Were people neutral or positive about the Covid Outbreak on Twitter?

Monisa Qadiri*, Muzafar Bhat** & Naffi Ahanger***

[Although there is less scare with the latest reports of Covid returning to the scene, the Covid-19 pandemic has stayed for more than three years killing almost seven million people across the world. The outbreak was compared to dreaded historic epidemics like the 'The Great Influenza' (1918 Spanish flu), or the Black Death (bubonic plague. The study analyses the emotions and themes discussed on Twitter about the pandemic to understand public perception using machine learning. Contrary to the expectation that people shared negative sentiments, the findings revealed that mostly neutral sentiments were displayed and diverse topics were discussed.]

Humans witnessed one of the deadliest disease outbreaks caused by a newer strain of Coronaviruses (Covid) discovered in December 2019, not previously identified in humans ("Coronavirus", 2020). On 30 January 2020, the WHO Director-General declared the outbreak a public health emergency of international concern a 'pandemic' (Taylor, 2021). This was unlike any previous health concerns as it affected a significant world population.

After infecting patient 0 in China, this disease spread to multiple countries in a few months and the worst affected were the United States of America, India, and Brazil. The issue grabbed media attention the world over as the cases grew every day making people follow the news related to it. The number of people infected from different countries went up with every passing day, peaking in the first half of 2020.

The total number of coronavirus cases is close to seven hundred million at present ("COVID Live Update," 2023). Many prominent people like political leaders, Hollywood actors, celebrities, and several sportspersons tested positive for COVID-19 (Al Jazeera, 2020). Even some doctors treating Coronavirus patients lost their lives taking the number of mortalities to millions.

Thus for months, the crisis emerged as the most prominent news for the world with a minute-by-minute update on the number of dead, infections, vaccine status, precautions, and measures by the governments. Round-the-clock special bulletins and online content were produced for creating much-needed awareness among the public. However, there have been concerns about the numerous challenges of reporting a rapidly evolving medical crisis like COVID-19, as any inaccurate information could cause multifarious levels of panic (H, 2020) and many media houses and outlets have been accused of distorted news and bias (Al-Burai, 2020).

Nevertheless, digital platforms were more accessible and available to people in the absence of many communication and news tools like newspapers, etc. and this interface was the avenue for searching and sharing information. Simultaneously, social media became a readily available source of vital information as well as a potent space for misinformation for many users around the world (Sokolov, 2020). After "Coronavirus" was mentioned across different social platforms and news media in the last week of February (Molla, 2020), this outbreak was the most trending online topic for many weeks.

During the last decade— the age of social media, at least three other international pandemics; the H1N1 (swine flu) pandemic, the Ebola epidemic, and the Zika outbreak have occurred, which were discussed before the Covid, but not like this. Thus, outbreaks get widely documented and have an impact on the way social media conversations take place (Sokolov, 2020).

Twitter, being a significant micro-blogging space today is used for socializing or for disseminating information (Vijayakumar, Umamaheshwar, Bambaataa, et al. 2015). It is shaping how 'cyber beings' communicate, whether we think of ordinary citizens or global leaders, there is a frequent exchange happening through Twitter (Bhat, Qadri & Kundra, 2019). This has emerged as a significant area of research globally, particularly during the Pandemic. Previously also, studies analysed the latent themes or topics from tweets (Bhat et al, 2020) or the sentiment of tweets using machine learning approaches (Agarwal & Mittal, 2016) or to understand the overall sentiment of a country like India (Berker, Vibha, & Kamath, 2020) regarding lockdown in India.

* Dept. of Journalism and Mass Communication, Islamic University of Science and Technology, J & K.
** Dept. of Computer Science, Islamic University of Science and Technology, J&K.
*** Dept. of Computer Science, Islamic University of Science and Technology, J&K.

The current paper attempts at analysing the bulk of Twitter posts through the most prominent hashtags concerning Covid 19 in the initial months of the pandemic. These user-generated metatags aid the topic modelling and sentiment analysis (Steinskog, Therkelsen, & Gambas, 2017) to help in understanding the topics, and sentiments mostly expressed by the users ever since this pandemic hit the world. This paper also attempts to understand and reveal hidden topics or patterns associated with the selected hashtags using Biters Topic Modelling.

Proposed Approach

For our study, the tweets were collected from Twitter and saved along with some associated data as flat files. Then, a standard data pre-processing process was conducted in various steps to clean the data and sentiment analysis was performed using textblob to obtain the sentiments associated with each tweet as well as performing topic modelling using Biterm Topic Modelling.

Experimentation

Data Collection

The data were collected via Twitter Search API through select keywords as well as from tweet ids. In this experimentation, tweet ids regarding coronavirus were obtained from an open-source dataset (Lamsal, 2020) and actual tweet data was fetched from Twitter API using those tweet ids. The research analysed a total of 1.3 Million tweets as data corpus from the first peak period of four months from 20 March, 2020 to 27 July, 2020 and daily 10,000 tweets were collected randomly.

Data Pre-processing

Most Twitter data is highly unstructured and user-generated and may have linguistic inconsistencies and semantic issues etc., like typos, slang usage, and grammatical mistakes. Thus, cleansing steps are applied to generate structured data. The main pre-processing involved:

- Conversion to lowercase letters
- Removing @user and links
- Removal of punctuation and digits
- Removal of stopwords
- Removing extra white spaces
- Stemming the documents i.e., eliminating affixes from words to convert the words into their base form, for example, words like "run", "runs" and "running" can be stemmed into a single word -"run".
- Saving files

The Languages and libraries used for data collection and pre-processing included:

- Language: Python (v.3.6)
- Libraries used are Tweepy, regex, NLTK, csv, Pandas, Pre-processor, and Textblob.

Analysis and Results

Sentiment Analysis

For sentiment analysis, we used the textblob library in python based on top of the NLTK library to compute sentiments expressed in each pre-processed Tweet for both the hashtags; #COVID19 and #coronavirus, and used sentiment analysis scales as positive, negative or neutral. After pre-processing, results about 1.11 million cleaned tweets related to #coronavirus and #COVID-19 showed that 43.43% of Tweets had neutral sentiments, 35.60% of Tweets were positive and about 20.97% of Tweets were negative.

Some randomly cleaned tweets along with their sentiment are:

Positive: *"Excellent way to use the power of the media to spread a message of hope and positivity...that India will defeat COVID-19" "We will fight together and win from corona with the blessings of almighty god. Let's pray"*

Neutral:*" We are brainstorming how to keep our futures lab community engaged during the pandemic" "Corona day 3: it just feels like Sunday again and…again" "#Corona Awareness: Soap or Hand Sanitizer, what should I use to wash my hands?"*

Negative: *"I do not know what the evil that controls the world is up to, but if this is a test, then we are all headed to the slaughtered" "While Corona Pandemic has been politicised in all countries- including United States - but worst politics is being played in Pakistan" "It remains shocking that the United States is the worst country in the world handling the pandemic."*

Topic Modelling

For topic modelling, values of hyperparameters for training the data were required:

a. The first hyperparameter is having an optimal number of topics for which, few metrics like Arun2010 (Arun, Suresh, Madhavan, & Murthy, 2010), CaoJuan2009 (Cao, Xia, Li, Zhang, & Tang, 2009), Deveaud2014 (Deveaud, SanJuan, & Bellot,2014), and Griffiths2004 (Griffiths & Steyvers, 2004) were used on pre-processed datasets. The results from these metrics established that both datasets could be modelled on three topics.

b. The other hyperparameters alpha (Symmetric Dirichlet prior of P (z)) and beta (Symmetric Dirichlet prior of P (w|z)) were calculated using the

following formulas as suggested by (Griffiths & Steyvers, 2004)

alpha = 50 /T (1)

where T is no. of topics

beta = 200/w (2)

where w is the no. of words in the vocabulary.

For the Topic discovery, java code or pure implementation of the original Biterm Topic Model (Yan, Guo, Lan, & Cheng, 2013) was used. The code was run for several topics = 3, alpha (Symmetric Dirichlet prior of P (z)) = 16.66, beta (Symmetric Dirichlet prior of P (w|z)) = 0.01, number of iterations = 2000.

After the execution of 2000 iterations, results were saved in text files. Four text files generated by the code are:

- Document Topic Matrix.
- Topic Word Matrix.
- Top 20 words of each Topic.
- Term Frequency.

Word clouds for each tweet retrieved for the study were also generated using Pythons Word cloud library by giving term frequency matrix as input to the library. Three topics for each of the hashtags were generated, which define the themes and sub-themes highlighted by the users on the microblogging site.

From the generated data files of coronavirus-related tweets, the most frequent words used were obtained from the corpus. The top ten words of the three topics are as below:

I. Topic 1: Covid, Coronavirus, Triumph, Pandemic, Corona, People, Health, President, Government, Virus.

II. Topic 2: Cases, New, Deaths, Positive, Tested, Day, death.

III. Topic 3: Virus, Like, Pandemic, Know, Please, Home, Mask

Further, the top twenty words and corresponding frequencies included Covid as most used with f=224069, followed by corona with f=173566, coronavirus with f=122173, and other words like people, virus, pandemic, Trump, death, and lockdown were most commonly used.

Conclusion

A paradigm shift was observed in how people led their day-to-day lives or how practices of prevention, physical distancing, and isolation were incorporated into their lives to prevent the spread of this virus. The fear of this disease was omnipresent, as it affected millions, either by infection or by disrupting their lives. Due to disruption in communication and physical engagements, social media became a source for socialisation as well as for information sharing. Multiple hashtags and trends were used, including #Covid19 and #Coronavirus, as the most prominent hashtags used by this research to assess the nature of global perception regarding the pandemic in the first half of 2020.

The analysis of over a million tweets by the application of Topic Modelling and Sentiment Analysis helped understand these sentiments and rather than an expected result of finding more negative sentiments, the study revealed that the perception was mostly neutral, followed by positive sentiments. These formed the bulk of emotions in the initial months when much uncertainty was seen Interestingly, negative sentiments were least expressed online but that may not mean they were not worried, instead, it could mean they were discovering ways and means of staying positive mentally and not physically. It defined the way they choose to approach the disease on social media, which makes studies based on user behaviour important.

While, Topic modelling through Bit Term was applied by using three topics, and word clouds and topic clouds were generated. This presented an idea about the most discussed topics and words used by Twitter users. The constructive expression also indicates support for helpful gestures and government-led steps to curb the disease. This highlights the role of social media analysis, AI approaches, and information and communication technology in matters of health communication while amplifying hope for the future.

References

1. Coronavirus. (2020, January 10). Retrieved from https://www.who.int/health-topics/coronavirus
2. Taylor, D. B. (2021, March 17). The Coronavirus Pandemic: A Timeline. Retrieved from https://www.nytimes.com/article/coronavirus-timeline.html
3. COVID Live Update. (2023). Retrieved March 31, 2023, from https://www.worldometers.info/coronavirus/
4. Al Jazeera. (2020, October 2). Coronavirus pandemic: Which politicians and celebs are

affected? Retrieved fromhttps://www.aljazeera.com/news/2020/9/20/coronavirus-pandemic-which-politicians-and-celebs-are-affected

5. H, C. (2020, March 11). Public media coverage of coronavirus. Retrieved from https://www.publicmediaalliance.org/public-media-coronavirus/
6. Al-Burai, A. (2020, March 17). Media bias is politicising the coronavirus pandemic coverage. Retrieved from https://www.middleeastmonitor.com/20200317-media-bias-is-politicising-the-coronavirus-pandemic-coverage/
7. Sokolov, M. (2020, March 3). The pandemic infodemic: how social media helps (and hurts) during the coronavirus outbreak. Retrieved from https://www.thedrum.com/opinion/2020/03/03/the-pandemic-infodemic-how-social-media-helps-and-hurts-during-the-coronavirus
8. Molla, R. (2020, March 12). Coronavirus dominates Facebook, Twitter, and Google searches. Retrieved from https://www.vox.com/recode/2020/3/12/21175570/coronavirus-covid-19-social-media-twitter-facebook-google
9. Vijayakumar, M., Umamaheshwar, T. M., Kambhampati, S., & Talamadupula, K. (2015, June). TweetSense: Context Recovery for Orphan Tweets by Exploiting Social Signals in Twitter. In *Proceedings of the ACM Web Science Conference* (pp. 1-3).
10. Bhat, M., Qadri, M., & Kundroo, M. (2019). *Twitter: Global Perspectives, Uses and Research Techniques (Media and Communications - Technologies, Policies and Challenges)* (1st ed., Vols. 293–316). NY, USA: Nova Science Pub Inc.
11. Bhat, M., Qadri, M., Beg, N. U. A., Kundroo, M., Ahanger, N., & Agarwal, B. (2020). Sentiment analysis of social media response to the Covid19 outbreak. *Brain, Behavior, and Immunity*, *87*, 136–137. https://doi.org/10.1016/j.bbi.2020.05.006
12. Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). Sentiment analysis using common-sense and context information. *Computational intelligence and neuroscience*, *2015*.
13. Barkur, G., Vibha, & Kamath, G. B. (2020). Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. *Asian Journal of Psychiatry*, *51*, 102089. https://doi.org/10.1016/j.ajp.2020.102089
14. Steinskog, A., Therkelsen, J., & Gambäck, B. (2017, May). Twitter topic modelling by tweet aggregation. In *Proceedings of the 21st Nordic conference on computational linguistics* (pp. 77-86).
15. Lamsal, R. (2020). *Coronavirus (COVID-19) Tweets Dataset* [Dataset]. Retrieved from https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset
16. Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010, June). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Springer, Berlin, Heidelberg.
17. Agarwal, B., & Mittal, N. (2016). Machine learning approach for sentiment analysis. In *Prominent feature extraction for sentiment analysis* (pp. 21-45). Springer, Cham.
18. Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011
19. Deveaud, R., San Juan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84
20. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101
21. Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445-1456).

◆◆