Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 177 (2020) 423–430

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2020)
November 2-5, 2020, Madeira, Portugal

Monitoring the Dynamics of Emotions during COVID-19 Using Twitter Data

Simranpreet Kaur*, Pallavi Kaul, Pooya Moradian Zadeh

School of Computer Science, University of Windsor, Windsor, Ontario, Canada
{kaur1a1,kaulp,moradiap}@uwindsor.ca

Abstract

The novel COVID-19 is one of the most serious health pandemics in our time. According to the World Health Organization (WHO), it has been spread over more than 150 countries and territories worldwide with thousands of deaths. In this research, we propose a framework to explore the dynamics and flow of behavioral changes among twitter users during the pandemic. In our framework, the related tweets are retrieved from the Twitter social network in three different time intervals and stored in our data repository. After cleaning and pre-processing the data, using natural language processing and social network analysis techniques, a set of emotions is extracted from them along with their sentiment characteristics. Further, the data is visualized in order to identify the changing patterns. The results of this project show significant connections between the infection and mortality rates and the emotional characteristics of the twitter users.

Keywords: COVID-19; Pandemic; Twitter data analysis; Emotional Analysis; Data Visualization; Social Network Analysis, Sentiment Analysis

1. Introduction

Social networks have a significant role in enhancing interaction among peers, family, and friends [11]. These platforms give users a voice to connect universally. The versatility of these networks proves to be beneficial in numerous forms, such as in digital marketing of products[9], data analysis[18], dissemination of helpful information linked with preventive measures[17], and daily updates regarding the number of cases and loss of life count in the ongoing pandemic situation [7]. Twitter is one of the platforms used by various organizations [20], professionals, students, and ordinary people [4]. The tweets posted by users may convey a lot more than just mere set of words [15]. It may

[*] Corresponding author.
E-mail address: kaur1a1@uwindsor.ca

10.1016/j.procs.2020.10.056

be valuable in identifying individuals suffering from depression, anxiety, or similar emotional disorders. Location of twitter users may assist in looking for a group or section of masses undergoing the same kind of problematic situation or a joyful experience. Researchers may put to use this data for analysis and search for patterns that may facilitate identical research problems in the future[12]. Thus, compiling such tweets and then performing an analysis may assist in detecting sections of people with similarities in this kind of behaviour[15]. Such data, as a mirror of reality, can also be used to identify events, natural disasters, or monitor the spread of information, news, and rumors around the world [16]. In addition, disease surveillance as a tool to monitor the spread of diseases has been the center of attention of both policymakers and educational researchers for quite some time. Commonly focused diseases include cancer, flu, different kinds of viruses, or any other widespread disease[7]. Furthermore, disease surveillance plays a vital role in mitigating the harm caused by outbreaks and pandemics by continuously monitoring its spread[7].

COVID-19 is an infectious disease caused by the recently found virus known as SARS-CoV-2(Severe Acute Respiratory Syndrome) [2]. Its outbreak is beyond the previous observations of this virus and is thus considered as the pandemic by the World Health Organization (WHO). Since the spread of COVID-19 has already affected millions of people worldwide [21], therefore, it becomes essential to study the trend of this disease beginning from its time of origin till present. It will be beneficial for medical research and government health organizations to determine the pattern and health status of citizens, and hence, there is a need for surveillance of this disease so that its spread can be put to halt or at least may help reduce its spread to some extent.

The main objective of this research study is to explore the dynamics and flow of behavioural changes among twitter users during the pandemic. Consequently, we propose a framework to collect and process tweets related to the ongoing pandemic situation from the Twitter social network in three different time intervals. Based on the locations given by the users, these tweets are segregated geographically from each other. The sentimental and emotional analysis is conducted to identify emotional changes in the status updates as given in[12]. The results of this process are visualized to show the trend and dynamics of the changes.

The rest of the paper is arranged as follows: Section II throws light on the existing state-of-the-art techniques used and studies conducted by researchers in the field of disease surveillance, Section III describes a detailed explanation of the proposed approach followed by Experimental Setup giving details regarding the implementation process in Section IV. Section V lists the deductions obtained during the implementation stage accompanied by the discussion stating some of the inferences drawn based on the results. Concluding remarks and future directions are given in Section VI.

2. Literature Review

In this section, we review some of the existing research works involved in the Twitter Data Analysis.

Merchant et al. [10] described how social media acts as an integral tool in handling the ongoing pandemic and reconstructing awareness and feedback in the future. Some of the benefits listed by them include guiding masses to verified and trusted resources. They also listed how rumors and misinformation may be spread in such a situation. Further, they pointed everyone's attention towards social media platforms becoming the only means to communicate amidst the situation where social distancing and lock-down are prevalent. Achrekar et al. [1] aimed at investigating the usage of tweets posted by Twitter users to determine and anticipate circumstances related to the influenza epidemic in the real world. They followed the conventional method adopted by the Centers for Disease Control and Prevention (CDC) to detect the disease by gathering data for ILI (Influenza Like Illness) activity from medical records. Usually, a delay time of one to two weeks is observed between the diagnosis period, and the results are obtained. The authors collected 4.7 million tweets from 1.5 million unique users who quoted flu-related description in their tweets. As a result, tweets collected revealed a high correlation with the influenza-like illness activity data obtained from CDC. The authors concluded that data obtained from Twitter contributes to achieve impending real-time estimation and prediction of influenza-related activity and ongoing ILI activity standards.

In [7], the authors proposed a new real-time disease monitoring system utilizing data obtained from twitter to record activities associated with influenza and cancer. They collected tweets using twitter streaming API and applied different models such as geographical, physical, and content-based to explore activities associated with flu and cancer, along with the increasing frequency of terms related to diseases. In light of the system's automation and updating of the analysis of outputs in real-time, their proposed system can detect outbreaks of the disease quite rapidly when compared to the conventional disease monitoring systems collecting data from medical records. In [19], the authors carried out a study focused on interpreting and mining the extent of information associated with the Ebola virus in social networks. They analyzed geographically labeled data to understand the opinion of individuals concerning Ebola. They set up information spread models to estimate the location of the source of information and examine different ways in which geographical, physical, and general properties of data are related to the propagation of information based on Ebola. Furthermore, in [13], Oscar et al. proposed a partially automated text coding technique and analyzed tweets related to Alzheimer's disease (AD) and dementia. They employed machine learning techniques for modelling stigmatization represented in over 30000 tweets relating to AD collected using twitter's API depending upon nine keywords related to AD. Results obtained showed that approximately twenty percent of the tweets related to AD contained keywords associated with AD to spread the public stigma, which could affect conventional ideas and give rise to a pessimistic attitude among those concerned with AD, thus increasing their disability further.

Masri et al. [8] focused their attention on tracking the Zika virus epidemic in the United States in general and Florida, particularly using Cloudberry to filter random samples of data collected from Twitter. They tried to determine weekly Zika virus cases in advance from weekly case counts and related tweets using two regression models. The authors claimed that usually, models tend to over predict at low case counts and under-predict at maximum counts. Kaila et al. [14] considered the flow of information on Twitter during the continuous period of COVID-19. They performed sentiment analysis and topic modeling using Latent Dirichlet Allocation (LDA) on tweets containing #coronavirus. Analysis using LDA was able to identify suitable and authentic topics related to the spread of the coronavirus outbreak. Also, sentiment analysis verified the presence of positive sentiments such as trust and negative emotions in the tweets given by users.

Kouzy et al. [6] pointed at analyzing the degree of false information spread about ongoing COVID-19 pandemic situation on Twitter. They began by searching for tweets containing fourteen commonly used hashtags and keywords having a connection with COVID-19, and then evaluated each tweet for finding any false information present in them. They considered around 600 tweets for their investigation. On evaluations, they found that very few belonged to verified Twitter accounts out of which, around half of them were either containing misinformation, or were incapable of verification. Tweets from informal or unverified accounts possessed a higher rate of misleading information than those coming from health-related organizations. Chen et al. [3] created a COVID -19 Twitter dataset containing diverse languages. They began their research in January 2020. The accumulation of tweets was done using Twitter's streaming API and Tweepy to acquire certain keywords, and Twitter accounts influential at the time of the pandemic. They have already circulated over 120 million tweets, out of which 60% were in English. Their documentation involved fundamental analysis showing the response and reaction in relation to twitter movement.

3. Proposed Architecture

As shown in Fig.1, our proposed system is composed of 4 different phases: data collection and extraction, data cleaning and pre-processing, sentiment and emotional analysis, and visualization and integration with other datasets (i.e., spread and mortality rate based on WHO datasets).

In the first step, the tweets related to the pandemic is retrieved from the Twitter network and stored in our local database. Therefore, various keywords associated with COVID-19 are identified to be sent to Twitter as a search query. The collection of tweets from Twitter is initiated using Twitter Search API, which is used to stream tweets related to supplied keywords. In order to identify the related tweets, we use the most trending keywords used by Twitter users. The retrieved raw tweets are then processed, and the required contents are extracted from them (e.g., tweet text, location, timestamp, etc.) and stored in a local database.

Tweets are collected five times a day for seven days in three-time intervals in February, May, and June. The reason for collection of these tweets in three different time-period lies in the fact that in the month of February, it was very close to the time that the pandemic became known to the public, and its impact started to be revealed gradually whereas in the month of May, it already grew out as a pandemic and affected many regions. On the other hand, by the month of June, some of the countries affected with the pandemic were able to mitigate/recover the impact of COVID-19.

After completing the data collection process, the tweets are being pre-processed. This step includes removing the duplicate and irrelevant tweets, along with the removal of all URLs in their contents. This process is done automatically and is being verified manually. Since, some tweets did not contain attached locations or were incorrect, tracking

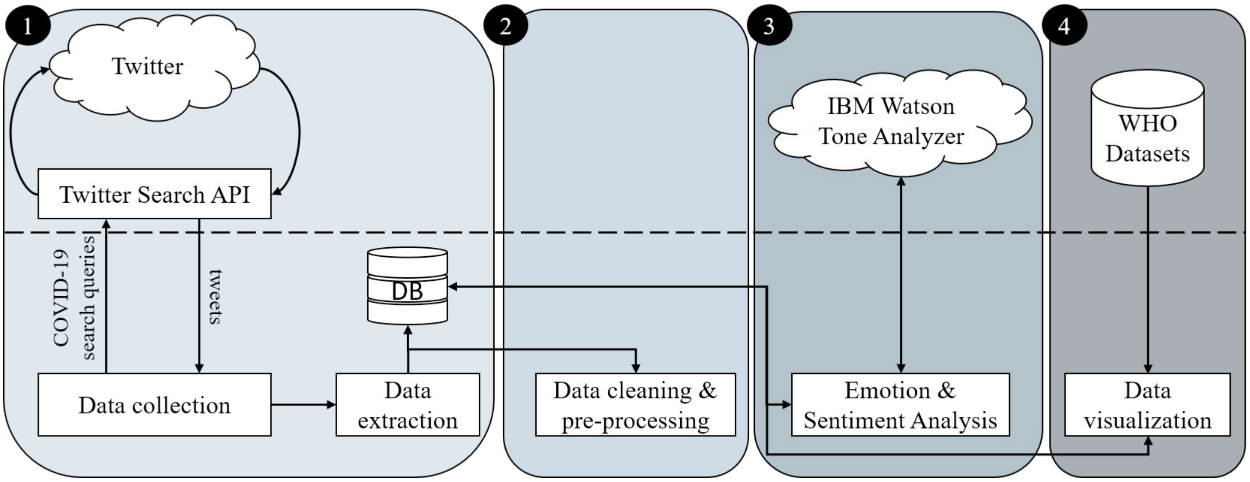

Fig. 1. The high level architecture of the proposed system

of user-profiles is done to find their most recent profile-location. In addition, the tweets without geo-location tags are separated in this step from the rest of them. Meanwhile, in case a tweet is not written in English language, it is translated into English language for the analysis.

The next step is to perform sentiment analysis to extract the polarity of the tweets and classify them based on three categories, namely: Positive, Negative, or Neutral. Side-by-side, emotional analysis is performed on the collected dataset to get a deeper understanding of the user's emotional engagement from the text. Anger, Analytical, Joy, Sadness, Tentative, Confident and Fear are the list of emotions considered in this research. IBM Watson Tone Analyzer is used for the emotion analysis. Therefore, the tweets are sent to the Watson Anlytics and as a result a set of emotional labels is assigned to each of them. For validation, each label received via automation is manually reviewed to confirm its correctness and modified accordingly.

In the last phase, statistical data regarding the pandemic's spread and mortality rates are gathered from the WHO datasets and are integrated with our twitter data. Various types of visualizations are conducted in order to demonstrate the flow and dynamics of emotional changes in the network.

4. Experimental Setup

For contriving the dataset, the proposed system was created using Python programming language that invokes the Twitter Standard Search API to stream tweets based on the keywords given in the python framework. The keywords used in this experiment were ”coronavirus”, ”coronavirusec”, ”coronavirusoutbreak”, ”COVID”, ”COVID19”, ”COVID-19”. The extraction script was executed according to selected keywords for three time periods from February 05 to February 11, May 21 to May 27 and June 15 to June 21 in the year 2020. The response was parsed using the Python code to extract the elements of interest like tweet timestamp, location the user, and the tweet text. As a result, a total of 16138 tweets was extracted. The extracted items were stored in the database along with the timestamp at which the script was executed.

From the tweets collected, it is observed that the locations entered by some of the users are inconsistent for knowing the exact location from where the tweets originated. Hence, the segregation of tweets was done based on the locations provided as input into their countries and continents, thus assisting in locating the tweets based on their actual geographical location. During the tweet collection and extraction phase, some tweets retrieved were in diverse languages. However, to perform the sentiment and tone analysis, the pre-condition to the used APIs is to have the text input in English. For translating the tweets to English, a JAVA automation script using Selenium Web Driver was employed to access Google Translate service. The translated tweets are then stored in our database. As mentioned in the previous section, for pre-processing, URL entities, duplicate and irrelevant tweets, and emoticons were excluded from the original tweets list. Following the pre-processing phase, sentiment analysis was performed to distinguish the tweets into

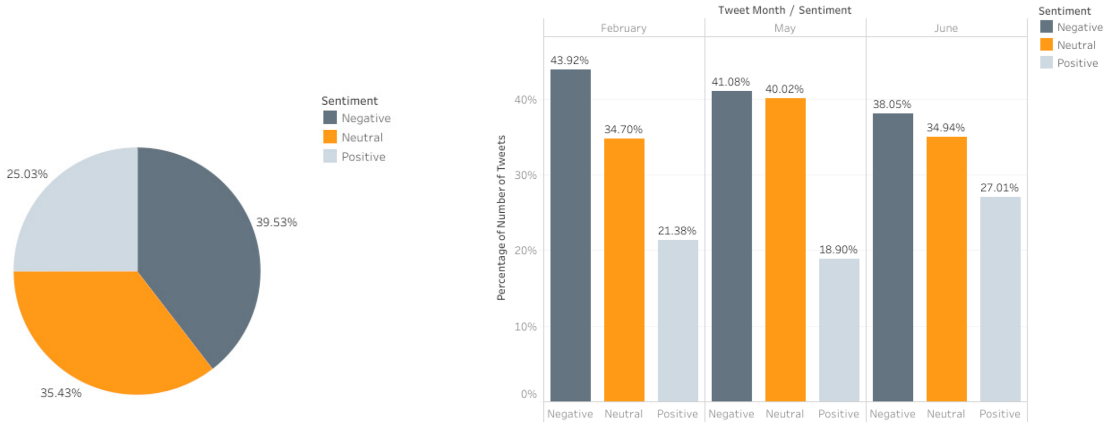

Fig. 2. (a) Tweets distribution with respect to Sentiment collectively; (b) Tweets distribution with respect to Sentiment for each interval

three broad categories, notably, Positive, Negative, and Neutral. It was implemented in Python using TextBlob. The script gets the tweets as the inputs and returns the text's polarity in terms of sentiment score. The sentiment score lies in the range of -1 to 1. Hence, the tweets are classified as 'Negative' if the score is less than 0, 'Neutral' if the score is equal to 0, and 'Positive' if the score is greater than 0. After that, the tweets' sentiments were validated manually to ensure the correctness of the output from the script.

Additionally, to extract information about the user's emotional engagement while writing the tweet, IBM Cloud is utilized to foretell the emotional tone associated with the text. A list of emotions and their corresponding scores were retrieved using the Python framework, and the output was saved in the database for additional processing. Each tweet depicts more than one emotion; however, only those emotions were considered whose corresponding score was either higher than 0.55, or it was the only emotion for that tweet. Similar to sentiment analysis, the emotions of the tweets were validated manually for its correctness. In addition, the daily mortality and disease spread rates were collected from WHO datasets and stored in our database. Meanwhile, the visualization was conducted using Tableau software.

5. Results

This section gives an insight into the results obtained from the above experiment. Fig.2(a) shows the percentage distribution of the total number of tweets based on the sentiments obtained over the three-time intervals. As can be observed, the highest percentage of tweets, that is, 39.53% belongs to the 'Negative' category followed by Neutral tweets with 35.43% and Positive tweets with the minimum percentage of 25.03%. Similar trends can be observed in Fig.2(b) where the percentage of tweets belonging to negative sentiment in all the intervals exceeds that of positive and neutral sentiment. Also, while negative tweets possess a continuous decline, neutral and positive tweets followed a fluctuating trend. Concerning the date-wise data for sentiment as in Fig.3, the highest percentage of negative tweets and lowest percentage of neutral tweets were obtained on February 9 whereas, the highest percentage of neutral tweets and lowest percentage of negative tweets were obtained on May 27. In the case of positive tweets, the highest was observed on June 17, whereas the lowest was found on May 21.

Fig.4 displays the percentage distribution of tweets based on their corresponding emotions. According to the results, the percentage of tweets belonging to the 'Sadness' attribute is higher in all time intervals than the remaining list of emotions. As far as the trend is concerned, tweets showing Confidence and Joy depicts a continuous incline in contrast to Tentative tweets with continuous decline and Anger, Fear and Sadness tend to fluctuate over the given time intervals.

It is apparent from Fig.5 that number of deaths and the number of tweets with sadness emotion in the month of February shows no consistency, in May shows consistency at some points and in June, shows full consistency. The result shows that, the tweets with sadness emotions reached its highest peak on May 26 with the differences in the number of deaths [21], reaching the lowest on this same day. Also, if the line graph in June is analyzed, it shows a

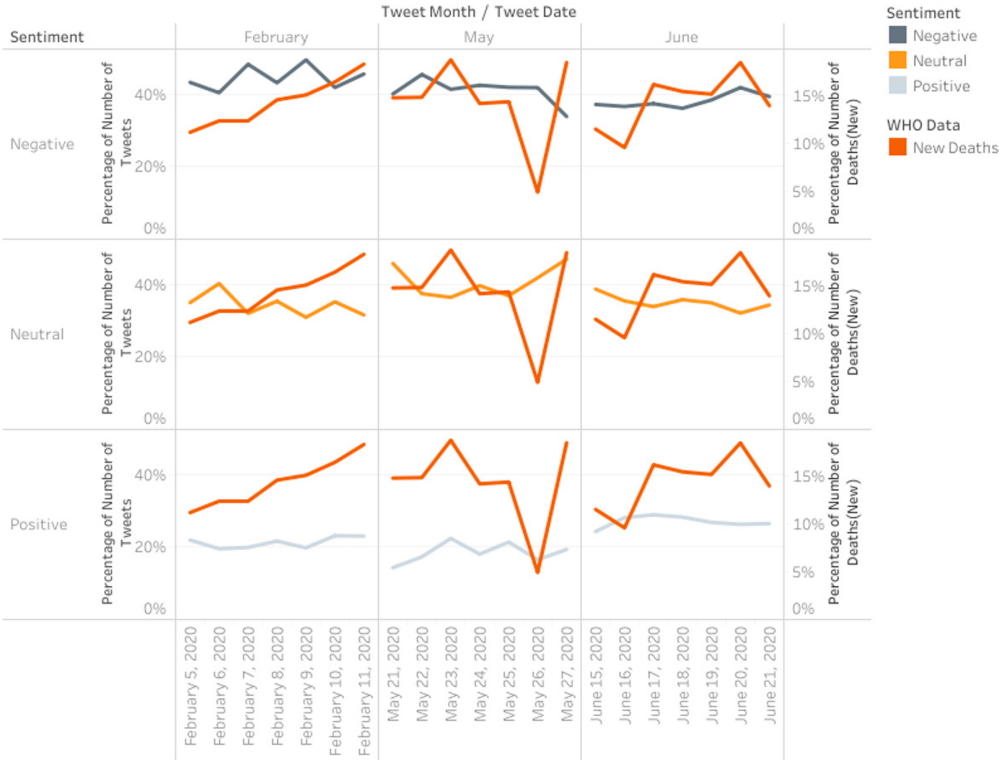

Fig. 3. Tweets distribution with respect to Sentiment for each date of the interval

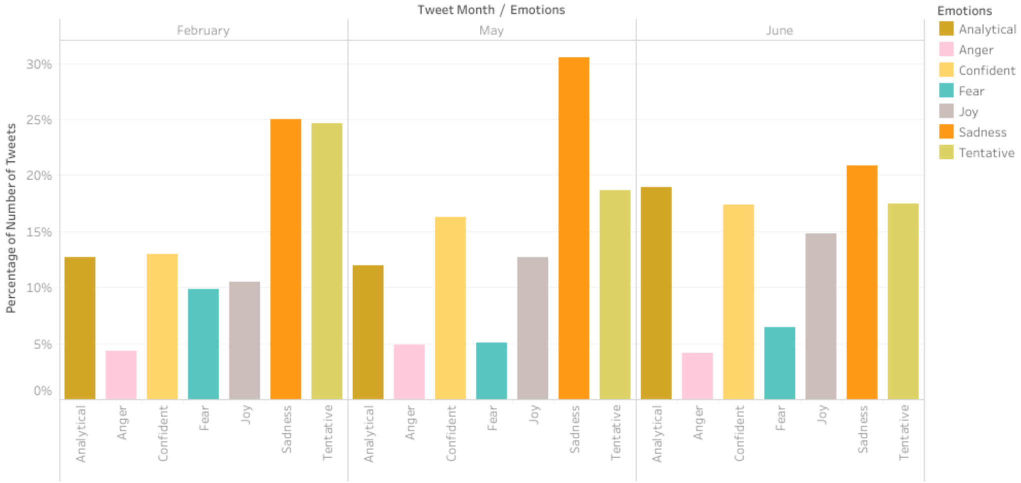

Fig. 4. Tweets distribution with respect to emotions for each interval

similar trend to the bars showing levels of sadness emotions. Whereas, the graphs in the remaining months show a contrasting pattern when compared with each other.

The geographical distribution of tweets with respect to emotions for seven profoundly affected countries is shown in Fig.6. Apparently, it represents the situation where apart from Italy and the United Kingdom (UK), the rest of the countries possess the highest percentage of ’Analytical’ tweets in the month of June. On the other hand, the maximum percentage of ’Tentative’ tweets appear in every country except Italy and Spain in the month of February.

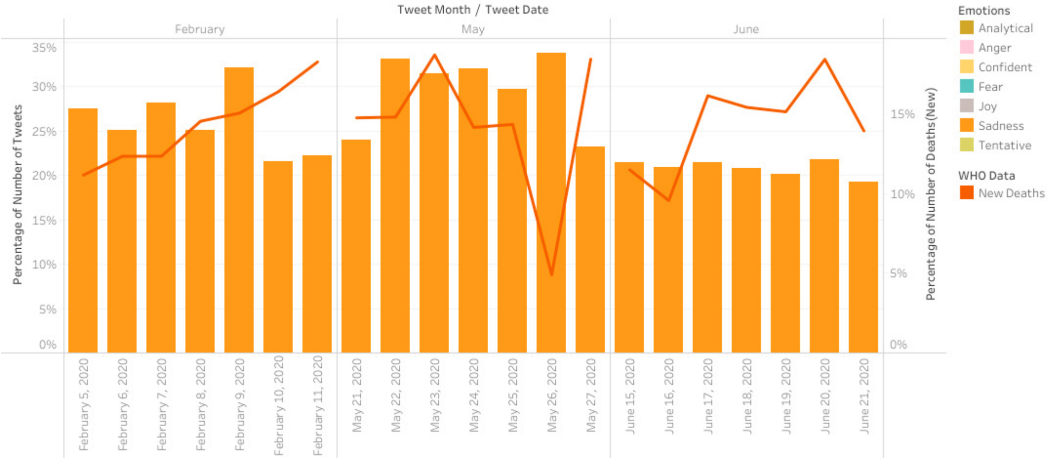

Fig. 5. Tweets distribution of 'Sadness' emotion with respect to the number of deaths for each date in the time interval

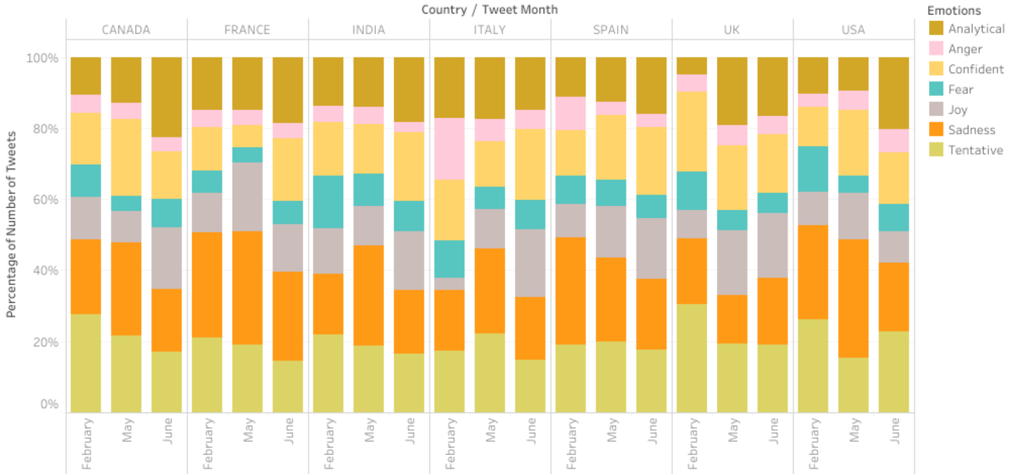

Fig. 6. Tweets distribution with respect to emotions for highly affected countries

5.1. Discussion

As observed from Fig.2(b), the highest percentage of tweets belongs to the 'Negative' category, illustrating the negative sentiments possessed by the majority of the masses. In February, when the mortality rate was not that high[21], we observe higher positive tweets compared to May, although our research is limited and more data are needed to analyze the root of this issue, one possible reason could be that people have not anticipated the destructive nature of the virus at that time. However, in May, when it began spreading throughout the world [21], optimistic (positive) nature of people might have changed to being pessimistic (negative) justified by the drop of 2.48% in tweets with 'Positive' sentiment. Also, when the pandemic situation turned into a daily routine, and people got accustomed to living with masks and carrying out daily chores amid lock-down[21], the rate of optimism might have risen.

It is worth mentioning that the highest number of tweets describing 'Sadness' emotion, shown in Fig.5, and the drastic decrease in the number of new deaths is observed on May 26. The brutal killing of George Floyd just the day before [5] may be one of the main reasons for this kind of trend. Also, while going through the tweets retrieved on this day, it is found that people from all around the world felt sympathetic towards him as well as the ones affected by Coronavirus. However, more comprehensive research is required to identify the underlying reasons behind our observations in this research.

6. Conclusion and Future Works

This research aims to study the varying trends in human behavior based on their tweets during the ongoing COVID-19 pandemic situation. First, a collection of tweets was extracted from the Twitter network based on the most commonly used hashtags related to coronavirus in three different time intervals. Pre-processing and cleaning of the data were done to prepare them for further analysis. After that, sentiment analysis and emotional analysis were performed, followed by the manual validation process. Finally, the results were visualized and analyzed to demonstrate the dynamics of emotional changes in the observed tweets. This research may assist in supporting future research work in the field to understand the variations in the attitude of people during any other similar pandemic situation. In the future, we will conduct our study on a larger set of social media data obtained from multiple social network websites.

Acknowledgements

The authors would like to thank Mr. Narinder Pal Singh for his support with this project.

References

[1] Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H., Liu, B., 2011. Predicting flu trends using twitter data, in: 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS), IEEE. pp. 702–707.

[2] Canada, P.H.A.o., 2020. Government of canada. URL: https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/symptoms.ht3ml.

[3] Chen, E., Lerman, K., Ferrara, E., 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR Public Health and Surveillance 6, e19273.

[4] Hanson, C.L., Burton, S.H., Giraud-Carrier, C., West, J.H., Barnes, M.D., Hansen, B., 2013. Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. Journal of medical Internet research 15, e62.

[5] Hill, E., Tiefenthäler, A., Triebert, C., Jordan, D., Willis, H., Stein, R., 2020. How george floyd was killed in police custody. URL: https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html.

[6] Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M.B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E.W., Baddour, K., 2020. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. Cureus 12.

[7] Lee, K., Agrawal, A., Choudhary, A., 2013. Real-time disease surveillance using twitter data: demonstration on flu and cancer, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1474–1477.

[8] Masri, S., Jia, J., Li, C., Zhou, G., Lee, M.C., Yan, G., Wu, J., 2019. Use of twitter data to improve zika virus surveillance in the united states during the 2016 epidemic. BMC public health 19, 761.

[9] McCorkle, D., Payan, J., 2017. Using twitter in the marketing and advertising classroom to develop skills for social media marketing and personal branding. Journal of Advertising Education 21, 33–43.

[10] Merchant, R.M., Lurie, N., 2020. Social media and emergency preparedness in response to novel coronavirus. Jama .

[11] Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N., 2011. Understanding the demographics of twitter users, in: Fifth international AAAI conference on weblogs and social media.

[12] Nambisan, P., Luo, Z., Kapoor, A., Patrick, T.B., Cisler, R.A., 2015. Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter, in: 2015 48th Hawaii International Conference on System Sciences, IEEE. pp. 2906–2913.

[13] Oscar, N., Fox, P.A., Croucher, R., Wernick, R., Keune, J., Hooker, K., 2017. Machine learning, sentiment analysis, and tweets: an examination of alzheimer's disease stigma on twitter. Journals of Gerontology Series B: Psychological Sciences and Social Sciences 72, 742–751.

[14] Prabhakar Kaila, D., Prasad, D.A., et al., 2020. Informational flow on twitter–corona virus outbreak–topic modelling approach. International Journal of Advanced Research in Engineering and Technology (IJARET) 11.

[15] Quercia, D., Ellis, J., Capra, L., Crowcroft, J., 2012. Tracking" gross community happiness" from tweets, in: Proceedings of the ACM 2012 conference on computer supported cooperative work, pp. 965–968.

[16] Singh, J.P., Dwivedi, Y.K., Rana, N.P., Kumar, A., Kapoor, K.K., 2019. Event classification and location prediction from tweets during disasters. Annals of Operations Research 283, 737–757.

[17] Song, P., Karako, T., 2020. Covid-19: Real-time dissemination of scientific information to fight a public health emergency of international concern. Bioscience trends .

[18] Takac, L., Zabovsky, M., 2012. Data analysis in public social networks, in: International scientific conference and international workshop present day trends of innovations.

[19] Tran, T., Lee, K., 2016. Understanding citizen reactions and ebola-related information propagation on social media, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE. pp. 106–111.

[20] Unsworth, K., Townes, A., 2012. Transparency, participation, cooperation: a case study evaluating twitter as a social media interaction tool in the us open government initiative, in: Proceedings of the 13th Annual International Conference on Digital Government Research, pp. 90–96.

[21] WHO, 2020. Coronavirus disease (covid-19) situation reports. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.