Public perception about vaccination for Covid-19: Social Content and Network analysis.

Francesca De Cola (819343) - Valentina Moretto (853744) - Alessandro Motta (812309)

Abstract

This historical moment is characterized by the pandemic that all the world has been involved in. All human life runs around the Covid-19, and all the aspects of human life are affected: education, politics, economy, healthcare. One of the possible solutions to fight this situation is to find a vaccine that protects people from the virus, and as expected nowadays it's one of the most discussed topics. We are witnessing a race to discover an effective vaccine. All states, companies and international organizations are involved because the consequences could change the fortunes. As usual, around a hot topic, opinions of people are different, if on the one hand someone agrees with an idea, one the other hand there's the other half that doesn't agree with. This article aimed to investigate the sentiments of the Twitter's user about this hot topic during the month of December 2020 and the analysis of the network that they generated.

Keywords

Social Content Analysis — Social Network Analysis — Sentiment Analysis — Community detection — Covid19 — Vaccination

Contents

Introduction

More than 2 million people have died for Covid 19 worldwide since the start of the pandemic, which sprang from Wuhan, China, in the first days of 2020. The situation is set to stop with the arrival of vaccines from different pharmaceutical companies in the world, but the problem that arises in today's society is that many people, due to bad information and ignorance in scientific matters, refuse to get vaccinated. This problem causes, in addition to social unease, also the opinion that if a considerable part of people do not get vaccinated, the desired herd immunity cannot be reached, allowing the virus to circulate and change, leading to the formation of new variants. In this context, vaccine misinformation on social media represents a significant and growing public health challenge. This work aims to disambiguate some questions: is the majority of the people willing to get vaccinated or not? Are people sure about the effectiveness? How have opinions changed during these days? What are the events and the actors that influenced people the most? How does politics and news spread influence the situation? To answer, the starting point are the opinions of the Twitter's user and the communities that they have generated through the network. This analysis was carried out through the investigation of approximately 270,000 tweets related to the vaccine collected from the 5 of December to the 21 of December 2020. The chosen period is essential to have a clear view about the sentiment, the events and the actors regarding the chosen topic in the last month of 2020 when the first vaccination campaign started. Furthermore, is crucial the research of the communities of twitter's users that were formed around a certain topic in order to be able to analyze it and find insights that would be helpful to understand how people confronted each other around such an important issue. In order to perform these tasks have been developed a Social Content Analysis and a Social Network Analysis that will be shown in details later in the paper.

1. Dataset

There are several social media where you can discuss and be able to declare your opinion and experiences about it. For this research was considered the Twitter social network for several reasons. The first is that is one of the few social media that has decided to implement security policies to prevent the spread of bad information about vaccines by monitoring and deleting all tweets that generate bad information. Secondly, Twitter makes its API available to developers, which through Python libraries allow streaming and listening to all posts made by users of the social network filtered for one or more hashtags of interest. Thirdly, Twitter was chosen as it focuses on keywords (hashtags) and allows people to post to a very large audience compared to other channels like Facebook. Moreover, it allows users to communicate directly in real time. Thus, reports on what is going on during an event as the incident unfolds. The API has certain limitations including:

- Only access tweets from the last 6-9 days;
- Only request 18,000 tweets in one call: you can stream tweets and collect them using ongoing protocols however there are limitations to how much data can collect.

Having considered all these aspects, was used the twitter API thanks to the Python library called Tweepy. In order to use this library it was first necessary to create a developer account to be able to use the API. For each tweets was downloaded:

- The unique code for each user;
- The date of publication of the tweets;
- The user's username;
- The text of the tweet;
- The place from where the tweet was written;
- If the tweet is a retweet or not.

Furthermore, it was decided to only consider tweets written in English originated from any nation in the world but containing hashtags regarding the Covid vaccine. In addition to that was also taken into account the tweets that spoke about the pharmaceutical companies running for the production of the vaccine. The following hashtags were used: #vaccine, #vaccines, #Pfizer, #Astrazeneca, #Jhonson&Jhonson, #Moderna, #Novavax, #Vaccinate, #covidvaccine. Subsequently, again thanks to some libraries available in python (KafkaProducer, KafkaConsumer), it was possible to resort to the use of Apache Kafka, an open source stream processing platform. This technology uses applications such as producer and consumer which respectively write and read streams of records on a particular structure called topic. The tweets obtained in Json format were processed through the MongoDB software. MongoDB is in fact a document-oriented non-relational DBMS, which moves away from the traditional table-based structure of relational databases, in favor of documents in JSON format. Then, thanks to the PyMongo library in Python, it was possible to directly access the database and the collection in MONGODB. This has been converted to a CSV file to allow for more immediate integration and manipulation.

1.1 Preprocessing

Once that was found the final dataset "df_tweet_clean.csv" is time to investigate the content. First, the retweets were eliminated because there was the interest in the first opinion of users and the will to discover interaction between users represented by mentions or hashtags. In a second time, were eliminated rows that contained null values in the columns username and text. Then was analyzed specifically who are the users that write tweets, which users tweets the most and where these users came from. It was discovered that there were a lot of users bot, therefore later there is the explanation of how those rows were eliminated using specific tools. Now, it's appropriate to make a special focus on the location column: this field it's an optional field that users can fill or not. The situation showed that there are 3 scenarios:

- The field is empty;
- The field is completely random (e.g. 'my house','moon');
- The current location.

Then the focus was moved onto the general analysis of tweets. Here was performed a first phase of cleaning and application of some text mining techniques like stop word removal, url removal, lowercase conversion, emoticons removal, tokenization to better understand and clean the content. Some basic analysis were carried out like seeing the most used words, hashtags, mentions and days where were a lot of activity. The most frequently mentioned users were (Figure 1):

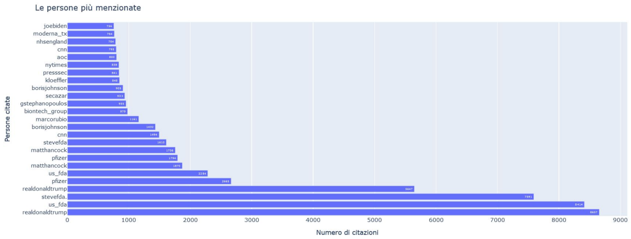


Figure 1. Most Mentioned Users

- @RealdonaldTrump who is an American politician who was the 45th president of the United States;
- @us_fda or United States Food and Drug Administration (FDA or USFDA) that is a federal agency of the Department of Health and Human Services;
- @SteveFDA who is the 24th Commissioner of Food and Drugs of USFDA;

- @pzifer who is an American multinational pharmaceutical corporation that in May 2020 began testing four different COVID-19 vaccine variations;
- @matthancock serving as Secretary of State for Health and Social Care.

The most frequently used hashtags excluding those used for the collection were (Figure 2):

- #nhs that is the National Health Service of England;
- #operationwarpspeed is a public private partnership initiated by the U.S. government to facilitate and accelerate the development, manufacturing, and distribution of COVID-19 vaccines, therapeutics, and diagnostics;
- #BREAKING;
- #learntherisk;
- #cdnpoli relating to Canadian Politics;
- #trumpisacompletefailure.

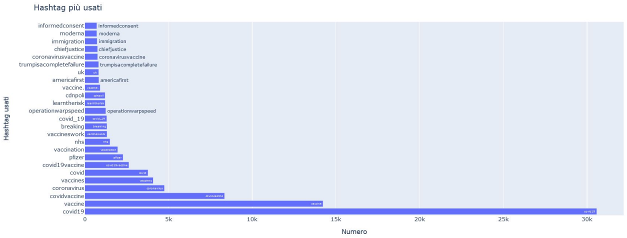


Figure 2. Most Used Hashtags

For this general analysis was decided to maintain all the dataset to better capture the characteristics of the data but later the focus moved on a specific location that's the United Kingdom in order to better answer the research questions. For this reason were filtered out data that can be traced back to this location by the application of a filter on the location to obtain all the rows that contained terms such as ”uk”, ”london”, ”england”, ”united kingdom”. There was the awareness that in this way were surely lost some tweets and these could represent useful information. But on the other hand this was the only way to filter information effectively and obtain a clearer dataset.

1.2 Bot Analysis

After the basic preprocessing was implemented a solution that would allow to get more quality data. In fact, analyzing the tweets it was noticed how many seemed repetitive and out of context despite having hashtags in reference to the covid vaccine. So, was decided to search and delete from tweet's dataset all the posts generated by so-called BOT accounts. Twitter bots are programs that compose and post tweets without human intervention, and they range widely in complexity. In python, was used the library called Botometer (formerly BotOrNot) to know if a particular tweet was made by a bot or not. The Botometer library uses a machine learning algorithm trained on tens of thousands of labeled data. The output is a probability on a scale of 0 to 1, where 1 indicates that a twitter account is managed by a bot, 0 otherwise. Botometer uses its API to help identify bots. The APIs are free if you sign up for an account that allows you to analyze up to 500 accounts every 24 hours. Were analyzed 14,239 Twitter accounts (500 per day) who had written at least one tweet regarding the covid vaccine. Were only considered accounts that were found to have a greater than 0.5 probability of being BOTs. The result is that of the total accounts considered about 9500 turned out to be potential BOTs, for a total of 32,160 tweets. Bot research focuses on understanding whether these automated accounts write tweets that have positive, negative or neutral sentiment. This is to have a general idea if the Bots have the goal of being noVax or proVax. For a better comprehension was produced a WordCloud on the tweets written by the Bots. This to get an idea of the content of the tweets. As can be seen from the Figure 3,

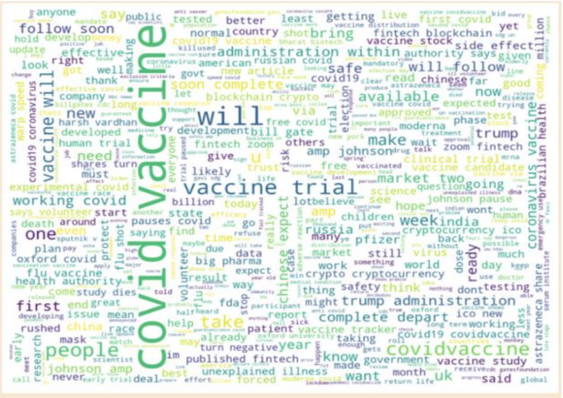


Figure 3. WordCloud of Bot's Tweets

the result does not seem to show a clear stance by the BOTs, whose words seem to be very generic and neutral.

2. Social Network Analysis

Talking about social media and particularly for Twitter, what is most interesting is that it creates connections; this connections forms Social Networks that can be studied to understand how users interact or how opinions and topics get spread. In order to answer that, were used the properties of graph theory, indeed Graph is a data structure used to represent and analyze connections between elements. The two elements that a graph needs are nodes, called vertices too, and edges that represent the connections between nodes. These structures can be undirected, when the edges doesn't have an orientation and directed, like in this case, where the edges have an orientation. Furthermore they can be bidirectional, like in Facebook, where A is a friend of B and viceversa or monodirectional, like in Tweeter, where if A follows or comment a post of B it's not necessarily true that B do the same.

2.1 Graph building and analysis

According to the purpose of this analysis, users represent the nodes and when the system finds an interaction between them an edge will be created to connect both nodes. For the development was used a Python library called NetworkX, a R-studio library called iGraph and Gephi that is an open source software for the visualization and the exploration of graphs. From the dataset were considered:

- author of the tweet;
- twitter's user @mentions in the text.

First of all is necessary to extract the mentions in order to create a different dataframe:

- User, Splitted @Mentions;

Is known that the mention, thanks to its structure, is preceded by the @, so was extracted the name directly from the text of the tweet using a regular expression. Once the necessary information has been obtained, we initialize the graph. In order to better understand and answer the research questions, were selected some key mentions:

- Directed graph of the entire dataset for user-mentions;
- Directed graph of @BorisJohnson, Prime Minister of Britain vs @MattHancock serving as Secretary of State for Health and Social Care;
- Directed graph for @SkyNews vs @CNN vs @bbcnews for the spread of news and information.

In a monodirectional social network the degree of a node can be further splitted in out degree of a vertex v, that is the number of edges starting from vertex v (outgoing edges), and in degree of a vertex v, that is the number of edges arriving at vertex v (incoming edges). An Hub is a node that has an high value of out degree and that is connected to nodes of small degree. Authorities are the opposite of Hubs. They represent nodes with an high in degree connected to nodes with a small degree. The coefficient of Assortativity assumes positive values when Hubs show a tendency to link to each other (Assortative Network), or negative values when Hubs tends to avoid linking to each other (Disassortative Network). All these measures will be explained in details for the three graphs generated.

2.2 Community Detection

One of the most important tasks when studying networks is that of identifying communities. They allow to discover groups of interacting objects and the relations between them. After some tests the method that brings the highest value in terms of Modularity was the Louvain Method implemented in Gephi and iGraph described below.

2.2.1 Louvain Method

Louvain is a method to extract communities from large networks. Is a greedy optimization approach that attempts to optimize the "Modularity" of a partition of the network. The optimization is performed in two steps. First, the method looks for "small" communities by optimizing Modularity locally. Second, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of Modularity is attained and a hierarchy of communities is produced. The Modularity measure quantifies the quality of an assignment of nodes to communities. This means evaluating how much more densely connected the nodes within a community are, compared to how connected they would be in a random network. The result will be explained in details for the three graph generated.

2.3 UK Directed Graph

In Figure 4 there is the graph generated by the entire dataset of tweet in UK, where the nodes are proportional to the in-degree and the colors are the partitions of the community detection. The layout used for the visualization is the 'Force Atlas 2' algorithm that places the nodes inside the graphic space and is able to handle large networks while keeping a very good quality. In details:

- Number of nodes 18279;
- Number of edges 20022;
- Average Degree 1.09;
- Modularity 0.86;
- Assortativity -0.06;
- Number of communities 2610.

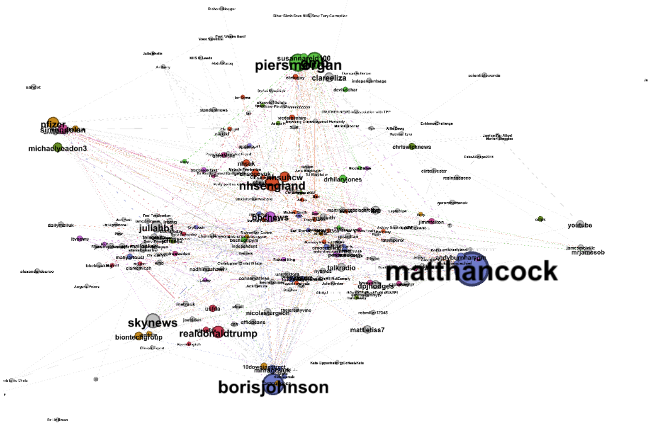


Figure 4. UK Directed Graph filtered by node degree

This high value of communities is due to the fact that there are a lot of sparse nodes.

2.4 Politic Directed Graph

In order to answer the question of how does the politic affects the network was generated the graph in Figure 5 where the nodes are proportional to the in-degree and the colors are the partitions of the community detection. The layout used for the visualization is the 'Yifan Hu' that is a very fast algorithm with a good quality on large graphs In details:

- Number of nodes 1397;
- Number of edges 2112;
- Average Degree 1.51;
- Modularity 0.64;
- Assortativity -0.24;
- Number of communities 53.

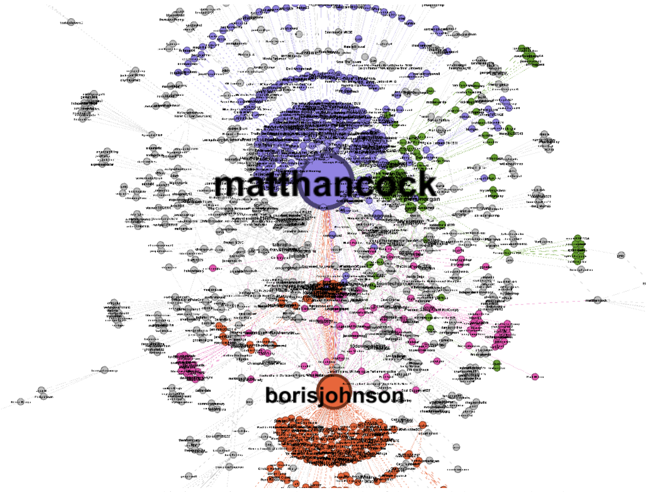
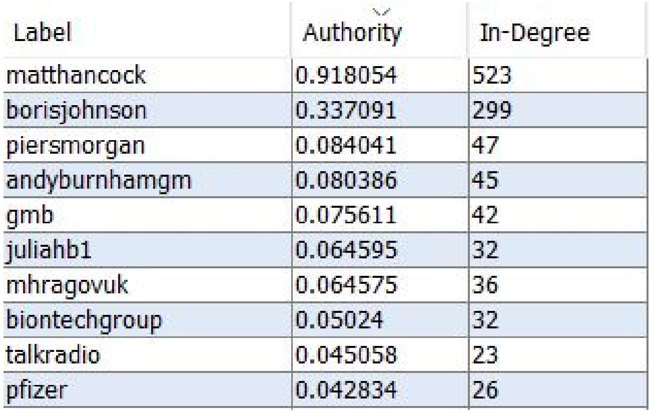


Figure 5. Politic Directed Graph filtered by node degree



Label	Authority	In-Degree
matthancock	0.918054	523
borisjohnson	0.337091	299
piersmorgan	0.084041	47
andyburnhamgm	0.080386	45
gmb	0.075611	42
juliahb1	0.064595	32
mhragovuk	0.064575	36
biontechgroup	0.05024	32
talkradio	0.045058	23
pfizer	0.042834	26

Figure 6. Authority for politics

According to the results in the Figure 6 it can be seen that the top nodes, those with the highest authority level, are @borisJohnson and @mattHancock, this shows that they represent nodes with an high in degree connected to nodes with a small degree. Considering the same measure (removing the two main nodes already explained) we see that have high importance also @PiersMorgan, British journalist and TV personality, and @AndyBurnhamGM, Labour mayor of Manchester very active on social and @GMB a famous English TV talk.

2.5 News Directed Graph

In order to answer the question of how does the news spread affects the network was generated the graph in Figure 7 where the nodes are proportional to the in-degree and the colors are the partitions of the community detection. The layout used for the visualization is the 'Yifan Hu'. In details:

- Number of nodes 2152;
- Number of edges 2109;
- Average Degree 0.98;
- Modularity 0.89;
- Assortativity -0.11;
- Number of communities 284.

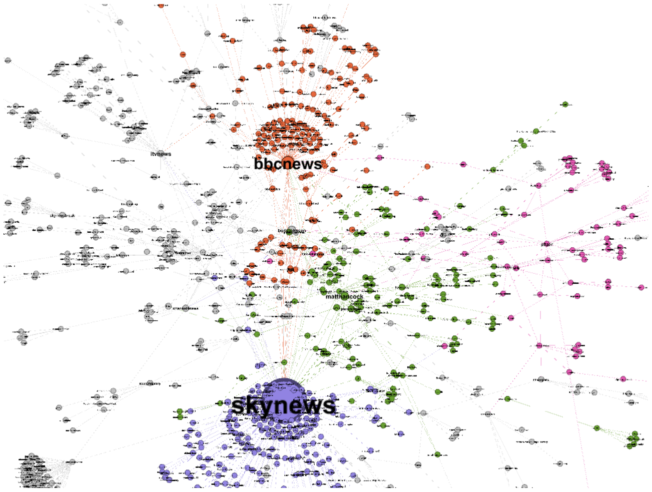


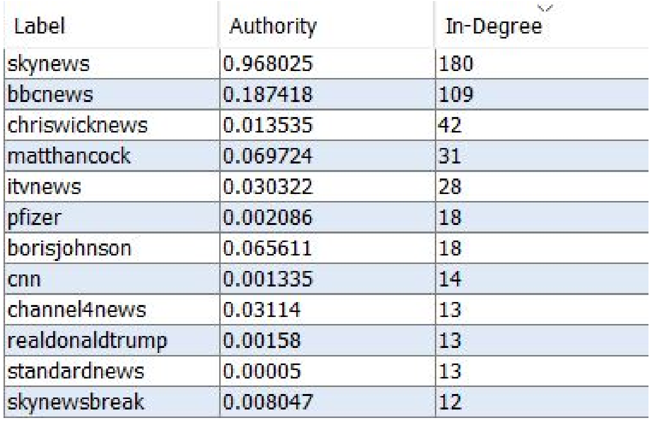
Figure 7. News Direted Graph filtered by node degree

The analysis of the graph regarding communication and media shows (Figure 8) that the highest level of authority is that of the node belonging to @SkyNews, which is expected. Interesting to see that the nodes we thought were more important, namely @bbcnews and @cnn have a much lower level of authority.

3. Social Content Analysis

For the analysis of Social Content was carried out a Sentiment Analysis based on Lexicon Approach described below.preceded by some basic text preprocessing:

- Text Normalization via elimination of links, hashtags, mentions and symbols;



Label	Authority	In-Degree
skynews	0.968025	180
bbcnews	0.187418	109
chriswicknews	0.013535	42
matthancock	0.069724	31
itvnews	0.030322	28
pfizer	0.002086	18
borisjohnson	0.065611	18
cnn	0.001335	14
channel4news	0.03114	13
realdonaldtrump	0.00158	13
standardnews	0.00005	13
skynewsbreak	0.008047	12

Figure 8. Authority for News

- Stop Word removal via NLTK;
- Lowercasing of ’location’ and ’text’ columns;
- Replace consecutive non-ASCII characters with a space;
- Stemming for reducing inflected words to their word stem.

3.1 Sentiment Analysis

Sentiment Analysis was applied via Lexicon Based Approach using the Vader library. VADER (Valence Aware Dictionary for Sentiment Reasoning) is a library used for sentiment analysis that is sensitive to both polarity of the text (positive / negative) and intensity of emotion. It belongs to a type of sentiment analysis that is based on lexicons of sentiment-related words. In this approach, each of the words in the lexicon is rated as to whether it is positive or negative, and in many cases, how positive or negative it is. Typically, we quantify this sentiment with a positive or negative value, called polarity. Polarity can range from -1 (extremely negative tweet) to +1 (extremely positive tweet). After calculating the polarity, in order to clarify the differences between days and topics, the final results are displayed via line charts, built with the mathplotlib library. The graph is built by calculating the average of the values of each tweet, for each day and topic. First was implemented the comparison of the sentiment of two key figures in British public health during the covid epidemic: Boris Johnson and Matt Hancock like was done for the network analysis.

As can be seen from the Figure 9, in the first part of the time frame in which the tweets were collected, the sentiment did not seem to deviate from an almost neautral polarity. Matt Hancock’s polarity plummets to an average of -0.75 on December 16th. In fact, on that date, Health secretary Matt Hancock announced that London, plus parts of Essex and Hertfordshire would be placed under the highest level of coronavirus restrictions from Wednesday 16 of December. This meant millions more people were in Tier 3, and businesses have had to reclose. In October, the prime minister introduced a three-tier system of local lockdown measures for England to help control the spread of coronavirus. Different parts of the country were split into medium (tier 1), high (tier 2), and very high (tier 3), local coronavirus alert areas, with respective sets of rules depending on the categorisation of risk. As for Boris Johnson, sentiment rose up to around 0.8 on December 18th. On date day, the British Prime Minister said that his team of Brexit negotiators ”will keep talking” but that the talks have been ”difficult” and that ”a gap” that needs ”to be filled” remains. Johnson reiterated his insistence that the UK was ready to exit the transition period on the 1th January without a trade deal in place, despite nightly tariffs and other additional obstacles this would mean for trade with its largest market. Statements that have been positively received by the British people on twitter.

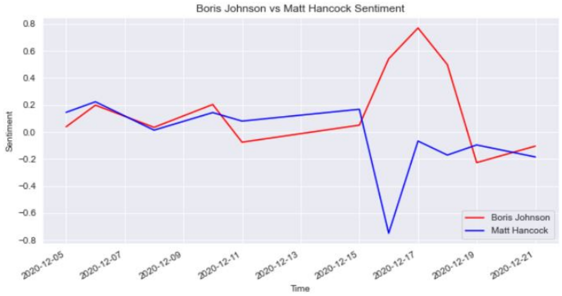


Figure 9. Sentiment Johnson vs Hancock

The second analysis was made by comparing the sentiment regarding the two large pharmaceutical companies that are delivering the largest number of vaccines in Europe: Pfizer and Moderna. Pfizer Inc. is a US pharmaceutical company. It is the largest company in the world operating in the research, production and marketing of pharmaceuticals. Moderna is a company specializing in the discovery and development of medicines based on messenger RNA. The sentiment on

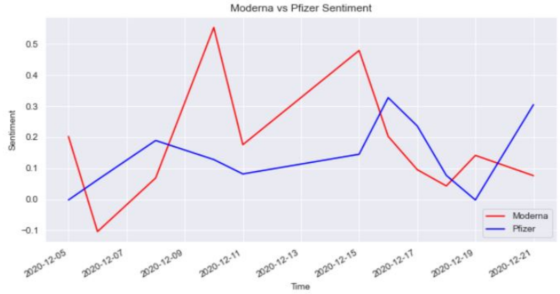


Figure 10. Sentiment Moderna vs Pzifer

Figure 10 shows Moderna undergoing a growth on 10 December 2020. It is the day that Moderna declared the start of phase 2/3 of the vaccine development. December 15th also corresponds to a peak in sentiment. During that day F.D.A. confirms Moderna’s earlier assessment that its vaccine had an efficacy rate of 94.1% percent in a trial of 30,000 people and for this reason is highly protective Against Covid-19. As for sentiment related to Pfizer, an anomaly was detected on December 19th. On that date, the sentiment related to tweets talking about the Pfizer vaccine suffered a relapse, infact the U.S. Food and Drug Administration investigated about five allergic reactions that occurred after people were given Pfizer Inc and the COVID-19 vaccine. Finally, we wanted to compare three of the most important world publishing houses: Sky, CNN and BBC. Analyzing the sentiment in Figure 11 it can be noticed that the main deviation from the neutral sentiment is on December 8th. In fact, December 8th is the day that a UK grandmother has become the first person in the world to be given the Pfizer Covid-19 jab as part of a mass vaccination program.

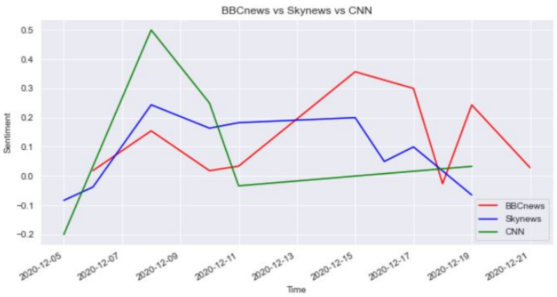


Figure 11. Sentiment Sky vs CNN vs BBC

4. Results and Discussions

Now that the network has been analyzed both from the point of view of its structure, social network analysis, and from the semantic point of view, social content analysis, it is possible to answer the research questions that opened this project. This positive mood suggests that people are pro-vaccine or that the vaccine may be a possible solution anyway. Perhaps it is too pretentious to say if the majority of people want to be vaccinated or not, but what is evident is that social networks, in this case Twitter, have a fundamental role both in spreading news and in being a mirror of people's opinion. Analyzing this network and seeing that information flows quickly, that the central nodes are represented by the main mass media and the main actors of the political scene and that the sentiment of tweets is influenced by external events, it is clear that the role of these nodes is fundamental in the information campaigns for citizens and to fight misinformation and fake news. It is important to remember that these tweets were collected at the same time as the administration of the first doses of vaccine in the UK: a moment of particular excitement and enthusiasm.

4.1 Limits and Future Developments

The collection of tweets has been limited both for reasons of permissions (limit imposed by Twitter itself) and for computational issues; a possible implementation of this work would be to enlarge the collection of users' opinions to different timelapse and countries in order to make comparison between different periods and actors. Another interesting analysis for this network is a deeper inspection of the users' profiles: it could be useful to add to each node further information, for example age, gender, education to investigate how the network is composed and find important correlations.

References

[1] Appunti del corso di Social Media Analytics a.a. 2020-2021, Elisabetta Fersini & Marco Viviani.