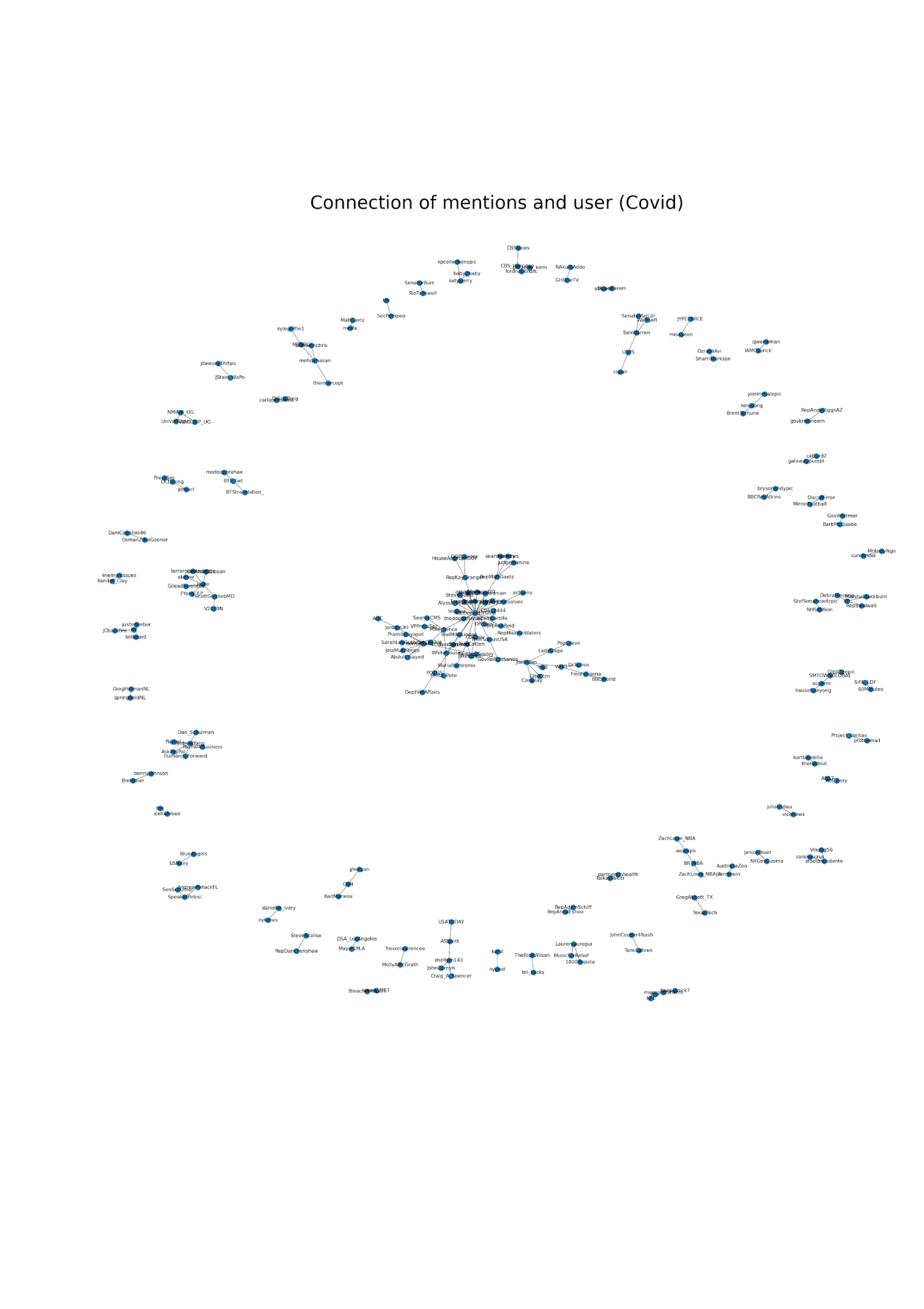
1. Anna Kruspe, Matthias Häberle, Iona Kuhn, Xiao Xiang Zhu: *Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic*, 2020
2. Jia Xue, Junxiang Chen, Ran Hu: *Twitter discussions and emotions about COVID-19 pandemic: a machine learning, 2020*
3. Kamaran H. Manguri, Pshko R. Mohammed Amin, Rebaz N. Ramadhan*:* *Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks, 2020*
4. Jim Samuel, G. G. Md. Nawaz Ali1 , Md. Mokhlesur Rahman, Ek Esawi1,and Yana Samuel: *COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification, 2020*

Connection of mentions and user (TripAdvisor)



Connection of mentions and user (Zoom)




Connection of mentions and user (Covid)