Sentiment Analysis and Topic Modelling of Indian Government's Twitter Handle #IndiaFightsCorona

Christina Sanchita Shah, M. P. Sebastian

Sentiment Analysis and Topic Modelling of Indian Government's Twitter handle *#IndiaFightsCorona*

Christina Sanchita Shah and M.P. Sebastian
Indian Institute of Management Kozhikode, Kozhikode, India
{christinas12fpm, sebasmp}@iimk.ac.in

Abstract. The purpose of this study was to conduct opinion mining on twitter data containing *"#IndiaFightsCorona"* to analyse public opinion. This was accomplished using sentiment analysis and topic modelling . First, sentiment analysis was done and positive and negative sentiments were separated. Then, on each sentiment topic modelling was done to discover hidden topics. Two approaches were used namely Latent Semantic Analysis (LSA) and (2) Latent Dirichlet Allocation (LDA)and then their results were compared. It was found that there were more positive sentiments than negative. For positive sentiments, LDA performed better and for negative sentiments, LSA performed better. While some topics were common between LSA and LDA for positive sentiments, there was very little overlap for negative comments.

Keywords: Topic modelling , sentiment analysis, opinion mining, #IndiaFightsCorona

1 Introduction

India reported its first case of novel coronavirus on 30th January 2020. On 11th February, the World Health Organization (WHO) announced that this coronavirus would be called COVID-19. A month later, on 11th March, WHO declared the COVID-19 outbreak as a pandemic. As of 1st April, there are almost 922,000 confirmed cases of COVID-19 out of which approximately 656,000 cases are active, 193,000 recoveries and 46,000 deaths[1]. In India, there were 1,238 confirmed cases with 32 deaths, according to the Ministry of Health and Family Welfare website. On 31st March 2020, the Ministry of Information and Broadcasting set up a dedicated Twitter handle called *@CovidnewsbyMIB* to share news and updates about novel coronavirus COVID-19. The account is named *#IndiaFightsCorona* with the handle *@ CovidnewsbyMIB*. This handle provides information on the latest number of coronavirus cases in India, and

[1]https://www.businessinsider.in/slideshows/miscellaneous/a-comprehensive-timeline-of-the-new-coronavirus-pandemic-from-chinas-first-covid-19-case-to-the-present/december-31-2019-chinese-health-officials-informed-the-world-health-organization-about-a-cluster-of-41-patients-with-a-mysterious-pneumonia-most-were-connected-to-the-huanan-seafood-wholesale-market-a-wet-market-in-the-city-of-wuhan-/slideshow/74721165.cms

various relief and economic measures[2]. While the handle *@CovidnewsbyMIB* was set up by the end of March, hashtag *#IndiaFightsCorona* has been trending even before that as indicated by Twitter data. With the rise of the COVID-19 pandemic, many countries placed restrictions on travel and movement and imposed "lockdowns". The world saw a rise in social distancing initiatives, travel bans, self-quarantines and business closures. As people could no longer access public spaces freely, a significant proportion of the dialogue on the COVID-19 pandemic shifted to online forums and social networking sites such as Twitter [1].

In light of the COVID-19 pandemic and the purpose of this study is to analyse public opinion surrounding hashtag *#IndiaFightsCorona*. Opinion mining, broadly speaking, uses text analytics to understand public sentiment. Twitter is a popular microblogging platform that people use to express themselves in real-time and can be used for opinion mining [2-4]. While there are multiple hashtags for novel coronavirus trending on Twitter, the reason for selecting *#IndiaFightsCorona* was its credibility as it is used by the Ministry of Information and Broadcasting, India. In this research, Twitter data containing *#IndiaFightsCorona* was extracted, and sentiment analysis using Support Vector Machine (SVM) classifier was done to classify data into positive and negative sentiments. The performance of the SVM classifier was measured using the Confusion Matrix. Once the sentiments had been segregated, topic modelling was conducted on each sentiment. Topic modelling discovers hidden topics/themes in each sentiment. Two approaches were used for conducting topic modelling, namely Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). Results from each method were then compared and evaluated to decide which method gave the most coherent topics within each positive and negative sentiment data. The preliminary findings of this paper are: (1) It was found that there were more positive sentiments than negative sentiments (2) For positive sentiments, LDA performed better, and for negative sentiments, LSA performed better (3) For negative sentiments, there was minimal overlap between the topics generated by LSA and LDA whereas all other topics were distinct.

1.1 Social media and previous public health crises

Research scholars have often used Twitter as a way of understanding trends visible in online social networks [2]–[5]. More specifically, Twitter offers researchers the opportunity to analyse the role of social media during a public health crisis, such as the latest COVID-19 pandemic [6]–[9]. This helps researchers investigate the social dimensions of the pandemic. While previous epidemics have demonstrated the relevance of researching social media information, there is a special significance for studying the role of social media during the current COVID-19 pandemic. In today's information age, social media is expected to play a much larger role as compared to previous health crises. For instance, during the Ebola outbreak in February 2014, Twitter comprised of approximately 255 million active users. However, Twitter had 330 million active users by 2019 [10]. Thus, a lot of people communicate online and get their news through social media sites like Twitter [11], [12]. Further, compared to past epidemics, there is

[2] https://www.theweek.in/news/india/2020/04/01/govt-launches-twitter-handle-to-smash-false-facts-on-coronavirus.html

a much higher risk of incorrect information circulating [10]. Studies have shown that the amount of misinformation available on Twitter regarding medical content is as high as 24% and is being circulated at an alarming rate [10], [13], [14]. Another study developed a multilingual COVID-19 Twitter data set to track misinformation and unverified rumours [1]. Thus, we see that social media sites can be used to spread all kinds of information. However, on the other hand, a positive aspect of the increasing presence of public opinions on social media sites is that policy makers can mine this data to understand popular discourse and develop measures to tackle the pandemic [15].

1.2 Sentiment Analysis

Opinion mining is also known as sentiment analysis which is an analysis of people's opinions, emotions, and sentiments from the written language. With the rapid growth of social media platforms such as Twitter, Facebook, Instagram, the relevance of sentiment analysis and opinion mining has increased. Opinions influence our behaviours and therefore, are essential to almost all human actions. The way we perceive and interpret the world is essentially influenced by our beliefs and interpretations of reality. This is why we always look for the inputs or opinions of others when we have to make a decision. It refers not only to individuals but to businesses as well [16]. Microblogs like Twitter are platforms where users can post their opinions, reactions, feelings, and thoughts in real-time. Some early analyses on Twitter data using sentiment analysis include Bermingham and Smeaton [6], Pak and Paroubek [4], Barbosa et al. [5], Bifet and Frank [6], Davidov et al. [7] and Agarwal et al. [19]. In this study, we use a support vector machine (SVM) classifier for sentiment classification. For the purpose of this study, we have classified sentiments only as being either positive or negative.

1.3 Topic Modelling on Twitter data

The topic model can be considered as a tool for addressing the enormous amount of data, to find hidden concepts, prominent features or latent variables [9]. It is an unsupervised natural language processing technique that extracts latent topics from a corpus of documents. There are many methods to implement topic modelling. In our study, we use two approaches: (1) Latent Semantic Analysis (LSA) and (2) Latent Dirichlet Allocation (LDA).

Latent Semantic Analysis (LSA). Latent Semantic Analysis is a single value decomposition (SVD) based algebraic process in which a bag-of-words (BoW) model is used to create a document term matrix [9]. LSA is one of the simplest topic models that are easy to understand and implement. More often than not, it gives better results than vector space models. It is also faster.

Latent Dirichlet Allocation (LDA). Latent Dirichlet Allocation learns how words, topics, and documents relate to each other by assuming that documents are generated using a particular probability model [21]. The purpose of LDA is to discover hidden topics based on data [22]. Topic in LDA is defined as "probability distribution over words". It is also a bag-of-words model [13].

Topic Coherence. Topic coherence is a method of evaluating topic models. It measures the degree of semantic similarity between its high scoring words. These measurements help distinguish between topics that are human interpretable and those that are artefacts of statistical inference. Greater coherence scores indicate greater interpretability by humans [24].

2 Research methodology

This study is divided into several stages. The first stage consists of data collection, followed by the pre-processing of data. Then comes the sentiment analysis stage and segregation of data into positive and negative sentiments. Then topic modelling is done on each sentiment using LSA, and LDA. Finally, both these approaches are evaluated using topic coherence scores. Implementation of each model is done in Python version 3.7.4.

2.1 Data Collection

Twitter data was collected using NodeXL, which is a network analysis software package for Microsoft Excel. Since the aim of the study was opinion mining of novel coronavirus COVID-19 in India, the Twitter database was searched for all tweets containing the hashtag "IndiaFightsCorona" or "indiafightscorona". 33378 tweets in all languages were collected out of which 16665 tweets were in English. For the purpose of this study, we have considered tweets written only in English. The tweets were dated between 16th March to 1st April 2020.

2.2 Data preparation

The purpose of this pre-processing of the data is to clean it and make it ready for further analysis. At the end of this stage, data is more structured and processed. This was done twice, once before sentiment analysis and once before topic modelling. The pre-processing steps were different before sentiment analysis and different before topic modelling. Combined, all the steps include [20], [24], [25]: *1) Data cleaning.* In this step, data is converted to lowercase, all punctuation and stop words such as "a"," and", "to", "the" and so on and forth, are removed. *2) Tokenization.* A tokenizer splits the document at the word level, and each word is labelled as a token. For our study, we used *word2vec* algorithm. *3) Stemming and Lemmatization.* Stemming and Lemmatization are text normalization techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing. Stemming is the process of reducing inflexion in words to their root forms.

2.3 Sentiment Analysis

To conduct sentiment analysis, SVM classifier was trained and tested. It was then used to predict the sentiment of words using their vector representation. The sentiment score

of each word was then calculated. Primarily, there are four steps in training and using the sentiment classifier: (1) Load a pre-trained word embedding. Word embeddings map words in vocabulary to numeric vectors. These embeddings capture semantic details of the words so that similar words have similar vectors. (2*)* Load an opinion lexicon listing positive and negative words[3]. (3) Train the SVM sentiment classifier to classify words into positive and negative categories. (4) Sentiment score of each word is calculated in the text, and the mean score is taken.

2.4 Topic Modelling

Once the corpus of tweets was divided into positive and negative sentiments, topic modelling was run on each sentiment to discover various topics and themes associated with each sentiment. Two approaches to topic modelling were used: Latent Dirichlet Analysis (LDA) and Latent Semantic Analysis (LSA). The results for both approaches were then compared for each sentiment using topic coherence scores to understand which approach gave better topics within each sentiment.

3 Results and Discussion

This study uses tweets containing "*#IndiaFightsCorona"* from 16th March to 1st April 2020. Only English tweets have been considered for this study amounting to 16665 tweets. We first conducted a sentiment analysis on the corpus of tweets, segregating it into positive and neutral tweets. Then topic modelling using LSA and LDA approaches was done on each sub-corpus of the sentiments. Finally, coherence scores of LSA and LDA were computed and compared to see which approach was better.

3.1 Sentiment Analysis

For the purpose of sentiment analysis, support vector machine (SVM) classifier was trained, which classifies word vectors into positive and negative categories. 10% of the corpus was set aside for testing purposes while the rest was used for training. A confusion matrix is a table that is often used to describe the performance of a classification model. Figure 1 shows the classification accuracy in a confusion matrix for SVM sentiment classifier. As can be seen, the classifier is an efficient classifier. Based on this confusion matrix, we found that Precision was 93.26%, Recall was 91.37% and F1 score was 92.30%. Next, the sentiment of all the tweets was calculated to predict the sentiment score of each word in the text. Scores greater than zero were considered as positive sentiments and scores less than zero were considered as negative sentiments. There were 13701 positive sentiments and 2964 negative sentiments.

[3] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

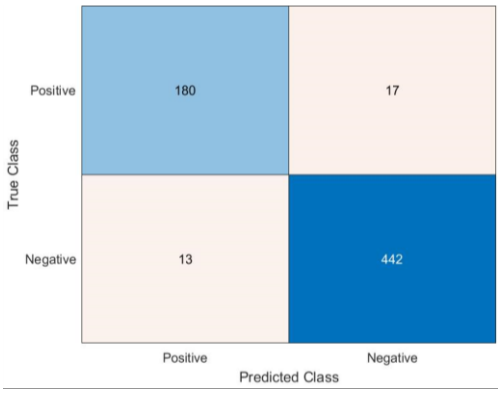

Fig. 1. Confusion matrix of SVM sentiment classifier

3.2 Topic Modelling

After the process of determining the sentiment, then the next step is to find topics within each positive and negative sentiments. Two approaches are used, namely LSA and LDA to find topics, and then their results are compared using coherence scores.

Positive Sentiments. To find the optimal number of topics, we ran the LDA model for a different number of topics' k'. Figure 2 shows the coherence chart in which we see that maximum coherence is achieved at k=13 with a coherence score of 0.426. We then ran the LDA model with K=13 and computed individual coherence of each topic. We see in Table 1 that topics 3,5,7,1,9, 6 and 11 have high coherences while the rest have lower coherence scores. Topic 3 and 5 refer to the positive reception of the news of the nationwide lockdown with many offices, Bollywood celebrities like Kartik Aryan being supportive. Topic 7 praises the lockdown as a response to the pandemic. Topic 1 refers to the State-Trait Anxiety Inventory (STAI), and how people are following the lockdown by working from home. Topic 9 is a favourable reaction to PM Cares fund by Prime Minister Narendra Modi and the measures taken by the government for its citizens. Topic 6 and 11 favour the direction of India's leadership, refer to the people at the forefront of this battle against the pandemic.

Similarly, we compute the coherence score for LSA model. As can be seen from Figure 3, the optimal number of topics is 12 with a score of 0.428. We then ran the LSA model with K=12 and then computed individual coherence of each topic. We see in Table 2 that topics 12,2,5,11,6 and 8 have high coherences while the rest have lower coherence scores.

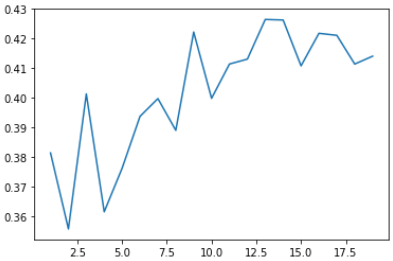

Fig. 2. Coherence chart of LDA model

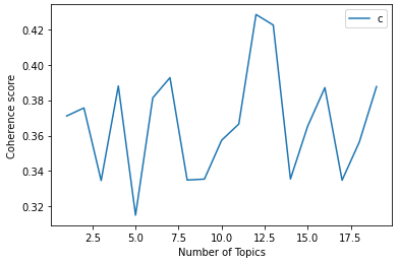

Fig. 3. Coherence chart of LSA model

Table 1. Topic words and coherence scores from LDA with k=13

Topic#	Wordlist	Coherence
3	district, given, health, privileg, office, organis, beauti, local, daily, gesture	0.000
5	celebr, modiji, small, go, bollywood, cycl, aaryan, kartik, acknowledg, virtuous	0.000
7	good, let, respons, tell, respect, sevasocieti, covid, indiathanksindiafightscorona, india, old	-0.147
1	stai, safe, inform, home, young, coronaviru, essenti, stayathomesavel, covid, work	-0.154
9	tax, govt, time, issu, citizen, ordin, pmcare, encourag, rebat, liber	-0.217
6	india, leadership, rise, associ, healthier, welfar, show, compass, coal, occas	-0.219
11	battl, happi, like, defeat, covid, forefront, hardwork, sportsperson, time, need	-0.240
12	tata, thank, nation, gestur, servic, indian, group, shri, life, develop	-0.338
13	covid, stayhom, stayhomestaysaf, coronavirusoutbreak, amp, retweet, stayathomesavel, indiafightscorona, minist, indiafightscoronaviru	-0.471
2	help, covid, indiafightscorona, coronaoutbreak, donat, swasthabharat, inform, dai, healthforal, fight	-0.507
4	contribut, care, thank, india, fund, covid, support, fight, pmcare, crore	-0.531
10	right, thank, covid, precaut, take, matter, proactiv, lead, effort, overcom	-0.545
8	countri, covid, fight, time, thank, initi, crisi, commun, food, indiafightscorona	-0.721

Table 2. Topic words and coherence scores from LSA with k=12

Topic#	Wordlist	Coherence
12	amp ,young ,ipledgetocontribut ,pmnrf ,donat ,letschaintofightcorona ,appeal ,food ,salari ,distribut	-0.445
2	india ,show ,contribut ,leadership ,rise ,coal ,healthier ,associ ,welfar ,occas	-0.446
5	tax ,celebr ,modiji ,time ,govt ,issu ,ordin ,citizen ,encourag ,rebat	-0.555
11	fight ,amp ,care ,like ,countri ,happi ,defeat ,sportsperson ,hardwork ,battl	-0.61
6	tata ,celebr ,modiji ,care ,group ,servic ,nation ,commit ,indian ,gestur	-0.616
8	inform ,stai ,help ,covid ,safe ,fund ,forefront ,battl ,contribut ,happi	-0.67
3	contribut ,tax ,celebr ,modiji ,pmcare ,time ,citizen ,india ,govt ,effort	-0.747

10	safe ,thank ,help ,indiafightscorona ,forefront ,india ,inform ,contribut ,food ,stai	-0.777
7	amp ,covid ,support ,inform ,dai ,donat ,salari ,pmnrf ,caus ,fund	-0.867
1	contribut ,care ,thank ,covid ,effort ,support ,right ,precaut ,proactiv ,matter	-0.985
9	help ,covid ,care ,amp ,indiafightscorona ,thank ,fund ,swasthabharat ,healthforal ,coronaoutbreak	-1.017
4	covid ,amp ,indiafightscorona ,fight ,help ,modiji ,celebr ,effort ,inform ,contribut	-1.042

Topic 12 talks about the altruistic efforts of people wherein people are donating money and contributing to the Prime Minister's National Relief Fund. Topic 2 looks at the leadership of India and how the measures are taken are for the people's welfare and benefit. Topic 5 and 11 refers to Prime Minister Narendra Modi and his encouragement to the citizens and the hard work of many people in fighting coronavirus. Topic 6 is a reference to Ratan Tata's donation of Rs. 500 crores for this cause. Topic 8 refers to the State-Trait Anxiety Inventory (STAI), people who are at the forefront of this battle against coronavirus and the need for more funds. Topic 3 is a reference to PM Cares fund by Prime Minister Narendra Modi.

In Table 3, we see the comparison of LDA and LSA coherence scores between their most coherent topics. We see that LDA produces greater coherent scores for each topic than LSA demonstrating that LDA performs better than LSA in finding hidden topics.

Table 3. Comparison of coherence scores per topic between LDA and LSA

LDA						
T3	T5	T7	T1	T9	T6	T11
0	0	-0.15	-0.15	-0.22	-0.22	-0.24
LSA						
T12	T2	T5	T11	T6	T8	T3
-0.44	-0.44	-0.55	-0.61	-0.62	-0.67	-0.75

Negative Sentiments. Similar to the process followed for positive sentiments, we ran the LDA model and compute coherence scores to find the optimal number of topics. As can be seen in Figure 4, the optimal number of topics is 8 with a score of 0.542. We then ran the LDA model with K=8 and then computed individual coherence of each topic. We see in Table 4 that topics 3,8,5 and 2 have high coherences while the rest have lower coherence scores.

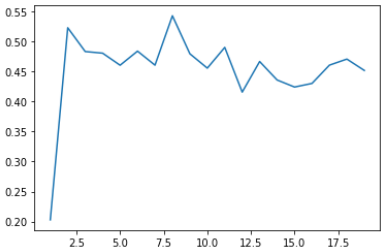

Fig. 4. Coherence chart of LDA model

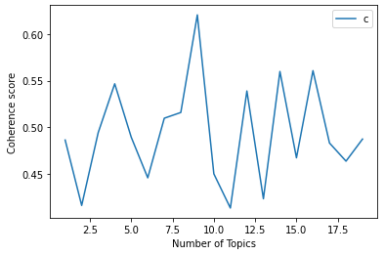

Fig. 5. Coherence chart of LSA model

Topic 3 refers to Tablighi Jamat incident in New Delhi. Topic 8 talks about the lockdown in Kashmir versus lockdown in India due to coronavirus. Topic 5 talks about food availability issues and the impact of workers. Topic 2 talks about the sacrifices made by people during the lockdown period with social life being affected.

Table 4. Topic words and coherence scores from LDA with k=8

Topic#	Wordlist	Coherence
3	lakh ,right ,corona ,countri ,entir ,it' ,tablighi ,fight ,crimin ,unit	-0.049
8	socialdistanc ,stand ,case ,make ,unit ,apart ,kashmir ,terror ,vallei ,struck	-0.2
5	let ,crisi ,face ,covid一 ,stopcoronaviru ,stand ,food ,worker ,amp ,packet	-0.316
2	covid ,distanc ,mean ,social ,life ,sanitis ,dai ,sacrific ,lock ,wrong	-0.358
6	indiafightscorona ,peopl ,covid ,spread ,amp ,stayhomestaysaf ,urg ,awar ,actor ,tamil	-0.469
7	pmcare ,covid ,fight ,fund ,donat ,salari ,month ,indian ,govt ,polic	-0.474
1	covid ,infect ,person ,fight ,support ,produc ,gear ,mumbai ,minim ,step	-0.489
4	amp ,covid ,peopl ,lockdown ,food ,distribut ,coronaviru ,road ,needi ,time	-0.637

In a similar vein, we compute the coherence score for LSA model to determine optimal number of topics. As can be seen from Figure 5, optimal number of topics is 9 with a score of 0.62. We then ran the LSA model with K=9 and then computed individual coherence of each topic. We see in Table 5 that topics 3,8,5 and 4 have high coherences while the rest have lower coherence scores. Topic 3 is a reference to the negative perception regarding lockdown and social distancing, lack of availability of sanitizer and how it is affecting people's daily lives. Topic 8 refers to people not following the protocol and staying at their homes. This will result in increased infections spread through cough droplets. There is mention of Software Technology Parks of India (STPI) which refers to IT offices being shut down and people working from home. Topic 5 is about facing coronavirus by staying at home. In some cities of India, police are asking people on the streets to go inside. Topic 4 refers to deaths due to pandemic and the drastic action of nationwide lockdown.

In Table 6, we see the comparison of LDA and LSA coherence scores. We see that LSA produces greater coherent scores for each topic than LDA. This means that LSA performs better than LDA in finding hidden topics here. Thus, we see that while LDA performed better for positive sentiments, LSA performed better for negative sentiments. This is because the corpus size is different for both sentiments. There are 13701 positive sentiments and 2964 negative sentiments. It is a known fact that for bigger data sets, LDA performs better while for smaller datasets, LSA performs betters [20]. This then explains the performance behaviour difference between positive and negative datasets.

Table 5. Topic words and coherence scores from LSA with k=9

Topic#	Wordlist	Coherence
3	decis ,daily ,distanc ,social ,life ,mean ,wrong ,sanitis ,open ,lockdown	0.00

8	person ,infect ,cough ,viru ,stpiindia ,stayathomesavel ,stai ,awai ,sneez ,speak	0.00
5	let ,stand ,face ,covid,crisi ,stopcoronaviru ,covid ,case ,polic ,stayhom	-0.073
4	lakh ,right ,lockdown ,time ,india ,come ,taken ,need ,death ,pandem	-0.108
9	road ,peopl ,action ,situat ,danger ,control ,kudo ,result ,team ,potenti	-0.195
2	fight ,countri ,corona ,unit ,it' ,crimin ,entir ,tablighi ,sin,talibani	-0.22
7	food ,distribut ,amp ,lockdown ,coronaviru ,poor ,indiafightscorona ,needi ,stayhom ,tamil	-0.263
1	covid ,amp ,peopl ,indiafightscorona ,spread ,stayhomestaysaf ,fight ,urg ,awar ,tamil	-0.613
6	covid ,amp ,peopl ,fight ,support ,gear ,minim ,produc ,naval ,hazard	-0.64

Table 6. Comparison of coherence scores per topic between LDA and LSA

LDA			
T3	T8	T5	T2
-0.05	-0.2	-0.32	-0.36
LSA			
T3	T8	T5	T4
0	0	-0.07	-0.11

4 Conclusion and Future Work

The purpose of this study was to conduct opinion mining on twitter data containing "*#IndiaFightsCorona"* to analyse public opinion. This was accomplished using sentiment analysis and topic modelling. First, sentiment analysis was done, and positive and negative sentiments were separated. Then, on each sentiment, topic modelling was done to discover hidden topics. Two approaches were used, namely LDA and LSA and then their results were compared. It was found that there were more positive sentiments than negative. For positive sentiments, LDA performed better, and for negative sentiments, LSA performed better. LDA revealed topics within positive sentiments which include positive reception of the news of the nationwide lockdown and praising the measures taken by the government, vocal support of lockdown by Bollywood celebrities like Kartik Aryan, STAI, work from home, PM Cares fund and the people who are at the forefront of the battle against the pandemic. Topics revealed by LSA for positive sentiments include the altruistic efforts of people, money donation, the measures are taken are for the people's welfare and benefit, to Ratan Tata's donation of Rs. 500 crores, STAI, PM Cares fund by Prime Minister Narendra Modi and the hard work of many people in fighting coronavirus. As we can see that while some topics are common between LSA and LDA like STAI, PM Cares fund and the people fighting coronavirus, some topics are distinct such as vocal support of lockdown by celebrities like Kartik Aryan, Ratan Tata's donation of Rs. 500 crores and work from home. For negative sentiments, topics revealed by LDA include Tablighi Jamat incident in New Delhi, the lockdown in Kashmir versus lockdown in India due to coronavirus, food availability

issues and its impact of workers and the sacrifices made by people during the lockdown period with social life being affected. Topics revealed for negative sentiments by LSA include a negative perception regarding lockdown and social distancing, lack of availability of sanitizer and how it's affecting people's daily lives, people not following the protocol and leaving their homes, increased infections spread through cough droplets, shut down of IT offices and actions taken by police to make the people follow lockdown protocol. Here, we see very little overlap between the topics generated by LSA and LDA for negative sentiments which include only the negative impact on people's social life. All other topics are distinct. Our findings are echo prior research [1] in which we see that public sentiment is inclined towards positivity.

This study has a few limitations. First, the data was collected only for *#IndiaFightsCorona.* Future work can include hashtags for broader coverage of public opinion. Second, only one evaluation criteria are used, namely topic coherence. Future work can include other criteria such as log-likelihood and perplexity. Also, other methods of topic modelling can also be used apart from LSA and LDA, and a comparison can be made. Third, the data collected is only a sample of the total Twitter data due to API restrictions. A bigger dataset can, perhaps reveal more insights.

References

1. E. Chen, K. Lerman, and E. Ferrara, "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set," *JMIR Public Health Surveill.*, vol. 6, no. 2, p. e19273, May 2020, doi: 10.2196/19273.
2. K. Lerman and R. Ghosh, "Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks," *ArXiv10032664 Phys.*, Mar. 2010, Accessed: Sep. 30, 2020. [Online]. Available: http://arxiv.org/abs/1003.2664.
3. D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *Proceedings of the 20th international conference on World wide web - WWW '11*, Hyderabad, India, 2011, p. 695, doi: 10.1145/1963405.1963503.
4. C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web - WWW '11*, Hyderabad, India, 2011, p. 675, doi: 10.1145/1963405.1963500.
5. E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jul. 2016, doi: 10.1145/2818717.
6. A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study," *J. Med. Internet Res.*, vol. 22, no. 4, p. e19016, Apr. 2020, doi: 10.2196/19016.
7. C. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *PLoS ONE*, vol. 5, no. 11, p. e14118, Nov. 2010, doi: 10.1371/journal.pone.0014118.
8. H. Liang *et al.*, "How did Ebola information spread on twitter: broadcasting or viral spreading?," *BMC Public Health*, vol. 19, no. 1, p. 438, Dec. 2019, doi: 10.1186/s12889-019-6747-8.
9. H. W. Park, S. Park, and M. Chong, "Conversations and Medical News Frames on Twitter: Infodemiological Study on COVID-19 in South Korea," *J. Med. Internet Res.*, vol. 22, no. 5, p. e18897, May 2020, doi: 10.2196/18897.

10. L. Singh *et al.*, “A first look at COVID-19 information and misinformation sharing on Twitter,” *ArXiv200313907 Cs*, Mar. 2020, Accessed: Sep. 30, 2020. [Online]. Available: http://arxiv.org/abs/2003.13907.
11. E. Shearer and K. Matsa, “News Use Across Social Media Platforms 2018,” *Pew Research Center*, Sep. 10, 2018.
12. S. Fischer, “Social media use spikes during pandemic,” *AXIOS*, Apr. 24, 2020.
13. R. Kouzy *et al.*, “Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter,” *Cureus*, Mar. 2020, doi: 10.7759/cureus.7255.
14. M. Cinelli *et al.*, “The COVID-19 Social Media Infodemic,” *ArXiv200305004 Nlin Physicsphysics*, Mar. 2020, Accessed: Sep. 30, 2020. [Online]. Available: http://arxiv.org/abs/2003.05004.
15. C. E. Lopez, M. Vasu, and C. Gallemore, “Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset,” *ArXiv200310359 Cs*, Mar. 2020, Accessed: Sep. 30, 2020. [Online]. Available: http://arxiv.org/abs/2003.10359.
16. B. Liu, “Sentiment Analysis and Opinion Mining,” *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
17. A. Bermingham and A. F. Smeaton, “Classifying sentiment in microblogs: is brevity an advantage?,” in *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM ’10*, Toronto, ON, Canada, 2010, p. 1833, doi: 10.1145/1871437.1871741.
18. A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.,” *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pp. 1320–1326, 2010.
19. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau, “Sentiment analysis of twitter data.,” *Proc. Workshop Lang. Soc. Media*, pp. 30–38, 2011.
20. P. Kherwa and P. Bansal, “Topic Modeling: A Comprehensive Review,” *ICST Trans. Scalable Inf. Syst.*, vol. 0, no. 0, p. 159623, Jul. 2018, doi: 10.4108/eai.13-7-2018.159623.
21. D. M. Blei, A. Y. Ng, and M. I. Jordon, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
22. A. Onan, K. Serdar, and B. Hasan, “LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis,” *Int. J. Comput. Linguist. Appl.*, vol. 7, no. 1, pp. 101–119, 2016.
23. K. Farrahi and D. Gatica-Perez, “Discovering routines from large-scale human locations using probabilistic topic models,” *ACM Trans. Intell. Syst. Technol. TIST*, vol. 2, no. 1, p. 3, 2011.
24. K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, “Exploring Topic Coherence over many models and many topics,” *Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Learn.*, pp. 952–961, Jul. 2012.
25. Moh. N. Aziz, A. Firmanto, A. M. Fajrin, and R. V. Hari Ginardi, “Sentiment Analysis and Topic Modelling for Identification of Government Service Satisfaction,” in *2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, Sep. 2018, pp. 125–130, doi: 10.1109/ICITACEE.2018.8576974.
26. J. H. Lau, D. Newman, and T. Baldwin, “Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality,” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, Apr. 2014, pp. 530–539,.