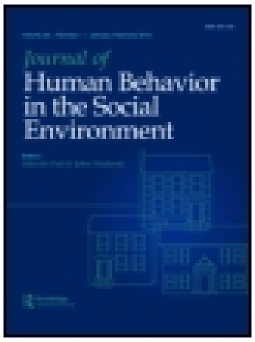


Journal of Human Behavior in the Social Environment

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/whum20

Characterizing public emotions and sentiments in COVID-19 environment: A case study of India

Subasish Das & Anandi Dutta

JOURNAL OF HUMAN BEHAVIOR IN THE SOCIAL ENVIRONMENT
https://doi.org/10.1080/10911359.2020.1781015

Routledge
Taylor & Francis Group

Check for updates

Characterizing public emotions and sentiments in COVID-19 environment: A case study of India

Subasish Das [a] and Anandi Dutta [b]

[a]Texas A&M Transportation Institute, Texas A&M University, College Station, Texas, USA; [b]Department of Computer Science, The University of Texas at San Antonio, San Antonio, Texas, USA

ABSTRACT

Coronavirus 2019, or COVID-19, is a contagious disease triggered by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). With origins in Wuhan, China, this disease has since spread globally, resulting in the ongoing 2019–2020 coronavirus pandemic. As of May 3, 2020, the Ministry of Health and Family Welfare confirmed a total of 39,980 positive COVID-19 cases and 1,301 deaths in India (more than 3.42 million positive COVID-19 cases resulting in more than 243,000 deaths worldwide). To flatten the curve, India has been locking down its country from March 24 to May 17, 2020. This study collected "COVID-19 in India" related tweets (totaling 410,643 tweets in English) from March 22 to April 21, 2020 to gauge the unknowns and contexts associated with public sentiments during the lockdown. This work contributes to the growing body of studies on COVID-19 social media mining by extracting emotions and sentiments over time, which could potentially shed some lights on the contexts of expressions during pandemic.

KEYWORDS
COVID-19; social media mining; Twitter mining; emotion mining; sentiment analysis; topic model

Introduction

Coronaviruses are a family of viruses that can trigger illnesses such as the Middle East respiratory syndrome (MERS) and the common cold, severe acute respiratory syndrome (SARS). In 2019, a disease outbreak that originated in China was identified as a new form of coronavirus. This virus is now referred to as severe acute respiratory syndrome coronavirus 2 or SARS-CoV-2. The resulting disease of SARS-CoV-2 is called coronavirus disease 2019 (COVID-19). In March 2020, the COVID-19 outbreak was declared a pandemic by the World Health Organization (World Health Organization [WHO], 2020). As of May 3, 2020, more than 3.42 million COVID-19 cases have been reported from 187 countries and territories. These cases have resulted in more than 243,000 deaths. India has tested 39,980 total positive cases and claimed 1,301 deaths as of May 3, 2020 (Dong et al., 2020). On March 22, 2020, a 14-hour voluntary public curfew (#JanataCurfew) at the instance of the prime minister Narendra Modi was observed. The curfew was followed up with lockdowns in 75 districts where COVID cases were rising as well as all major cities. Furthermore, on March 24, the Prime Minister of India ordered a nationwide lockdown lasting 21 days. On April 14, the ongoing nationwide lockdown was extended until May 3, which was later extended until May 17, 2020 (Press Information Bureau [PIB], 2020).

CONTACT Subasish Das s-das@tti.tamu.edu Texas A&M Transportation Institute, Texas A&M University, College Station, TX, USA

A United Nations (UN) report projected a potential trade impact of 348 USD million USD on India due to COVID-19 pandemic. Asian Development Bank (ADB) projected that the outbreak could cause losses of up to 29.9 USD billion USD to India's economy (BBC News, 2020). A recent survey study (Roy et al., 2020) showed that the anxiety levels of the participants were high. Over 80% of the participants were concerned with the impact of COVID-19 in their lives.

In recent years, social media platforms have become very popular discussion forums for many users around the world. Twitter allow users to post texts (known as tweets) with up to 280-characters expressing thoughts, information, comments, and observations. In many cases, the texts associated with the tweets are trended based on the real-time incidents and events. A study conducted by Barkur and Vibha (2020), using 24,000 tweets during March 25–28, showed that the overall sentiment trend is positive during the early days of the lockdown. Many Indian celebrities widely participated #JanataCurfew event and tweeted positive and affirmative attitudes to tackle this pandemic. Majority of the influential people supported the actions taken by the Indian Government to handle this pandemic. As the celebrity tweets are retweeted and reshared by many of their followers, an abundance of these tweets may influence the overall positive emotion scores. There is a need for a follow-up study to reexamine the findings by using a comprehensive database. This study collected "COVID-19 in India" related tweets for 30 days (March 22 to April 21, 2020) to perform a comprehensive emotion mining and sentiment analysis to explore the contexts and trends associated with this pandemic.

The current research aims to answer two major questions: **RQ1**: What are overall sentiment trends during COVID-19 in India? and **RQ2**: What are the contexts of expressions while tweeting either positive or negative sentiments during COVID-19 in India?

Relevant studies

Studies on social media analytics during pandemic and natural disasters

Social media analysis is one of the most popular research areas in the recent days. A plethora of studies apply different natural language processing (NLP) techniques. Out of these techniques, sentiment analysis is one of the most researched NLP topic. Interested readers can consult several comprehensive literature survey studies on sentiment analysis and emotion mining (Cambria et al., 2013; Feldman, 2013; Liu, 2012; Medhat et al., 2014; Pang & Lee, 2008). The scope of the current literature review is only limited to pandemic and natural disaster related social media studies.

The Zika virus pandemic has gained huge social media attraction. It is known that precise information helps in decreasing any pandemic spread. Sharma et al. (2017) examined the effective use of the social media as an information source during this pandemic. The study found that the misleading and fake posts gained more popularity compared to resourceful and accurate information sharing via social media. Liu and Kim (2011) investigated how 13 organizations differently framed the 2009 H1N1 flu pandemic crisis using both social media and conventional media responses. The findings show that organizations did not use the opportunity of information sharing via social media. Jain and Kumar (2015) examined important issues related to 2015 H1N1 pandemic in India. The results show that Twitter can be used as an information sharing tool effectively.

Many studies examined the usefulness of social media for natural disaster communication (Imran et al., 2015; Kryvasheyeu et al., 2016; Li et al., 2018; Wang & Ye, 2018, Wang et al., 2019). Some of these studies also examined emotions and sentiments using social media data during and after natural disasters. Chen, Mao et al. (2020) investigated sentiment and retweet patterns of disaster-affected areas and disaster-unaffected areas during Hurricane Harvey. The results show that off-site tweets generated more negative sentiments than on-site tweets. One interesting finding was that negative tweets spread faster than positive tweets. It discloses that social media users were more sensitive to negative information in disaster situations. Neppalli et al. (2017) performed a sentiment analysis of tweets during Hurricane Sandy and visualized online users' sentiments on a geographical map centered around the hurricane. Badmus (2020) examined sentiment measures from various virtual communities formed in the post-Hurricane Dorian period.

Studies on COVID-19 social media analytics

There has been a recent surge of COVID-19 studies; however, exploration on "COVID-19 and social media" is limited. With the promotion of citizen engagement during the COVID-19 crisis, Chen, Min. et al. (2020) systematically investigated how Chinese central government organizations used social media platform during unprecedented events. Using data gathered from "Healthy China," an official Sina Weibo account of the National Health Commission of China, this study examined several issues including contexts, content type, dialogs, sentiments and emotions. The results showed that citizen engagement is positively affected by the government's handling of event and the information relating to the latest news about the crisis through government social media. Concerns about the misinformation generated on social media during a pandemic like COVID-19 were examined by Limaye et al. (2020) as a commentary article. Lu and Zhang (2020) analyzed the largest social media in China, WeChat, to determine the trends about COVID-19. Rajkumar (2020) conducted a literature survey on "COVID-19 and mental health" and chose four studies to be the original research articles relevant to the topic. The remaining 24 articles consisted of letters to the editor and editorials (n = 16) or commentary related to mental health and COVID-19 (n = 8). The findings show that symptoms of depression and anxiety (16–28%) and self-reported stress (8%) were common psychological responses to the COVID-19 that could be associated with disrupted sleep.

The key findings of the literature review show that the research trends of sentiment analysis using social media data during pandemic and natural disasters are still evolving. There is a critical need for practice-ready studies so that authorities can make data-driven decision making from the insights of social media mining in real-time. This study is contributing to the growing body of this research area by analyzing a larger set of Twitter data during the early months of COVID-19 outbreak in India.

Data collection and exploratory data analysis

Data collection

Barkur and Vibha (2020) collected tweets using two hashtags "#IndiaLockdown" and "#IndiafightsCorona." With at least one of the hashtags biased toward positive sentiments, this study used a different approach to collect the relevant Twitter data. This study collected relevant data from a large-scale COVID-19 Twitter chatter dataset (Banda et al., 2020). This study collected all tweet IDs, during March 22, 2020-April 21, 2020, from this study. According to the data sharing policy of Twitter, only tweet IDs can be publicly shared. This study used Hydrator, a free open source Desktop application developed by Ed Summers (2017), which allows users to extract detailed information about a tweet from the tweet IDs. Data collected from hydration includes user profile information such as location, tweet information like interactions and content, as well as much more. From the large set of collected twitter data, this study applied text extraction methods to identify the tweets related to India by using keywords including "India" and "Indians." The current data collection is limited to "English" language tweets only. The final dataset used in this study contains 410,643 unique tweets.

Exploratory data analysis

Figure 1 illustrates the number of tweets, retweets, and favorites in the collected dataset. The frequencies of tweets (maximum 16,163 tweets on April 5, 2020) and retweets (maximum 97,124 retweets of the original tweets on April 4, 2020) show that peaks occurred during the first week of April 2020. The maximum number of favorites was 374,802, which occurred on April 9, 2020. The preliminary exploration also shows that the top public officials can reach a wide audience of citizen engagement through Twitter. The official Twitter handle (@PMOIndia) of the honorable Prime Minister of India has 34.9 million followers (with an additional 56.3 million followers in his personal Twitter handle, @narendramodi). Many of his COVID-19 tweets have been extensively retweeted and favorited by his followers. Many of the Indian cinema (known as Bollywood) and sports celebrities have large Twitter fan bases. India, as a nation, is deeply driven by its cinema and sports celebreties. Positive and inspiring tweets from these influencial accounts usually reach to wide Twitter audience due to the high number of retweets.

Before proceeding emotion mining and sentiment analysis, the tweets are pre-processed to produce clean texts for NLP tasks. Additionally, tweets contain hashtags (#), mentions (@), and website links (www., http://). In the text cleaning work, these elements are removed from

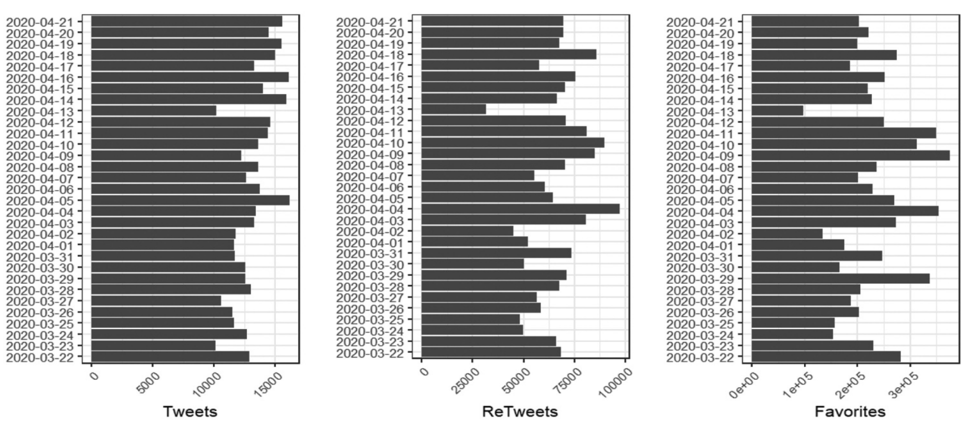


Figure 1. Number of tweets, retweets, and favorites of the collected Twitter dataset.

the text. Emoticons (i.e. shorthand for facial expressions) and emojis (i.e. pictographs) were also removed from the current study design. As the emoticons and emojis are often needed to be cross-examined against the sentiment score generated from corpus, the current study did not explore this option due to the larger set of data and associated processing time. Additional basic text cleaning procedures (removing punctuation marks and extra spaces; removing high-frequency words such as India, Indians, COVID, COVID-19, corona, coronavirus; removing conventional stop-words) were conducted to prepare the dataset ready for lemmatization. In many cases, some of the words (e.g., include, includes, including, inclusion, included) mean similar thing. Instead of using all of these words, a single representative word would be sufficient. Lemmatization, an improved version of word stemming, uses morphological analysis of words to remove inflectional endings. This study used open source R software "korpus" to perform lemmatization (Michalke et al., 2018). After lemmatization, the dataset was ready for the next step.

Emotion mining and sentiment analysis

To answer **RQ1**, this study performed two analytical procedures: 1) perform a generalized emotion mining using larger texts such as daily level total tweets, and 2) develop sentiment scores by analyzing each tweets with associated sentiment scores.

Corpus-level emotion mining using NRC lexicon

Initially, this study conducted a preliminary analysis on corpus (a set of texts based on criteria; for example, all tweets based on a date is considered as a corpus in the study design in this section) level emotion mining. The NRC Emotion Lexicon contains a list of English words and their associations with eight basic emotions (anticipation, anger, surprise, fear, trust, joy, sadness, and disgust) and two sentiments (positive and negative). Ekman (1992) argued that there are only six basic emotions: joy, sadness, anger, fear, disgust, and surprise. However, Plutchik (1994) proposed a theory that identified eight basic emotions, including Ekman's six emotion categories along with trust and anticipation. Plutchik organized the emotions into a wheel where the radius indicates intensity (the intensity was higher closer to the center). Plutchik claimed that the eight basic emotions form four opposing pairs: joy–sadness, trust–disgust, anger–fear, and anticipation–surprise (Mohammad & Turney, 2013).

Table 1 shows that the mean ratio of sentiment pairs (negative/positive) from is 0.666 (standard deviation = 0.046). The mean ratios of the emotion pairs (sadness/joy, fear/anger, disgust/trust, surprise/anticipation) are 0.773 (standard deviation = 0.082), 1.124 (standard deviation = 0.095), 0.307 (standard deviation = 0.032) and 0.413 (standard deviation = 0.043), respectively. The corpus-level analysis shows that positive sentiments are much higher compared to negative sentiments. Similarly, negative emotions are less frequent than positive emotions. Barkur and Vibha (2020) also found similar findings.

Table 1. Word counts based on NRC sentiment categories.

Date	positive	negative	joy	sadness	anger	fear	trust	disgust	anticipation	surprise
3/22/2020	17,030	10,708	5,679	2,965	6,490	6,350	9,605	2,140	5,634	2,035
3/23/2020	11,419	7,423	3,482	2,491	3,768	4,726	6,630	2,114	4,074	1,376
3/24/2020	14,391	8,472	4,491	2,787	4,254	5,519	8,744	2,168	5,057	1,845
3/25/2020	12,574	8,500	4,040	3,220	4,362	5,986	7,538	2,223	4,523	1,810
3/26/2020	13,650	7,989	4,349	3,009	4,255	5,263	8,198	2,439	4,918	1,828
3/27/2020	11,789	7,716	3,569	2,717	4,352	5,128	7,063	2,053	4,075	1,676
3/28/2020	14,935	9,393	4,586	3,077	5,305	5,826	8,703	2,361	5,096	2,042
3/29/2020	14,585	9,411	4,415	3,200	5,411	6,152	8,552	2,387	4,892	1,967
3/30/2020	14,095	8,832	4,345	3,270	4,691	5,932	8,158	2,590	4,820	2,022
3/31/2020	12,938	8,868	4,029	3,148	4,819	5,696	7,942	2,456	4,445	1,620
4/1/2020	12,654	8,937	4,024	3,201	4,994	5,513	7,594	2,585	4,197	1,788
4/2/2020	13,076	9,379	4,069	3,250	5,177	5,782	7,785	2,902	4,787	1,784
4/3/2020	15,486	10,177	4,664	3,291	5,854	6,277	9,604	2,643	5,482	1,930
4/4/2020	14,761	9,857	4,706	3,330	5,653	6,464	8,945	2,869	5,456	2,453
4/5/2020	18,261	13,897	6,659	4,848	8,580	8,086	11,347	3,629	7,683	3,386
4/6/2020	14,634	9,966	4,653	3,813	5,770	6,299	9,163	2,798	6,002	2,447
4/7/2020	13,582	9,467	4,684	3,826	5,346	5,825	7,854	2,615	5,046	2,780
4/8/2020	14,393	9,261	4,524	3,497	5,365	5,920	8,504	2,834	5,460	2,285
4/9/2020	12,967	8,724	4,120	3,406	4,798	5,369	7,740	2,336	4,766	1,927
4/10/2020	14,582	9,826	4,566	3,874	5,625	6,397	8,646	3,093	5,309	2,322
4/11/2020	15,812	9,713	4,933	3,989	5,353	6,197	9,194	3,055	5,632	2,295
4/12/2020	15,394	10,107	5,221	4,174	5,659	6,431	9,351	3,067	5,863	2,584
4/13/2020	11,180	7,166	3,512	2,668	3,886	4,413	6,475	1,948	4,404	1,756
4/14/2020	17,740	10,934	5,250	3,820	6,668	6,818	11,193	2,852	5,997	2,149
4/15/2020	14,828	10,033	4,631	3,894	5,632	6,221	9,184	2,988	5,263	2,172
4/16/2020	17,308	11,834	5,499	4,609	6,796	7,063	10,648	3,332	5,827	2,552
4/17/2020	13,899	9,176	4,403	3,618	5,552	5,768	8,946	2,695	4,588	2,034
4/18/2020	15,699	11,423	5,077	4,160	7,228	6,942	9,788	3,029	5,920	2,622
4/19/2020	15,855	12,655	5,324	5,004	7,511	7,906	10,067	3,608	6,057	2,790
4/20/2020	14,871	10,316	4,511	4,071	5,783	6,687	9,434	2,909	5,308	2,315
4/21/2020	17,851	11,196	5,668	4,815	6,265	7,064	11,234	3,179	5,698	2,644

Sentiment analysis

Sentiment analysis using individual tweet sentiment score

The previous section (emotion mining) used a conventional approach to provide a broad picture of the corpus-level emotion mining. Conventional emotion mining and sentiment analysis techniques typically measure emotions and sentiments in the form of eight emotions or three sentiments (positive/negative/neutral) by matching words using a sentiment lexicon or dictionary. In many cases, this approach can provide some context. Language is a complex system, so simplified emotion mining or sentiment analysis would not be sufficient in gauging the complex sentiments during a pandemic like COVID-19. Without counting the valence shifters, words that can alter the meaning of sentiment associated with a word, the generated sentiment score can be biased. The open source R software package "sentimentr" incorporates several key valence shifters that conduct an additional layer of recalculation on the sentiment scores derived from conventional lexicon-based sentiment analysis (Rinker, 2016). Some of the common valence shifters are: negators (e.g., no, don't, didn't), amplifiers (e.g., surely, absolutely), de-amplifiers (e.g., nearly, hardly, almost), and adversative conjunctions (e.g., but, though, being said that). Table 2 lists some of the example tweets that generate borderline sentiment scores. Table 3 lists the average and standard deviations of the sentiment cores by date.

Table 2. Example tweets with sentiment (Score).

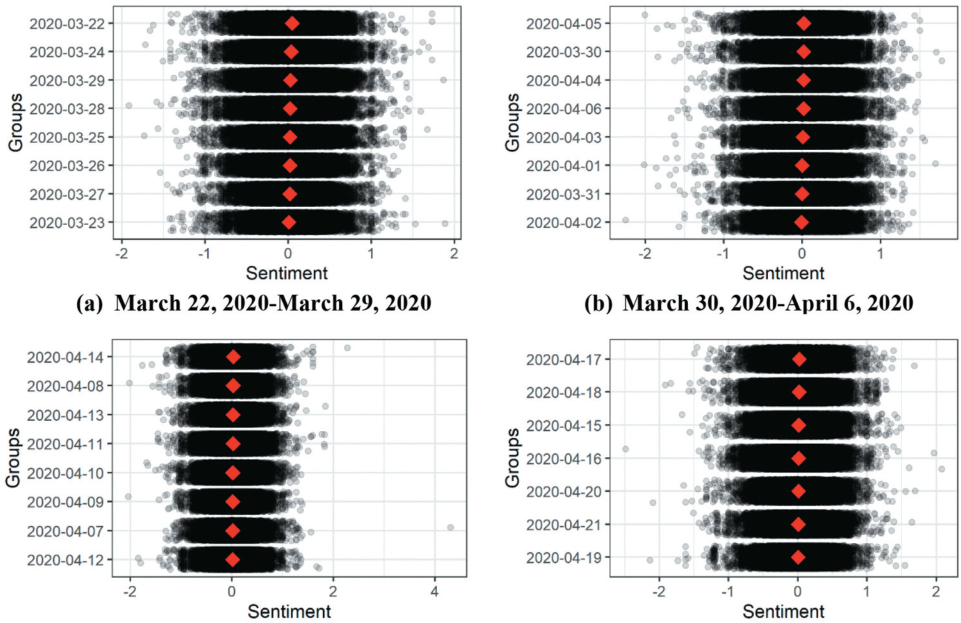
Tweet[a]	Date	Sentiment (Score)
We love u sid #stay safe stay at home keep safe #sideart #safe at home india keep safe n stay home love love love much love to u sid	Apr 16, 2020	2.083
Sir We all pray for you and mother India you are doing a great effort and doing very well we all support you Folded handsall will be fine very soon Wishing you all the very best what ever you do,and lucky India to have such a wonderful leader God bless you	Apr 13, 2020	1.846
Buying a protective mask can be out of reach to poor people due to high prices in india. at 8 centers, 3,200 masks have been stitched and distributed to those in need in neighboring villages.#covid19 #india #coronavirus #facemasks #salesian #wearedonbosco	Apr 2. 2020	0.00
Some purposely ignored police warnings Action with extreme prejudice be taken as fr murder n life threatening assault Severe financial penalties be imposed	Apr 1, 2020	−1.855
If #covid19 were a human, though most citizens would have tried their best to get it to quit india but xxxx2 would have been busy expression his grief on ill treatment and anguish to #covid19 by the yyyy2 …	Apr 9, 2020	−2.039

[a]Exact tweets are used as example (grammatical edits are not done).
[b, c]Removed Twitter mentions and emoticons from the text.

Table 3. Average and standard deviation of sentiment scores by date.

Date	Number of Tweets	Number of Retweets	Average Sentiment	Std. of Sentiment
2020–03-22	12,888	68,321	0.051	0.242
2020–03-23	10,122	66,155	0.012	0.250
2020–03-24	12,683	49,670	0.038	0.247
2020–03-25	11,627	48,086	0.024	0.248
2020–03-26	11,517	58,477	0.023	0.241
2020–03-27	10,617	56,437	0.023	0.239
2020–03-28	12,995	67,492	0.025	0.249
2020–03-29	12,594	71,276	0.029	0.250
2020–03-30	12,575	50,065	0.022	0.249
2020–03-31	11,730	73,457	0.001	0.247
2020–04-01	11,633	52,250	0.002	0.250
2020–04-02	11,755	45,096	−0.007	0.253
2020–04-03	13,284	80,676	0.012	0.239
2020–04-04	13,406	97,124	0.019	0.244
2020–04-05	16,163	64,282	0.022	0.240
2020–04-06	13,726	60,346	0.019	0.240
2020–04-07	12,653	55,182	0.016	0.251
2020–04-08	13,609	70,472	0.029	0.237
2020–04-09	12,220	85,158	0.019	0.240
2020–04-10	13,627	89,786	0.020	0.248
2020–04-11	14,419	81,185	0.024	0.244
2020–04-12	14,588	70,766	0.015	0.248
2020–04-13	10,205	31,491	0.026	0.242
2020–04-14	15,934	66,347	0.033	0.236
2020–04-15	14,010	70,511	0.014	0.242
2020–04-16	16,126	75,671	0.013	0.246
2020–04-17	13,311	57,687	0.025	0.239
2020–04-18	15,004	85,834	0.020	0.241
2020–04-19	15,516	67,765	0.008	0.248
2020–04-20	14,503	69,638	0.010	0.243
2020–04-21	15,603	69,510	0.009	0.245

Figure 2 illustrates the scatter plots (each dot indicates sentiment score of a tweet) of individual tweet sentiment scores by date. Four different temporal groups (March 22, 2020-March 29, 2020; March 30, 2020-April 6, 2020; April 7, 2020-April



(a) March 22, 2020-March 29, 2020 **(b) March 30, 2020-April 6, 2020**

(c) April 7, 2020-April 14, 2020 **(d) April 15, 2020-April 21, 2020**

Figure 2. Scatter plot of the sentiment scores by date.

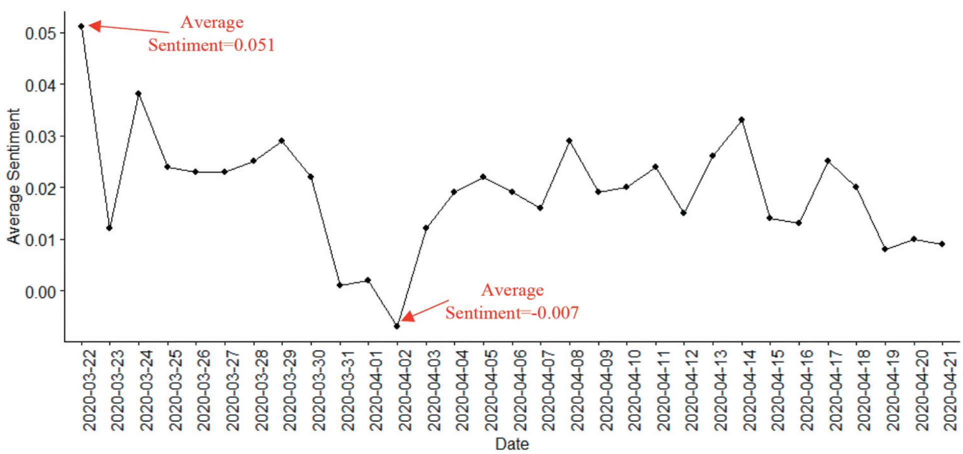


Figure 3. Line chart of daily average sentiment scores.

14, 2020; April 15, 2020-April 21, 2020) were created to show the distribution of the scores. Each red dot indicates the mean value of the sentiment scores for the date. The y-axis of the illustrations is sorted in descending order based on the average sentiment score of the daily scores. For example, April 2, 2020 has the lowest average sentiment score (−0.007), which is shown at the bottom of Figure 2(b). The average sentiment score by date is presented in Figure 3. There were seven peaks of the scores during the period. The first peak (n = 0.0501), which is also the highest, was on March 22, 2020.

The overall trend is declining until April 2, 2020 (with −0.007 as the lowest sentiment score). The trend after April 2, 2020 is upwards until April 17, 2020. Starting from April 18, 2020, the trend is again declining.

Contexts of expressions

To answer **RQ2**, this study developed Latent Dirichlet Allocation (LDA) topic models using two broader groups: 1) tweets with a tweet level positive sentiment score (sentiment score greater than zero) and 2) tweets with a tweet level negative sentiment score (sentiment score less than zero).

Topic models based on positive and negative sentiment corpus

This study used Python tools to develop the sentiment-based topic models. This study used an additional lemmatizer from "nltk" package and used "CountVectorizer" in scikit-learn to perform the token counting. This study also consulted spaCy to perform easy-to-use tokenization and lemmatization functions (Amazon Web Service Labs [AWSLabs], 2020). While performing the preliminary topic models, it was noticed that there is presence of some Islamophobic hashtags (e.g., #coronajihad) in the collected tweets. In a news article, Time magazine mentioned that several Islamophobic hashtags (e.g., #coronajihad, #tablighijamaatvirus, #Nizamuddinidiots) began circulating shortly after a false news claiming that a Muslim man from the Delhi congregation (a gathering by the

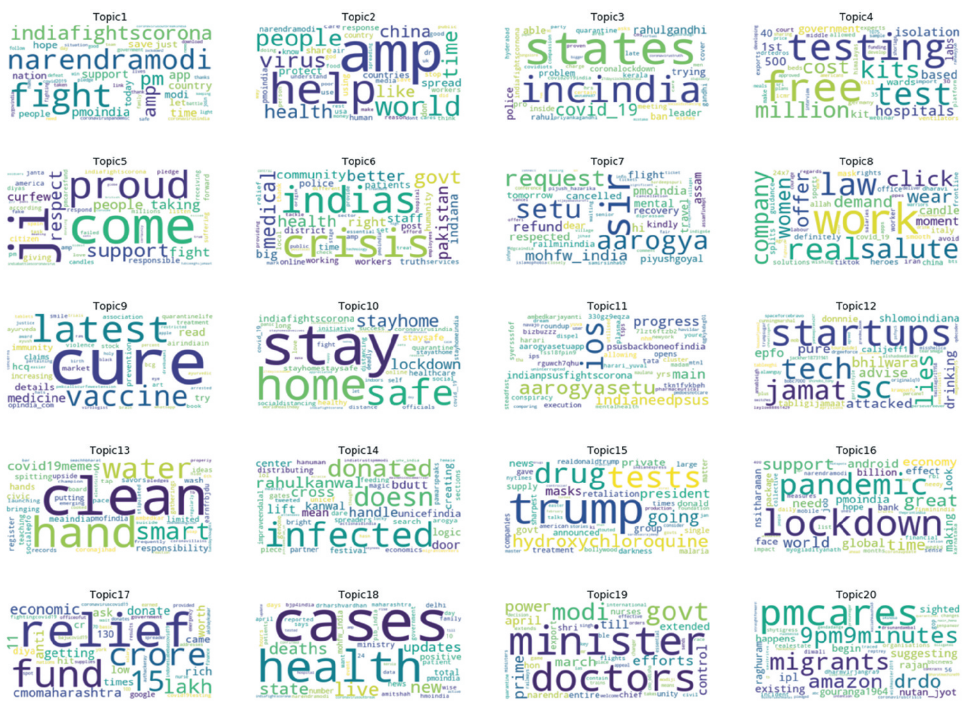


Figure 4. Wordcloud of top 20 topics from positive tweet group.

Figure 5. Wordcloud of top 20 topics from negative tweet group.

Tablighi Jamaat, an Islamic missionary movement, occurred at Delhi's Nizamuddin area) intentionally coughed on somebody (Time Magazine, 2020). This study removed all tweets (associated with the most frequent Islamophobic keywords) before performing the final topic model development because these tweets overall do not represent the general contexts of expressions of mass people of India. This study used Latent Dirichlet Allocation (LDA) topic model to perform the analysis. In 2003, Blei et al. (2003) developed the LDA model to report the issues found in the probabilistic latent semantic analysis (PLSI) mode. Improving upon the PLSI model, the LDA model uses a K-dimensional latent random variable. The LDA model has been proven as one of the most widely used topic models (Das et al., 2017, 2020, 2016).

In Figures 4 and 5, the size of a word shown in a topic's word cloud is proportional to the probability of the word occurring within the topic. The top 20 most likely topics are displayed in each figure based on the respective sentiment groups (positive or negative). The labels were chosen by looking at the occurrence probability of the top 50 words for each topic and examining the tweets that are most relevant per topic.

Figure 4 contains the top twenty topics gathered from positive tweets. The Indian Prime Minister, Narendra Modi, is included in four of the topic groups (Topic 1, Topic 5, Topic 19, and Topic 20). As mentioned earlier, many celebrities broadly accepted Modi's #JanataCurfew and other COVID-19 actions and widely circulated his tweets and messages via Twitter, encouragement and positive sentiments were expressed vigorously in these tweets. The U.S. President, Donald Trump, is mentioned in a topic that also includes "drug," "tests," and "hydroxychloroquine" (Topic 15). The word "testing" is also shown in Topic 4 along with "free," "kits," "million," and "test," demonstrating the interest of Indian citizens with the cost and availability of tests (Topic 4). Some topics include words such as "lockdown" (Topic 16) and "stay home" (Topic 10), while others include positive emotions such as "proud" (Topic 5) and "appreciation" (Topic 8). The other top keywords in the topics are Aarogya Setu, an Indian COVID-19 mobile tracking app developed by the National Informatics Center (Topic 7, Topic 11), startups (Topic 12), hand cleaning (Topic 13), relief fund (Topic 17), stay home (Topic 10), vaccine (Topic 9), and appreciation (Topic 8).

Figure 5 shows the top twenty topics gathered from negative tweets. As with the positive topics, Prime Minister Narendra Modi and politics are mentioned in four topics (Topic 3, Topic 5, Topic 6, Topic 19). Other nations are mentioned in several topics; these include Pakistan (Topic 3), China (Topic 11, 12, and 17), and the United States (Topic 10). The continent Africa is also included in a topic (Topic 4). Although frequent racial and communal keywords are removed in this analysis, Topic 17 shows some of xenophobic and racial contexts in the form of words such as "chinesevirus19," "evils," "communalizing," and "nizamuddinmarkaz." As mentioned earlier, these contexts are limited to only a certain group of people. Other words include "death" and "risk" (Topic 2, Topic 7, Topic 9, and Topic 13), "news" and "app" (Topic 16, and Topic 20), "lockdown" (Topic 1, and Topic 18), "new cases" (Topic 2, and Topic 14), and "economy" (Topic 18).

Conclusion

It is important to note that the average middle-class citizen is currently turning to social media for information and advice. Exploration of sentiment and emotions using social media mining has become a trending research in the recent years. This study extended Barkur and Vibha (2020) study with inclusion of larger dataset and advanced algorithms. This study makes the following two contributions to the literature. First, to the best of our knowledge, this study used the largest data samples possible to examine the contexts and unknowns associated with sentiments and emotions of Indians during COVID-19. Second, it used topic modeling and sentiment analysis together to produce topics in different sentiment groups.

This study used two methods to answer **RQ1** by performing the emotion mining and sentiment analysis: 1) simplified emotion mining using broad corpus-level data, 2) complex sentiment analysis at individual tweet level using valence shifters. For the broad corpus-level analysis (corpus based on date), the context of positiveness is significantly higher than negative sentiments. This finding is in line with Barkur and Vibha (2020). Barkur and Vibha (2020) explained that "it can be seen that Indians have taken the fight against COVID19 positively and majority are in agreement with the government for announcing the lockdown to flatten the curve." However, the positive sentiment trends are not so different from the negative sentiment trends when the analysis was performed at individual tweet level. The results of the second analysis show that the discussion of "COVID-19 in India" on Twitter contains slightly more positive sentiments than negative sentiments. The second analysis supplies a rigorous view of the overall sentiments, as it considers valence shifters to remove biases in sentiment scores. Additionally, LDA topic models were developed to answer **RQ2**. The wordclouds showing 20 topic models from each sentiment group provide additional insights and ongoing contexts of expression during the lockdown. Overall, the topics capture different contexts of expressions (for example, political leadership, availability of testing, hygiene practice, popular COVID-19 app Aarogya Setu, vaccine, international relationship with other countries, economy, risk and death, new cases) of the sentiment based corpora.

By tracing the variation of sentiments and emotions during May-April of 2020, this study examined the evolution of public attitude toward this crisis in India. This work contributes to the growing body of studies on social media mining during COVID-19. This study initiates an approach in extracting emotions and sentiments over time, which could potentially shed some lights on the contexts of expressions during pandemic.

One major limitation of this study is the nature of the Twitter dataset itself. First, this dataset only considers English language tweets. Additionally, the current data is limited to March 22 to April 21, 2020. Second, additional works are needed to measure the contexts associated with the topic. Furthermore, the association of topics with the timeline of the tweets can provide additional perceptions about the nature and trends of the tweets. Third, only a small percentage of the tweets contain geo-locations. It might be a future research area to perform context mining with the usage of the spatio-temporal COVID-19 cases and fatalities can unearth local contexts and public sentiments. Fourth, added investigation is needed to understand the perceptions and contexts associated with the negative sentiment and emotion associated tweets. Limitations of the current study can be improved in the future studies.

Acknowledgments

The authors like to thank Ly-Na Tran and Magdalena Theel for their initial review and feedback.

Funding

This research was not supported by any funding.

ORCID

Subasish Das http://orcid.org/0000-0002-1671-2753

Declaration of interest statement

There is no conflict of interest.

Author contribution statement

The authors confirm the contribution to the paper as follows: study conception and design: Subasish Das, Anandi Dutta; data collection: Subasish Das; analysis and interpretation of results: Subasish Das, Anandi Dutta; draft manuscript preparation: Subasish Das, Anandi Dutta. All authors reviewed the results and approved the final version of the manuscript.

References

Amazon Web Service Labs (AWSLabs). (2020). *Amazon Sagemaker examples*. Amazon Web Serice Labs. Retrieved May 3, 2020, from https://github.com/awslabs

Badmus, M. O. (2020). When the storm is over: Sentiments, communities and information flow in the aftermath of Hurricane Dorian. *International Journal of Disaster Risk Reduction*, *47*, 101645. https://doi.org/10.1016/j.ijdrr.2020.101645

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., & Chowell, G. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research–an international collaboration. *arXiv Preprint*, *arXiv:2004.03688*.

Barkur, G., & Vibha, G. B. K. (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

BBC News. (April 3, 2020). *Coronavirus: India's bailout may not be enough to save economy*. British Broadcasting Corporation (BBC). https://www.bbc.com/news/world-asia-india-52117704

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15–21. https://doi.org/10.1109/MIS.2013.30

Chen, Q., Min, C., Zhang, W., Wang, G., Ma, X., & Evans, R. (2020). Unpacking the black box: How to promote citizen engagement through government social media during the COVID-19 crisis. *Computers in Human Behavior*, *110*, 106380. https://doi.org/10.1016/j.chb.2020.106380

Chen, S., Mao, J., Li, G., Ma, C., & Cao, Y. (2020). Uncovering sentiment and retweet patterns of disaster-related tweets from a spatiotemporal perspective – A case study of Hurricane Harvey. *Telematics and Informatics*, *47*, 101326. https://doi.org/10.1016/j.tele.2019.101326

Das, S., Dixon, K., Sun, X., Dutta, A., & Zupancich, M. (2017). Trends in transportation research: Exploring content analysis in topics. *Transportation Research Record: Journal of the Transportation Research Board*, *2614*(1), 27–38. https://doi.org/10.3141/2614-04

Das, S., Dutta, A., & Brewer, M. (2020). Transportation research record articles: A case study of trend mining. *In Transportation Research Record*. (in press).

Das, S., Sun, X., & Dutta, A. (2016). Text mining and topic modeling of compendiums of papers from transportation research board annual meetings. *Transportation Research Record: Journal of the Transportation Research Board*, *2552*(1), 48–56. https://doi.org/10.3141%2F2552-07

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82–89. https://doi.org/10.1145/2436256.2436274

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, *47*(4), 1–38. https://doi.org/10.1145/2771588

Jain, V. K., & Kumar, S. (2015). An effective approach to track levels of influenza-A (H1N1) pandemic in India using twitter. *Procedia Computer Science*, *70*, 801–807. https://doi.org/10.1016/j.procs.2015.10.120

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*(3), e1500779. https://doi.org/10.1126/sciadv.1500779

Li, Z., Wang, C., Emrich, C. T., & Guo, D. (2018). A novel approach to leveraging social media for rapid flood mapping: A case study of the 2015 South Carolina floods. *Cartography and Geographic Information Science*, *45*(2), 97–110. https://doi.org/10.1080/15230406.2016.1271356

Limaye, R. J., Sauer, M., Ali, J., Bernstein, J., Wahl, B., Barnhill, A., & Labrique, A. (2020). Building trust while influencing online COVID-19 content in the social media world. *The Lancet Digital Health*, *2*(6), e277–e278. https://doi.org/10.1016/S2589-7500(20)30084-4

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016
Liu, B. F., & Kim, S. (2011). How organizations framed the 2009 H1N1 pandemic via social and traditional media: Implications for US health communicators. *Public Relations Review*, *37*(3), 233–244. https://doi.org/10.1016/j.pubrev.2011.03.005
Lu, Y., & Zhang, L. (2020). Social media WeChat infers the development trend of COVID-19. *The Journal of Infection*. 81, 82-83. https://doi.org/10.1016/j.jinf.2020.03.050
Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011
Michalke, M., Brown, E., Mirisola, A., Brulet, A., & Hauser, L. (2018). *koRpus: An R package for text analysis*. The Comprehensive R Archive Network. Retrieved May 3, 2020, from https://cran.r-project.org/web/packages/koRpus/koRpus.pdf
Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x
Neppalli, V. K., Caragea, C., Squicciarini, A., Tapia, A., & Stehle, S. (2017). Sentiment analysis during Hurricane Sandy in emergency response. *International Journal of Disaster Risk Reduction*, *21*, 213–222. https://doi.org/10.1016/j.ijdrr.2016.12.011
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011
Plutchik, R. (1994). *The psychology and biology of emotion*. HarperCollins College Publishers.
Press Information Bureau (PIB). (2020). *PIB's Special Webpage on COVID19*. Government of India. https://pib.gov.in/newsite/coronaviruss.aspx
Rajkumar, R. P. (2020). COVID-19 and mental health: A review of the existing literature. *Asian Journal of Psychiatry*, *52*, 102066. https://doi.org/10.1016/j.ajp.2020.102066
Rinker, T. W. (2016). *sentimentr: Calculate text polarity sentiment. University at Buffalo/SUNY, Buffalo, New York*. version 0.5, 3.
Roy, D., Tripathy, S., Kar, S. K., Sharma, N., Verma, S. K., & Kaushal, V. (2020). Study of knowledge, attitude, anxiety & perceived mental healthcare need in Indian population during COVID-19 pandemic. *Asian Journal of Psychiatry*, *51*, 102083. https://doi.org/10.1016/j.ajp.2020.102083
Sharma, M., Yadav, K., Yadav, N., & Ferdinand, K. C. (2017). Zika virus pandemic—analysis of Facebook as a social media health information platform. *American Journal of Infection Control*, *45* (3), 301–302. https://doi.org/10.1016/j.ajic.2016.08.022
Summers, E. (2017). DocNow Hydrator: GitHub repository. Github. https://github.com/DocNow/hydrator.
Time Magazine (April 3, 2020). It was Already Dangerous to Be Muslim in India. Then Came the Coronavirus. Time Magazine. Retrieved May 3, 2020 from https://time.com/5815264/coronavirus-india-islamophobia-coronajihad/
Wang, Z., Lam, N. S., Obradovich, N., & Ye, X. (2019). Are vulnerable communities digitally left behind in social responses to natural disasters? An evidence from Hurricane Sandy with Twitter data. *Applied Geography*, *108*, 1–8. https://doi.org/10.1016/j.apgeog.2019.05.001
Wang, Z., & Ye, X. (2018). Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, *32*(1), 49–72. https://doi.org/10.1080/13658816.2017.1367003
World Health Organization (WHO). (January 12, 2020). *Novel coronavirus – China*. World Health Organization. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/