

Root cause analysis of COVID-19 cases by enhanced text mining process

Sujatha Arun Kokatnoor, Balachandran Krishnan

Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bengaluru, India

Article Info

Article history:

Received Mar 17, 2021

Revised Aug 3, 2021

Accepted Sep 1, 2021

Keywords:

Dunn index

Feature engineering

Feature hashing

Hierarchical dirichlet process

K-means

Latent dirichlet allocation

Latent semantic analysis

ABSTRACT

The main focus of this research is to find the reasons behind the fresh cases of COVID-19 from the public's perception for data specific to India. The analysis is done using machine learning approaches and validating the inferences with medical professionals. The data processing and analysis is accomplished in three steps. First, the dimensionality of the vector space model (VSM) is reduced with improvised feature engineering (FE) process by using a weighted term frequency-inverse document frequency (TF-IDF) and forward scan trigrams (FST) followed by removal of weak features using feature hashing technique. In the second step, an enhanced K-means clustering algorithm is used for grouping, based on the public posts from Twitter®. In the last step, latent dirichlet allocation (LDA) is applied for discovering the trigram topics relevant to the reasons behind the increase of fresh COVID-19 cases. The enhanced K-means clustering improved Dunn index value by 18.11% when compared with the traditional K-means method. By incorporating improvised two-step FE process, LDA model improved by 14% in terms of coherence score and by 19% and 15% when compared with latent semantic analysis (LSA) and hierarchical dirichlet process (HDP) respectively thereby resulting in 14 root causes for spike in the disease.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sujatha Arun Kokatnoor

Department of Computer Science and Engineering, School of Engineering and Technology, CHRIST

(Deemed to be University)

Bengaluru, India

Email: sujatha.ak@christuniversity.in

1. INTRODUCTION

The coronavirus is a family of viruses capable of causing a variety of diseases that are life threatening to humans, including common and more severe forms of cold. The signs and symptoms of the disease may occur within two to 14 days after exposure. This time referred to as the incubation period is the time after exposure and before symptoms. The general signs and symptoms include fever, cough, tiredness, breathing difficulty, sore throat, running nose, headache, and chest pain. Other less common signs also include rash, nausea, vomiting, and diarrhea. Some people may only have a few symptoms and some may not have any symptoms at all. These cases are referred to as cases, symptomatic and asymptomatic respectively [1], [2].

As per the World Health Organization (WHO), data have shown that the virus spreads from person to person (about 6 feet or 2 meters) among the people in close contact. The virus spreads through respiratory droplets when someone is coughing, sneezing, or talking. Such droplets may be inhaled or landed in a nearby person's mouth or nose. It can also spread when a person touches a surface and touches his or her mouth, nose, or eyes, but this is not a major way of spreading the virus as per WHO reports [1]. In the case of

symptoms (symptomatic), a person with the virus is the most infectious, and this is the time that they are most likely to transmit the virus, according to the center for disease control and prevention (CDC) [2] trusted source. But even before they start showing symptoms (asymptomatic) of the disease itself, someone can spread it.

As per WHO for India region, there was a spike in new COVID-19 cases from 1907 to 7025 according to the data collected between 18th June to 29th June 2020 and to 22075 new cases as on 16th July 2020 as shown in Figure 1. There were 2003 deaths observed as on 16th June 2020. To understand the reason behind the increase in new COVID-19 cases during those days, apart from the information available in WHO [1], CDC [2], mohfw.gov.in [3] and mygov.in [4], data was collected from the online social media (OSM) as well to understand and analyze public's opinion on the same. Twitter was chosen for the people's opinions as it is one of the popular OSM as per Data Never Sleeps 7.0 report, where people post more than 500000 posts per minute.

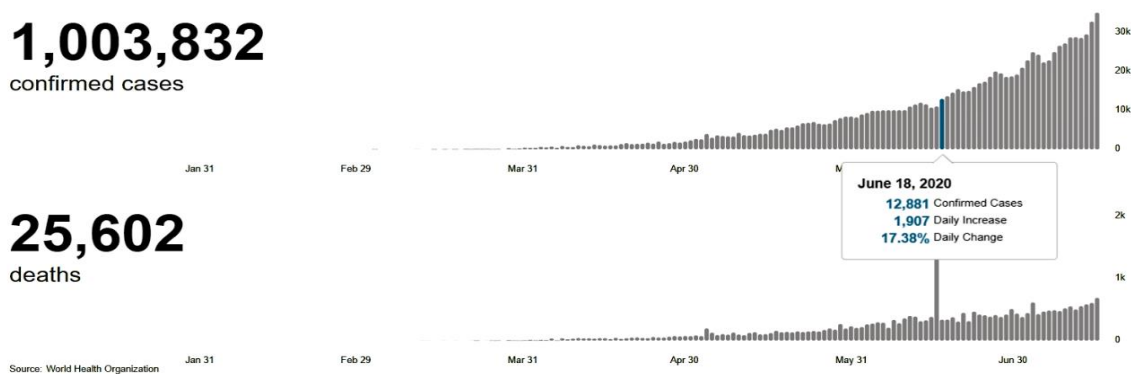


Figure 1. COVID-19 statistics as per World Health Organization [1]

The collected tweets were pre-processed and improvised feature engineering was applied to create an efficient vector space model (VSM) [5]. The tweets were then clustered using the proposed enhanced K-means clustering algorithm to group the dataset into five different clusters. It was observed manually that all the user's discussion on how the disease was spread was found in the third cluster and accordingly the third cluster was considered for topic modeling. Topic modeling was done using the latent dirichlet allocation (LDA) model [6] which identified 41 root causes through the help of trigrams. These 41 causes were communicated to six medical professionals across Karnataka, India and thus was validated to 14 important and prioritized root causes.

There are many clustering algorithms including partitioning, hierarchical, spectral, density based and grid based [7]–[9]. K-means is one of the partitioning clustering algorithms. A regular K-means is sensitive to outliers and to the contents present in it. The center or the mean of the data points belonging to the clusters is represented by each cluster. The text data posted by users in OSM for a particular issue or concern, when converted into a feature vector, the majority of the data points tend to be near first centroid due to the similar contents. This indicates a drawback for the K-means clustering algorithm. For the majority of tuples, the sum of squared error (SSE) becomes nil for the first center and only a smaller number of tuples have high values of SSE which put these into other clusters. This is mainly because the text dataset lacks various topics [10]. Hierarchical clustering has ambiguity in the termination process, encompasses a lot of illogical decisions and works better with numerical datasets [11]. Density based clustering is sensitive in setting the input parameters and has poor cluster descriptors [12]. In grid-based clustering, the boundaries of the clusters are either vertical or horizontal and there is no detection of diagonal boundaries. Also, they are computationally expensive for the datasets having distinct data points [13]. In this research work, an enhanced K-means clustering was implemented to address the grouping of the input dataset into five clusters based on the contents present in the user's tweets.

Feature engineering was applied in two steps. In the first step, forward scan trigrams (FST)-based term frequency and inverse document frequency (TF-IDF) was applied to reduce the high dimensional feature vector into an efficient VSM [5]. In the second step, all the weak features present in the text dataset were removed using the proposed feature hashing method. This VSM was input to the LDA topic model [14] for identifying the most relevant trigram topics which were considered as root causes for the spike in new COVID-19 cases for the dataset considered.

2. RELATED WORK

For anticipation of the disease epidemiological trend and rate of COVID-19 in India, linear regression (LR), multilayer perceptron (MLP) and the vector auto regression (VAR) models were used [15]. The possible COVID-19 impact trends in India were predicted on the basis of data collected from Kaggle. The prediction model was based on the cases which were in primitive stages and the Spearman's correlation was used to find the similarity between the features present in the dataset [15]. As the dataset considered is non-linear, and dependent on each other, Spearman's Rank coefficient [16] has led to inaccurate forecasting of the spread of the disease.

Several technologies including blockchain technology, internet of things (IoT), artificial intelligence (AI), machine learning (ML), 5G and unmanned aerial vehicles (UAF) were used to reduce the impact of corona virus disease outbreak by analyzing the datasets available [17]. Although studies on the pathophysiological properties of COVID-19 exist, it remains somewhat elusive about its spreading mechanism. Machine learning was used to quantify COVID-19 contents of establishment of health guidance, especially vaccines amongst online opponents in [18]. User's posts on Facebook were analyzed for both anti vaccination and pro-vaccination communities. Snowball's approach was used for scraping user's posts which discussed either vaccines or policies about vaccination or an argument on pro and anti-vaccination for the COVID-19 disease. Later LDA algorithm was used for analyzing the appearance and involvement of topics on COVID-19 [18]. The limitations included the study of other social media data and the feature engineering process in dealing with the text dataset.

In [19] situational information from social media data on COVID-19 was identified, analyzed, and classified using natural language processing techniques into seven types of situational information. They were cautions and advice, measures taken, donations, emotional support, seeking help, criticizing and rumor spreading. The dataset was manually labeled, and later SVM, naïve Bayes (NB), and random forest (RF) algorithms [19] were used for the classification. The limitations of this technique are that the social media data doesn't come with a label and manual labeling is very time consuming and is limited to one's domain expertise.

To analyze multivariate time series evolution, a cluster-based method named Hierarchical clustering was used for the COVID-19 pandemic in [20]. Countries were divided into clusters on a daily basis, according to their cases and death numbers. Algorithmically, the total number of clusters and the membership of individual countries were determined. This analysis gave new insights into COVID-19's spread across countries and through time [20]. Hierarchical clustering seldom provides the best solution, as it involves a lot of arbitrary choices, which include the fact that it does not work with missing data, works poorly with mixed data types, is doesn't work well on huge data sets, and is commonly misinterpreted with its main output, the dendrogram.

In the conventional K-means clustering process [21], significant improvements in defects of poor results of the clustering were found due to the optimum local value and large intra-cluster variance when calculating the density of a data set by means of a weighted distance density calculation method. Experimental results showed that the intra-cluster variance of the clustering results was reduced in comparison with the traditional method by applying the enhanced method which improved the performance of the algorithm [22]. A weighted distance density method fails to identify the outliers present in the dataset.

In the data set from different regions of China, obtained from the WHO [1], the K-means clustering based machine learning method was used. Within the original WHO data set the temperature area was included to demonstrate the effect of temperature on each region within three separate COVID-19 perspectives—suspected, verified, and death [23]. It is observed that temperature is not the only factor for the spread of the corona disease. There are several other factors for the spreading and if these are included as attributes for the data analysis, a better model of avoidance can emerge.

Latent dirichlet allocation [24] was used in the grouping of similar tweets which occurred in the same user to user communication channel in [6]. Cosine Similarity was used for extracting the topmost ten tweets in this technique. The grouping done by considering hashtags caused duplication of the tweets and thus took a longer training time thereby reducing the performance of the model.

The important topics posted by the public in Twitter were identified using the online LDA in [25]. A total of twelve different topics were identified which were consolidated into four main categories: i) virus origin, ii) virus resources, iii) virus impact factor on the public, and iv) countries and the economy and the last category was the identification of ways of mitigating the risk of infection. The regular online LDA uncorrelated topics could not be captured due to the topic's distribution in the tweets collected. The number of topics in the dataset was specified by the authors which is subjective and doesn't always highlight the true distribution of topics.

In [26], the public messages from Reddit based on coronavirus disease discussion were first extracted. Then, fifty discussion topics were chosen and the data analysis was done using natural language processing (NLP) and LDA algorithms [14]. Out of the chosen topics, three topics were found to be more

relevant to the corona virus discussion including public health measures, daily life impact and sense of pandemic severity. The topics were predicted according to the multinomial distribution, followed by a further multinomial distribution trained on a particular topic. However, this may be unsuitable if the real structure is more complex than a multinomial distribution or if the data to be trained is not sufficient [27].

3. RESEARCH METHOD

The main objective of this research work is to identify the root causes for spike in new coronavirus cases through identification of mode of virus transmission from one person to another using machine learning approaches. People have expressed their views and opinions on the onset of the coronavirus in the OSM. When the spike was observed in an increase of fresh cases, people took to OSM for expressing their views in their domain knowledge and the reasons behind its increased cases. Those posts were extracted, the reasons were analyzed and validated with help of medical professionals. One among the most popular OSMs is Twitter from where tweets were extracted for the study. The extracted text has imprecise grammar, leading to lexical, syntactic, and semantic ambiguities. This causes inappropriate analysis and identification of patterns. The text therefore needs to be pre-processed before it can be used for clustering and followed by topic modeling [6]. Figure 2 shows the architectural model considered for the study. After extracting the tweets from Twitter, natural language tool kit (NLTK) 3.1 version is used for initial data preprocessing. Then the first level of improvised feature engineering (weighted TF-IDF in combination with Forward Scan Trigrams [5]) is applied to create an efficient VSM. This VSM is input to an enhanced K-means clustering algorithm to yield clusters based on the similarity of the data elements. The cluster consisting of relevant root causes addressing the coronavirus further undergoes a second level of improvised feature engineering including removal of weak features by feature hashing method. The outcome of this step is input to the LDA Topic Model to identify topics and its associated words on the possible root causes identified.

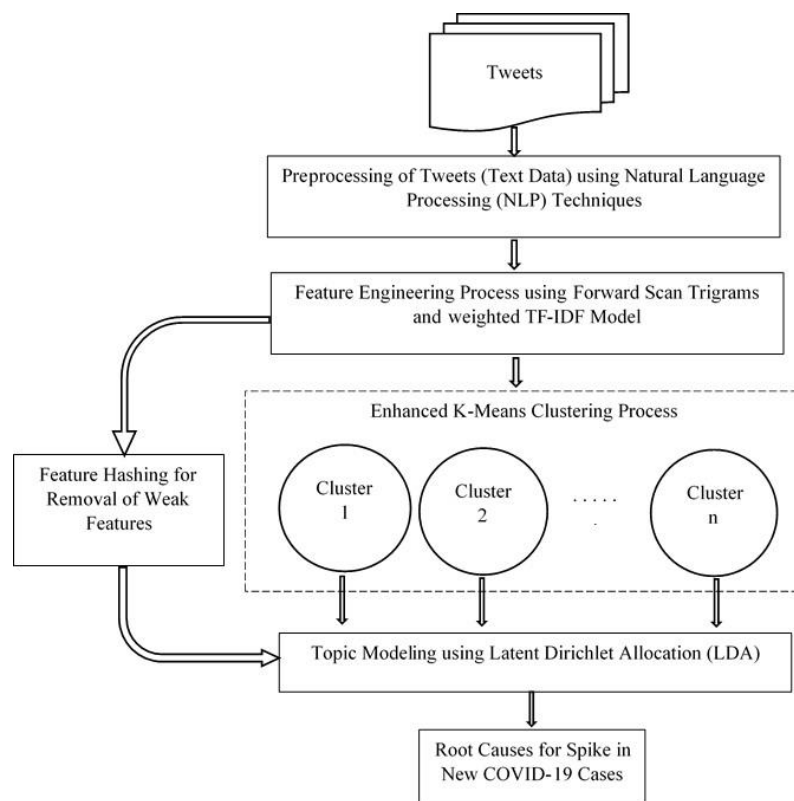


Figure 2. Proposed architecture for identification of root causes for spike in new COVID-19 cases

3.1. Corpus creation

As per data never sleeps 7.0 report [28], Twitter is one of the most popular OSMs where more than 500000 messages are posted by users per minute. Twint [10] comes with search filters which can be

combined with useful methods to display precise data. Twint is a Python library that can be written to carry out more specific and complex actions. The ability to write and scale Twitter 270 searches with Twint enables easy but efficient data extraction from social media. Twint has the opportunity to collect information about Twitter events of recent times.

COVID-19 statistics was taken from the WHO official website [1]. It was observed that there was a sudden spike in new cases as shown in Figure 1 between 18th June 2020 to 29th June 2020, when the Government announced Unlock 1.0 with ease of public movement restrictions. Using Twint, 7674 tweets were collected during the mentioned duration for India region using keywords such as “spike in COVID-19”, “spike in coronavirus”, “increase in COVID-19”, “increase in new coronavirus cases”, “root cause for corona spread” and “recent spike in new corona cases”.

3.2. Data pre-processing

This sub-section explains the pre-processing of text data used in the research work. Twitter comments were pre-processed before analysis in order to achieve better results. The proposed method utilized NLP tools. For the creation of a standard text data set, the following process was used. The textual tweets format was converted to a lower case in order to reduce text dataset volume. Punctuation has been removed to avoid different forms of the same word and white spaces too were removed. The data set stop words with little contribution to the calculation of the semantic quality of the document were deleted. These stop words included “is”, “the”, “and”, “a,” “an”. The original words with similar semantic characteristics but different forms (also called stemming words) were reduced to a common root term. For example, causes and causing, were reduced to its root term “cause”.

3.3. Feature engineering

Feature engineering [29], [30] is the domain knowledge approach used to extract raw data features using data mining techniques. These features can be used to enhance machine learning algorithms. In this proposed work, the tweets extracted from users on how there was an increase in new coronavirus cases between 18th June to 29th June 2020 for India specific, were converted into an efficient VSM using two steps of feature engineering process [5]. Weighted TF-IDF with forward scan trigrams approach was used in the first step [5] and weak features were removed using improvised feature hashing in the second step.

3.3.1. Improvised feature hashing

The second step of the feature engineering process was applied to the processed dataset. The output of it was fed to the LDA topic model. This process helped in analyzing and using the relationship between the topics and the words associated with it. These topics contained only those words relevant to the reasons behind the spike in new COVID-19 cases. The second step involved removal of weak features. In a text dataset, words with higher frequencies appear more often than words with low frequencies. The words with low frequency are essentially poor corpus characteristics, making it a good practice in removing them. All the weak features were removed from the preprocessed corpus in the second stage. This was accomplished by feature hashing technique where the terms were mapped with Hash function indices. The standard default feature size of 262,144 with Murmur3 hash function [31] was used in the study. Algorithm 1 illustrates the steps to remove weak features using improvised feature hashing. The VSM generated after removing all the weak features, was input to the LDA topic modeling algorithm [6].

Algorithm 1. Weak features removal using improvised feature hashing

```
//Input: Pre-Processed Text Corpus (Trigrams)
//Output: Corpus without Low Frequency Terms
1: for i in trigrams:
    value = murmur3(i)
    Calculate frequencies from the mapped indexes
2. Arrange every trigram according to its frequency calculated from Step 1
3. label(x)=1 [x is a trigram]
4. Calculate Median=(N+1)/2 [N is the total number of trigrams generated]
5. If (frequencycount(x) < Median)
    label(x)=0
6. Remove x with label value=0
7: Return
```

3.4. Enhanced k-means clustering

K-means is an unsupervised machine learning clustering algorithm [8]. Its main task is to divide the input dataset into K clusters based on the similarity of the data elements present in the dataset. It is unsupervised in nature. The K-means clustering algorithm indicates that the clusters are entirely dependent on the selection of the original clusters. K data elements as the initial centroids are selected and Euclidean

distance measure calculates the distances of all the data elements with the chosen centroids. Data elements less far from centroids are moved to the corresponding cluster. The process continues until no further changes in the centroids take place. The conventional K-means algorithm [8], however, is computationally expensive since centroids provide results in the quality of the clusters and converge to minimum local levels. Conventional K-means is also sensitive to original seed value (likely to be $k!/k^k$ where k is the total number of clusters chosen) as well as outliers in the dataset.

The initial centroids are systematically determined to produce the clusters more accurately during the first phase, while in the second phase, data points are assigned to the appropriate clusters described in the proposed algorithm. These two phases are iterated for $k = \sqrt{\frac{n}{2}}$ times, where n is the total number of feature vectors. In each iteration the Dunn index (DI) [32] is calculated using the equation 1 and the details of clusters and the DI values are stored in temporary lists. The total number of clusters and their details are considered for further analysis based on the highest DI value.

$$DI = \frac{\text{MinValue (Inter Cluster Distance)}}{\text{MaxValue (Intra Cluster Distance)}} \quad (1)$$

Algorithm 2 illustrates the procedure of enhanced k-means clustering algorithm. *calculate_centroids* and *determine_clusters* are shown in Figure 3 and Figure 4, respectively.

Algorithm 1. Enhanced k-means clustering algorithm

```
//Input: Pre-Processed Text Corpus
//Output: K clusters
1:  $K = \sqrt{n/2}$ 
2: Initial Cluster  $C = \{\text{Random Data Element}\}$ 
3: for  $i = 1$  to  $K$ :
 $C1 = \text{Calculate\_Centroids}(C)$ 
 $C2 = \text{Determine\_Clusters}(C1)$ 
Calculate DI and store DI value and the corresponding clusters details in a temporary list
4: for  $i = 1$  to  $K$ :
 $M = \text{Choose Maximum DI}$ 
5: Return the Clusters and the corresponding data elements for the  $M$ 
```

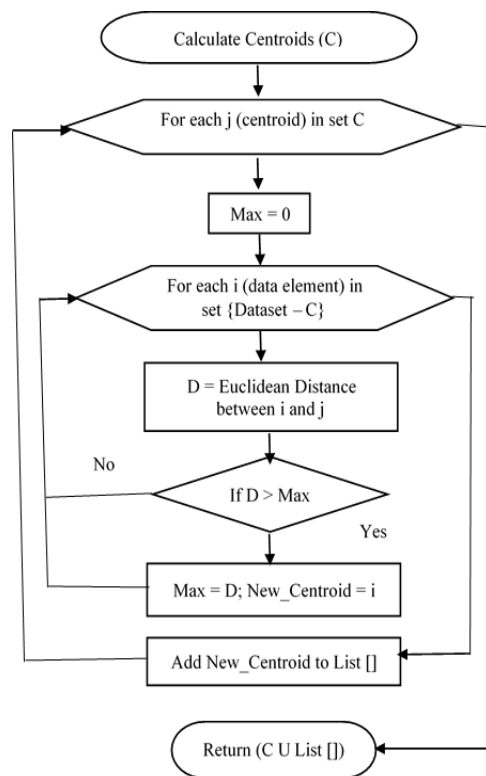
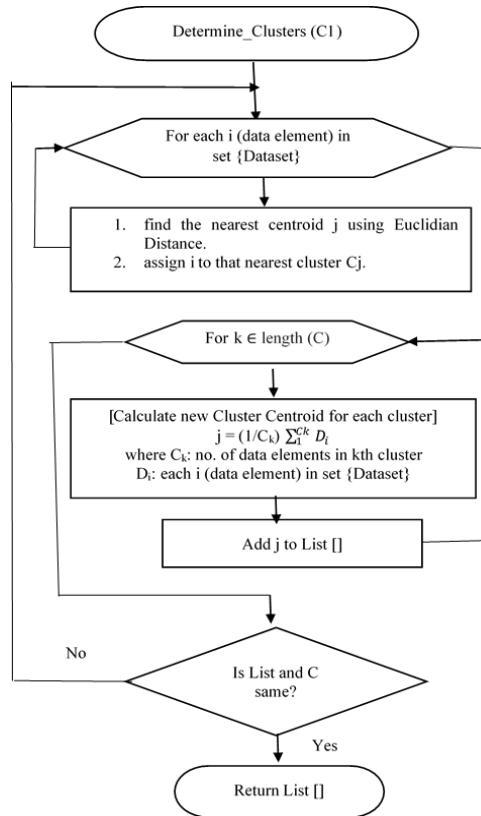


Figure 3. Flowchart for calculate_centroids

Figure 4. Flowchart to *determine_clusters*

3.5. Latent dirichlet allocation algorithm (LDA)

The VSM obtained after the feature engineering process as described in the Section 4.3.1 was fed as input to LDA topic modeling model [6] for further analysis. LDA is an unsupervised machine learning technique generally used for analyzing text documents for determining cluster words for a particular topic. In this paper, LDA topic modeling has been used to determine the reasons behind the increase in new COVID-19 cases between certain periods of time as per WHO reports based on the public discussions in OSM.

4. RESULTS AND DISCUSSION

This section provides a performance assessment of the research methodology. The data for the study included Indian public discussions on the day-to-day increase in COVID-19 cases on Twitter. The tweets were collected between 18th June and 29th June 2020, using a Python library named Twint as explained in section 3.1. The tweets were further processed using NLP tools and the feature engineering process as discussed in sections 3.2 and 3.3.

NLTK 3.1 [33] was used for building python programs as the work involved human generated textual dataset. After pre-processing the dataset, the input was fed to enhanced K-means clustering algorithm as discussed in section 3.4. Initially, regular K-means clustering [8] was applied with K=5 clusters. It was observed that there was overlapping of the clusters and the data points were not distributed accurately. Although K-means was iterated for 200 times, no changes were observed in the centroids. The pre-processed dataset was input to the enhanced K-means algorithm. It was observed that the cluster centroids were well separated in enhanced K-means clustering when compared to the conventional one.

DI [32] was used to validate and identify sets of clusters which were compact, with small variations between cluster members and with sufficient distance between other clusters centroids. The optimal number of clusters K=5 was chosen for this research work with maximum Dunn Index value (0.2797) as shown in Figure 5. The DI for conventional K-means with K=5 resulted in 0.0986. It was observed from clustering results done through enhanced K-means clustering algorithm that only one cluster among 5 clusters contained the reasons for increase in new cases as posted by the public. The cluster comprising root causes was validated through a statistical measure namely Cohen's Kappa measure of agreement [34]. Out of

7,674 tweets collected through manual annotation, it was found that 1639 tweets consisted of reasons for spike in the disease. But through the experimentation process, it was observed that 1,335 data elements were clustered consisting of the root causes in the 3rd cluster. As the value of C was 0.92, there was a strong agreement between the manual identification and the experimental results.

The third cluster after removal of the weak features through Feature Hashing Process as discussed in section 3.3.1, was input to the LDA model. Number of topics chosen were 10 and the experimentation process resulted in 300 trigrams with 30 trigrams per topic. The following 41 reasons as listed in Table 1, were found to be the root causes for a sudden spike in new COVID-19 cases between the specified time duration as per public’s opinion and discussion in OSM. These reasons were communicated to six medical professionals for the validation process where they consolidated the reasons to 14 main root causes which are shown in Table 2.

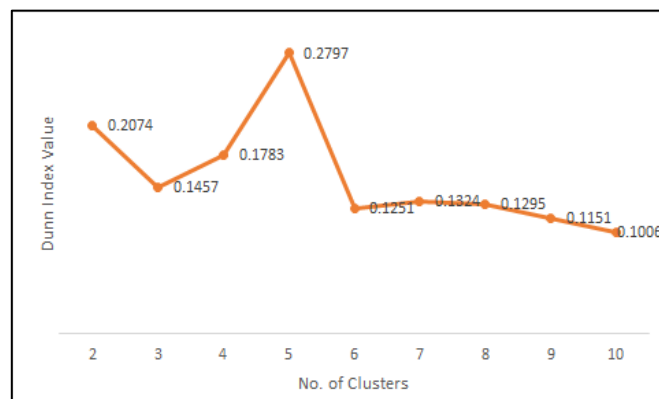


Figure 5. Validation of clusters with Dunn index

Table 1. LDA consolidated results 41 root causes for COVID-19 increase

Identified root causes for a sudden spike in New COVID-19 cases		
1. More family gathering	15. People moving freely	29. Late visiting to hospital
2. No proper social distancing	16. Do not Care other People	30. Early intervention and self-medication
3. Low quality sanitizers	17. Hospital was seized	31. Social workers lack
4. Poor quality masks	18. Hospitals are closed	32. Blame game attacking
5. Protests are leading spikes	19. Police and health	33. to work together
6. Hospital lock down	20. People are scared	34. Less people coordination
7. Late visiting Hospital	21. Avoid hospital checkup	35. Mislead other causes spikes
8. Less people coordination	22. Social stigma result	36. Permanent Doctors Nurses
9. People mislead others	23. Scared social stigma	37. Lack of doctors nurses
10. Do not follow home quarantine	24. Avoiding barricade during covid	38. Democrats keep yelling
11. Do not wear mask while outside	25. High influential people	39. Lying about corona
12. Infected-late presentation	26. Put proper mask	40. Hospital Lock Down
13. Spread across community	27. How to wear mask	41. Panic other group
14. Regular working people	28. Remove mask while outside	

Table 2. Consolidated root causes validated by medical professionals

Consolidated Root Causes for New COVID-19 Cases as Validated by Medical Professionals
1. Family gatherings-**
2. No proper social distance-*****
3. Low quality sanitizers- *
4. Low quality masks-*
5. Not wearing masks appropriately when in public-****
6. Removing masks while speaking-***
7. Hospital lockdowns due to a covid case thereby preventing treatment for rest identified-*
8. Group protests, people are moving freely, do not care about other people, avoiding home quarantine-***
9. Self-intervention, hiding the illness, social stigma, using highly influential people for hiding the disease, late visit to the hospitals, lying about their illness, avoiding barricades.-***
10. Lack of social workers, nurses, doctors and health facility-*
11. Misleading and panicking other groups of people-**
12. Lack of people coordination while treating people with Covid-*
13. Infections from police and health care sectors-*
14. No regular working people like nurses doctors for serving people who have corona-*

Each * provided by the medical professionals specified an extra weightage for the consolidated point mentioned, where * has least weightage and ***** has highest weightage, respectively. Feature Engineering through TF-IDF+forward scan trigrams [5] and removal of weak features through Feature Hashing has helped improve the model's performance by 12% in terms of coherence scores. The coherence score was used in the experimentation process for assessing the quality of the identified topics. It outputs a value for each topic by calculating the degree of similar semantic words for the chosen topic. The comparison results are shown in Table 3. Figure 6 gives the comparative analysis of LDA with latent semantic analysis (LSA) and hierarchical dirichlet process (HDP) topic models. LSA partially captured polysemy with several symbolic significance. It worked especially successfully for long documents because, based on their contextual meaning, a small number of vectors represented each document. Due to the high volume of data which lowered the efficiency of LSA, however, more storage and calculation times were consumed. The findings of the LSA were also found difficult to understand terms related to topics. The HDP calculated a simple probability and required various multinomial distributions, leading to an infinite number of topics. This allowed the number of samples taken into account to be increased or decreased. HDP restricted its option to an optimum degree of granularity for the number of topics, because of which it performed less. LDA included documents relevant to a certain number of topics and that each document derived from a combination of probabilistic samples. First the distribution of possible topics for the dataset was considered and second the list of potential words in a chosen topic was considered. Also, LDA easily distinguished between global and local variables. The global variables are topics and local variables are document level indices.

Table 3. Comparative coherence values for LDA model with feature engineering process

No. of Topics	Feature Engineering 1st Step: TF-IDF+Forward Scan Trigrams	Feature Engineering 1st Step+Feature Hashing
10	0.5302	0.6158

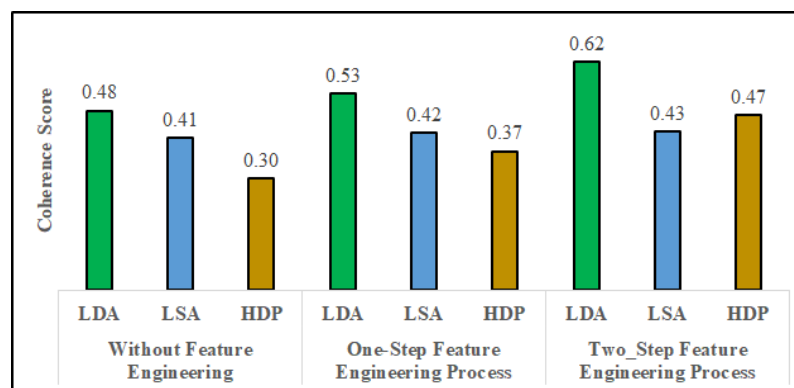


Figure 6. Comparative analysis of LDA with LSA and HDP topic models with and without feature engineering process

5. CONCLUSION

The analysis and description of the transmission of coronavirus disease by considering public posts in Twitter and its validation with the aid of medical professionals was a major focus of this article. The overview description provided a better understanding of COVID-19 and its exponential growth in the recent days. The Twitter® posts analysis on the disease provided relevant and useful topics in three-word sequences which specified the probable root causes of its transmission from one person to another.

In this proposed work, a three-step model was proposed. In the first step, the text corpus extracted between 18th June and 29th June 2020 consisting of COVID-19 discussions and their propagation were preprocessed using NLP tools and feature engineering process. The calculation of trigrams (forward scan trigrams) was modified by considering the probabilities of not only the words in backward direction for a target word for a given text but also the probabilities of words in the forward direction as well. This modification helped in retaining the context and meaning of the sequence considered as the user's text is a natural language text written naturally without considering the grammar aspects. This modified information was passed on to the weighted TF-IDF model thereby reducing the dimensionality of the VSM and was further processed by an additional task of removing weak features based on a process combining feature

hashing and a median value. This improved the detection of hidden topics with the simultaneous increase of the coherence Scores of the model. The general variance model added more weight for the outliers because they were far from the calculated mean, and median value was therefore taken into consideration for the removal of weaker features. This step enhanced the proposed model's performance by 12% in terms of coherence Scores. In the second step, an enhanced K-means clustering algorithm was applied for calculation of the centroids thereby grouping the processed dataset into clusters. The cluster validation was done through DI and five clusters were chosen based on the highest DI value obtained (0.2797). In the last step, LDA Topic Modeling was used for identifying all the root causes relevant for increase in new infected cases of coronavirus. Out of five clusters resulted from the second step, only one cluster attributed to the root causes and the same was validated through Cohen's kappa coefficient which resulted in 92% indicating a strong agreement between the manual identification and the experimental results. The last step of the proposed model yielded 43 root causes out of which 41 causes were validated by five medical professionals across Karnataka. The other two reasons ('recent travel history' and 'direct contact infected') were validated with WHO reports. The 41 reasons considered for the spread in disease were further consolidated to 14 prioritized root causes by the medical professionals. The future scope may include automated identification of root causes not only from the textual posts, but also from the emoticons, images and videos posted by the public in OSM.

ACKNOWLEDGEMENTS

Authors wish to acknowledge the technical and infrastructural help rendered by the faculty members of CSE department of CHRIST (Deemed to be University), Bangalore, India. The authors would also like to thank the Medical Practitioners across Karnataka, India, for providing a detailed discussion and report on what COVID-19 is, its possibility of transmission from one to another living being and the necessary precautions to be taken.




REFERENCES

- [1] "Coronavirus disease (COVID-19) pandemic," *World Health Organization*. <https://www.who.int> (accessed Mar. 2020).
- [2] "Coronavirus (COVID-19)," *Centers for Disease Control and Prevention*. <https://www.cdc.gov/> (accessed Jul. 2020).
- [3] "Coronavirus (COVID-19)," *Ministry of Health and Family Welfare*. <https://www.mohfw.gov.in/> (accessed Jul. 2020).
- [4] "Coronavirus (COVID-19)," *Government Online Services*. <https://www.mygov.in/> (accessed March 2020).
- [5] S. A. Kokatnoor and B. Krishnan, "A two-stepped feature engineering process for topic modeling using batchwise LDA with stochastic variational inference model," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 4, pp. 333–345, Aug. 2020, doi: 10.22266/IJIES2020.0831.29.
- [6] D. Alvarez-Melis and M. Saveski, "Topic modeling in Twitter: aggregating tweets by conversations," in *ICWSM*, 2016, vol. 20, no. 1.
- [7] S. Kodati, R. Vivekanandam, and G. Ravi, "Comparative analysis of clustering algorithms with heart disease datasets using data mining weka tool," in *Advances in Intelligent Systems and Computing*, vol. 900, Springer Singapore, 2019, pp. 111–117.
- [8] M. Z. Rodriguez *et al.*, "Clustering algorithms: a comparative approach," *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. 0210236, doi: 10.1371/journal.pone.0210236.
- [9] J. Khalfallah and J. B. H. Slama, "A comparative study of the various clustering algorithms in e-learning systems using weka tools," in *Proceedings of 2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing, JCCO: TICET-ICCA-GECO 2018*, Nov. 2018, pp. 76–82, doi: 10.1109/ICCA-TICET.2018.8726188.
- [10] S. A. Kokatnoor and B. Krishnan, "Self-supervised learning based anomaly detection in online social media," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 3, pp. 446–456, Jun. 2020, doi: 10.22266/IJIES2020.0630.40.
- [11] F. Ros and S. Guillaume, "A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise," *Expert Systems with Applications*, vol. 128, pp. 96–108, Aug. 2019, doi: 10.1016/j.eswa.2019.03.031.
- [12] A. Sharma, R. K. Gupta, and A. Tiwari, "Improved density based spatial clustering of applications of noise clustering algorithm for knowledge discovery in spatial data," *Mathematical Problems in Engineering*, vol. 2016, pp. 1–9, 2016, doi: 10.1155/2016/1564516.
- [13] B. Wu and B. M. Wilamowski, "A fast density and grid based clustering method for data with arbitrary shapes and noise," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1620–1628, Aug. 2017, doi: 10.1109/TII.2016.2628747.
- [14] W. Anwar, I. S. Bajwa, M. A. Choudhary, and S. Ramzan, "An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution," *IEEE Access*, vol. 7, pp. 3224–3234, 2019, doi: 10.1109/ACCESS.2018.2885011.
- [15] R. Sujath, J. M. Chatterjee, and A. E. Hassanien, "A machine learning forecasting model for COVID-19 pandemic in India," *Stochastic Environmental Research and Risk Assessment*, vol. 34, no. 7, pp. 959–972, May 2020, doi: 10.1007/s00477-020-01827-8.
- [16] P. Radhakrishnan and B. Vignesh, "A note on rank correlation and semi-supervised machine learning based measure," in *2017 Innovations in Power and Advanced Computing Technologies, i-PACT 2017*, Apr. 2017, vol. 2017-January, pp. 1–8, doi: 10.1109/IPACT.2017.8245035.
- [17] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact," *IEEE Access*, vol. 8, pp. 90225–90265, 2020, doi: 10.1109/ACCESS.2020.2992341.
- [18] R. F. Sear *et al.*, "Quantifying COVID-19 content in the online health opinion war using machine learning," *IEEE Access*, vol. 8, pp. 91886–91893, 2020, doi: 10.1109/ACCESS.2020.2993967.




- [19] L. Li *et al.*, "Characterizing the propagation of situational information in social media during COVID-19 epidemic: a case study on Weibo," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 556–562, Apr. 2020, doi: 10.1109/TCSS.2020.2980007.
- [20] N. James and M. Menzies, "Cluster-based dual evolution for multivariate time series: analyzing COVID-19," *Chaos*, vol. 30, no. 6, Jun. 2020, Art. no. 61108, doi: 10.1063/5.0013156.
- [21] Y. Xu, X. Fu, H. Li, G. Dong, and Q. Wang, "A k-means algorithm based on feature weighting," *MATEC Web of Conferences*, vol. 232, 2018, Art. no. 3005doi: 10.1051/mateconf/201823203005.
- [22] W. Yang, H. Long, L. Ma, and H. Sun, "Research on clustering method based on weighted distance density and k-means," *Procedia Computer Science*, vol. 166, pp. 507–511, 2020, doi: 10.1016/j.procs.2020.02.056.
- [23] M. K. Siddiqui *et al.*, "Correlation between temperature and COVID-19 (suspected, confirmed and death) cases based on machine learning analysis," *Journal of Pure and Applied Microbiology*, vol. 14, no. suppl 1, pp. 1017–1024, Apr. 2020, doi: 10.22207/JPAM.14.SPL1.40.
- [24] L. Zou and W. W. Song, "LDA-TM: a two-step approach to Twitter topic data clustering," in *Proceedings of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2016*, Jul. 2016, pp. 342–347, doi: 10.1109/ICCCBDA.2016.7529581.
- [25] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hai, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: a surveillance study," *Journal of Medical Internet Research*, vol. 22, no. 4, Apr. 2020, doi: 10.2196/19016.
- [26] D. C. Stokes, A. Andy, S. C. Guntuku, L. H. Ungar, and R. M. Merchant, "Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling," *Journal of General Internal Medicine*, vol. 35, no. 7, pp. 2244–2247, May 2020, doi: 10.1007/s11606-020-05889-w.
- [27] B. X. Tran *et al.*, "Studies of novel coronavirus disease 19 (COVID-19) pandemic: a global analysis of literature," *International Journal of Environmental Research and Public Health*, vol. 17, no. 11, pp. 1–20, Jun. 2020, doi: 10.3390/ijerph17114095.
- [28] "Data Never Sleeps 7.0," *Domo*, 2020. <https://www.domo.com/learn/data-never-sleeps-7> (accessed Dec. 2020).
- [29] R. N. Waykole and A. D. Thakare, "A review of feature extraction methods for text classification," *International Journal of Advance Engineering and Research Development*, vol. 5, no. 4, pp. 351–354, 2018.
- [30] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*, Mar. 2016, pp. 2264–2268, doi: 10.1109/WiSPNET.2016.7566545.
- [31] M. Jiang, C. Zhao, Z. Mo, and J. Wen, "An improved algorithm based on bloom filter and its application in bar code recognition and processing," *Eurasip Journal on Image and Video Processing*, vol. 2018, no. 1, Dec. 2018, doi: 10.1186/s13640-018-0375-6.
- [32] M. Misuraca, M. Spano, and S. Balbi, "BMS: an improved dunn index for document clustering validation," *Communications in Statistics - Theory and Methods*, vol. 48, no. 20, pp. 5036–5049, Oct. 2019, doi: 10.1080/03610926.2018.1504968.
- [33] S. Seshathriathithyan, M. V. Sriram, S. Prasanna, and R. Venkatesan, "Affective - hierarchical classification of text - An approach using NLP toolkit," in *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2016*, Mar. 2016, pp. 1–6, doi: 10.1109/ICCPCT.2016.7530228.
- [34] R. Adhitama, R. Kusumaningrum, and R. Gernowo, "Topic labeling towards news document collection based on latent dirichlet allocation and ontology," in *Proceedings - 2017 1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, Nov. 2017, pp. 247–251, doi: 10.1109/ICICoS.2017.8276370.

BIOGRAPHIES OF AUTHORS



Sujatha Arun Kokatnoor    has completed her Doctoral Thesis from CHRIST (Deemed to be University) in the Machine Learning Area of Anomaly Detection in Social Media Networks. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering, in CHRIST (Deemed to be University). Her area of research includes in the field of data science, machine learning, and deep learning. He can be contacted at email: sujatha.ak@christuniversity.in.



Balachandran Krishnan    has completed his Doctoral Thesis from Anna University in the Data Mining area of Ensemble Modeling. He is currently working as a Professor in the Department of Computer Science and Engineering, in CHRIST (Deemed to be University). His area of research includes data science, image processing and parallel computing. He can be contacted at email: balachandran.k@christuniversity.in.