Semantic Pattern Detection in COVID-19 Using Contextual Clustering and Intelligent Topic Modeling

Pooja Kherwa, Maharaja Surajmal Institute of Technology, Delhi, India
Poonam Bansal, Maharaja Surajmal Institute of Technology, Delhi, India

ABSTRACT

The COVID-19 pandemic is the deadliest outbreak in our living memory. So, it is the need of hour to prepare the world with strategies to prevent and control the impact of the pandemic. In this paper, a novel semantic pattern detection approach in the COVID-19 literature using contextual clustering and intelligent topic modeling is presented. For contextual clustering, three level weights at term level, document level, and corpus level are used with latent semantic analysis. For intelligent topic modeling, semantic collocations using pointwise mutual information (PMI), and log frequency biased mutual dependency (LBMD) are selected, and latent dirichlet allocation is applied. Contextual clustering with latent semantic analysis presents semantic spaces with high correlation in terms at corpus level. Through intelligent topic modeling, topics are improved in the form of lower perplexity and highly coherent. This research helps in finding the knowledge gap in the area of COVID-19 research and offered direction for future research.

KEywORDS

Latent Dirichlet Allocation, Latent Semantic Analysis, Log Frequency Biased Mutual Dependency, Mutual Information, Point Wise, Vector Space Model

INTRODUCTION

The Coronavirus family comprises of a wide range of animal and human viruses. Coronaviruses are positive-sense RNA viruses and are classiðed into four genera: Alpha, Beta, Gamma, and Delta-coronaviruses. (Weiss & Leibowitz.,2011; Burrell et al., 2016). Alpha coronaviruses and beta-coronaviruses are found exclusively in mammals, whereas gamma coronaviruses and deltacoronaviruses primarily infect birds. Prior to 2003, members of this family were believed to cause only mild respiratory illness in humans.

The 2003 epidemic of SARS-Cov prompted an intensive research for novel coronaviruses, resulting in the detection of a number. of novel coronaviruses in humans, domestic animals and wildlife. This research finds the greatest discovery, which suggest that bat and avian species are the natural reservoirs of the viruses (Guo,2020). Recent studies also discover that these coronaviruses are the result of recent cross species transmission events.

DOI: 10.4018/IJEHMC.20220701.oa7

The emerging of novel corona virus(2019-nCov) has awakened the echoes of SARS-Cov from nearly two decades ago (Gralinski & Menachery 2020). This zoonotic human coronavirus of the century emerged in Dec 2019, with a cluster of patients with connection to Huanan south China sea food market in Wuhan, Hubei Province China. Similar to severe acute respiratory syndrome (SARCov) and Middle east respiratory Syndrome coronavirus (MERS-Cov) infections patients exhibited symptoms of viral, pneumonia including fever difficulty in breathes and bilateral lung infiltration in the most severe cases (Wuhan Municipal Health Commision,2020).

Since its emergence in China in Dec 19, the coronavirus is spreading very fast in the entire world. Till 8th June globally there have been 6,881,352 confirmed cases of COVID-19, including 399,895 deaths, reported to WHO. According to country wise detail data from WHO dashboard on 8th June 2020, United states of America has the highest number of confirmed cases of 1915712, And at second largest confirmed cases in Brazil with 672846, and then Russia is at third position with 467673 and United Kingdom with 284,872 confirmed cases. Till 8th June India has total 265740 confirmed corona cases, out of which 129358 are active cases and 128894 recovered successfully, and unfortunately 7473 deceased [https://www.covid19india.org/]

India as the 2nd largest most populated country of world after china, where the first Covid-19 case emerged in Kerala on Jan 30,2020, which originally originated from China. Till 20 March, India observed around 223 confirmed cases and out of 4 lost their lives due to this pandemic. Indian Government has taken all the necessary step to tackle the pandemic in our country.

Till today the Covid -19, pandemic shows no sign of abating, as vaccine is yet to found. Although all the countries are trying to control with lockdown and local and global social distancing. Even in some countries situation is under control, Government started unlocking the country in phases with necessary precautions.

The researchers in different part of worlds are trying their best in different research labs and individually in different fields of medicine, bioinformatics, virology, technology, Data analytics, artificial intelligence to help the humanity to tackle this horrible epidemic with minimum loss.

Data scientist and analytics with advanced machine learning, deep learning algorithms try to predict the number of infected people in the future, also try to predict number of susceptible populations, so that government can take necessary action like implementation of lockdown, building necessary healthcare infrastructure.

In this paper, our approach towards Covid-19 pandemic is using distributional semantics, here the emphasis is to present semantic pattern in available literature of Covid-19, through contextual hierarchical clustering and intelligent topic modeling. For contextual hierarchical clustering implementation latent semantic analysis with novel three level weights at term level, document level and corpus level are used. We choose two three level semantic space, ATC – Augmented weighting at term level, log term frequency at document level and Cosine normalization at corpus

level and, NPC-Neutral at term level, probabilistic weighting at document level and Cosine normalization at corpus level.

Intelligent topic modeling is implemented using semantic collocations selection using point wise mutual information (PMI) and log frequency biased mutual dependency (LBMD) and then latent dirichlet allocation is applied. To show the effectiveness of our proposed methodology, both the approaches are compared with neutral weights at three level in contextual hierarchical clustering and traditional topic modeling algorithm latent dirichlet allocation.

The paper begins with data collection understanding, followed by 4 stages of analysis

1. Keyword trend analysis.
2. Contextual Hierarchical clustering in three semantic spaces
3. Cosine Similarity score analysis of term pair in three semantic spaces
4. Topic Modeling of Dataset using Intelligent Latent Dirichlet Allocation.

Related work

A study analyzed, trips from Wuhan to other parts of China including different mode of transport (air, road, train) between 370 cities in china and special administrative region of Hong Kong and Macau from data Dec3,2019 -Jan 24, 2020.Here Non homogeneous poison process model is constructed to predict the risk of infection in the traveler coming to Wuhan city and resident of Wuhan city (Lim et al.,2020).

In another study clinical finding of the patient, who was the first person become a carrier of territory transmission outside china. This is medical study where use of medicine on this patient with medicine lopinavir/ritonavir in treatment, in different stages of treatment and its effects are analyzed (Kim et al.,2020).

A model developed at John Hopkins University uses a stochastic simulation model which aims to mitigate pandemic at the one-set of the outbreak. The Meta population model connect airport network at global scale. In each airport a discrete time susceptible exposed infected recovered model is implemented to model the 2019-Covid spread (Biswas & Sen,2020; Li et al.,2020).

In other study, to predict the impact of disease at several level, many preliminary mathematical models are formulated by various research organizational groups. These insights will help as input for designing strategies to control the epidemics. In a study susceptible exposed infected framework is formulated to prevent epidemics during large events. e. g During parties or concert with huge crowd (Du et al.,2020).

A research in south Korea was conducted which attempts to isolate the pathogen from Covid-19 patients. In this study upper and lower respiratory tract secretion sample from putative patients with Covid-19 were inoculated on the cells to isolate the virus. Full genome sequencing and electron microscopy were used to identify the virus (Yunlu,2020).

In another approach where the three variants of genome sequence of Corona virus-Covid-19 by using amino acid, which are named as A, B, C, with A being the ancestral type are analyzed using phylogenetic network analysis (Forster et al.,2020).

A non-pharmaceutical intervention for preventing and controlling this deadly Covid -19 infectious disease is desirable. IoT (internet of things) and machine learning methods has sufficient potential to contribute in this time (Chakraborty,2019). A machine learning model based study has been done to predict the infection in Mexico city (Muhammad1 et al.,2020).

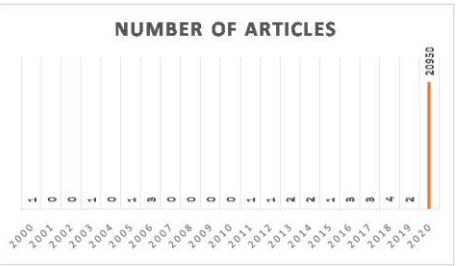
A secure, privacy concerned IOT (internet of things) inspired model for monitoring of epileptic patients are proposed (Gupta et al.,2019), It can also be utilized in this critical pandemic of SAR-Cov-2. In this pandemic major factor to contain the disease is social distancing and movement controlling. Automated digital contract tracing is effective and efficient technology. A hardware based model that capture movement information and contract of object is developed using IoT techniques (Garg et al.,2020).

Methodology

Data Collection

In this paper authors have used a collection of 21323 articles published on Covid-19 till on 21st May 2020.This is collected from WHO-Covid-19 Database, Who-Covid-19 database is an open research dataset growing through resources of scientific papers published by various researchers in the entire world on Covid-19 and related to historical coronavirus resources. In this collection, articles on coronavirus from year 2000 to 2020 are collected, and the publication trend as shown in figure 1 is observed. Before the outbreak of coronavirus in Dec 2019 in Wuhan city of China, very few publications exist in the dataset.

Figure 1. Publication Trend in Data set Covid-2019



Preprocessing of Dataset

The dataset is preprocessed before topic modeling to reveal the semantic themes in the huge data set collection. The data set is cleaned by using basic text mining tools. The dataset contains some articles in the Covid-19 data set in languages other than English like Chinese and German, for better interpretation of results only those articles written in English are considered and others are removed from the dataset. 568 articles of other languages are removed. The abstract of all 20755 documents are extracted as a single corpus object, and then all stop word are removed, all punctuation symbols are removed, all capitals are converted into lowercase, all numbers are removed and finally the corpus object is converted into a document term matrix for final topic modeling. To consider the importance of each keyword in dataset the term frequency and inverse document frequency weight mechanism during document term matrix construction.

Parameter Setting

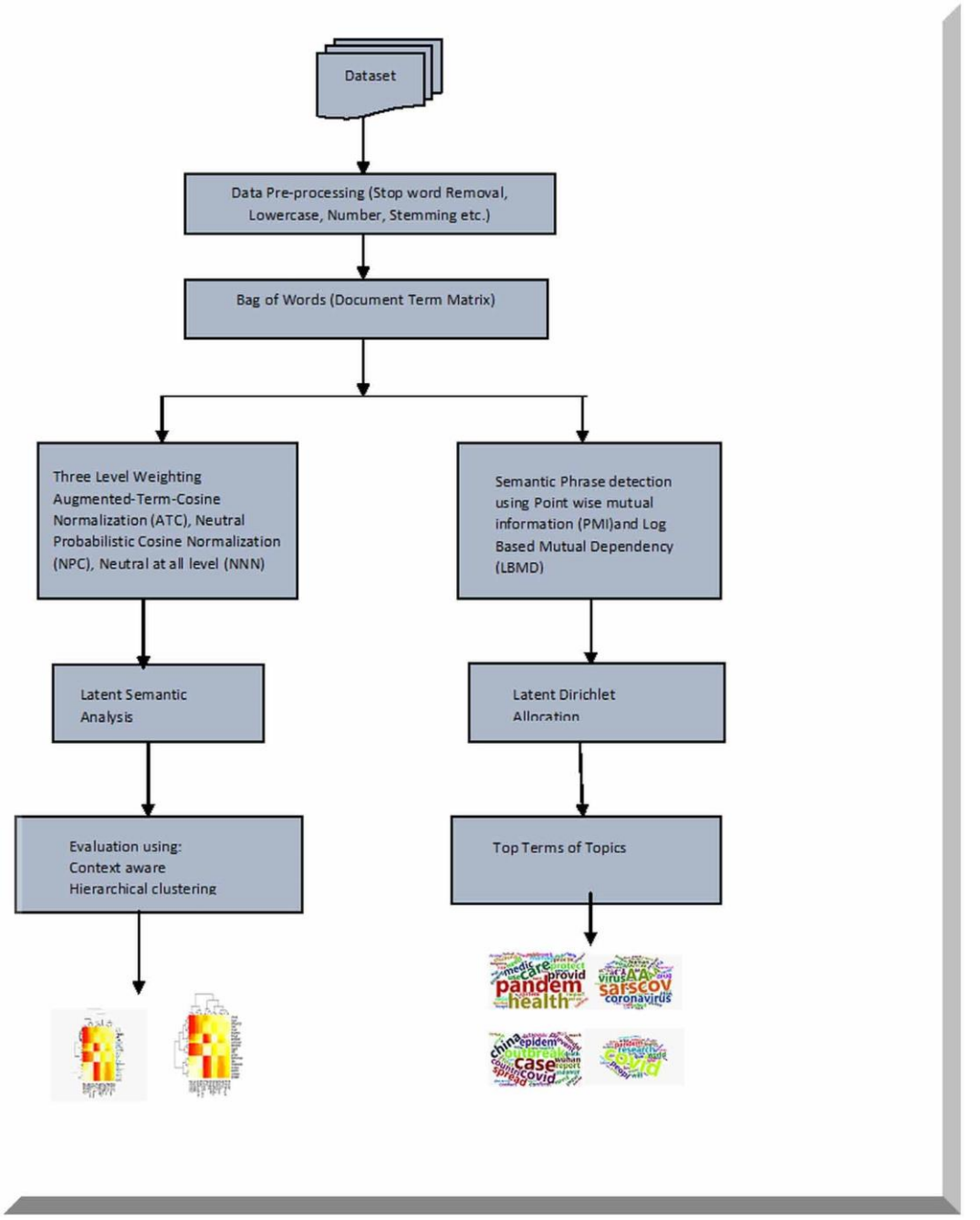
In this semantic theme detection research of Covid-19 Data set, we have used two technique known as Latent semantic analysis and Latent Dirichlet Allocation. For significant semantic pattern detection through Latent semantic analysis, three level weights at term, document and corpus level are used. After preprocessing step from collection of 20755 documents, we get the document Term matrix of 786*20755, with term frequency weights. After this our novel three level weight are used on document term matrix (Deng et al.,2004; Debole & Sebastiani 2003). In second phase of analysis, we applied topic modeling algorithm latent Dirichlet allocation, with novel intelligent phrase detection using point wise mutual information(PMI) and log biased mutual dependency(LBMD) for enhanced semantic themes in data set.

Number of Topics: In topic modeling techniques, the most important factor is choosing the number of topics. Topics are chosen in such a way that truly explore the dataset, and also able to find existing semantic themes in data set as accurate as human. In probabilistic method, there are many techniques exist [Cao Juan et al.,2009; Deveaud et al.,2014; Griffiths and Steyvers,2004) but which one to choose for dataset in hand is again a big question. Most of the techniques use likelihood method, and when executed with certain range of topics, they converge either on the

lowest value of topics i.e. at the starting point and in some case converges at the highest number of topics. In both the cases it become very confusing to choose the right approach, choosing too few numbers of topics will not be able to explore the dataset, and so many numbers of topics result in overlapping of topics. So in this work we use a very efficient techniques given by Arun (Arun et al.,2010), in this approach the normalized form of matrices generated from latent Dirichlet output, known as word topic matrix, and document -topic matrix is used and the K-L divergence between these matrices are calculated, best values of topics are chosen at the point where this divergence is minimum.

So, we have chosen 6 as the number of topics, and semantic themes broadly exist in dataset.

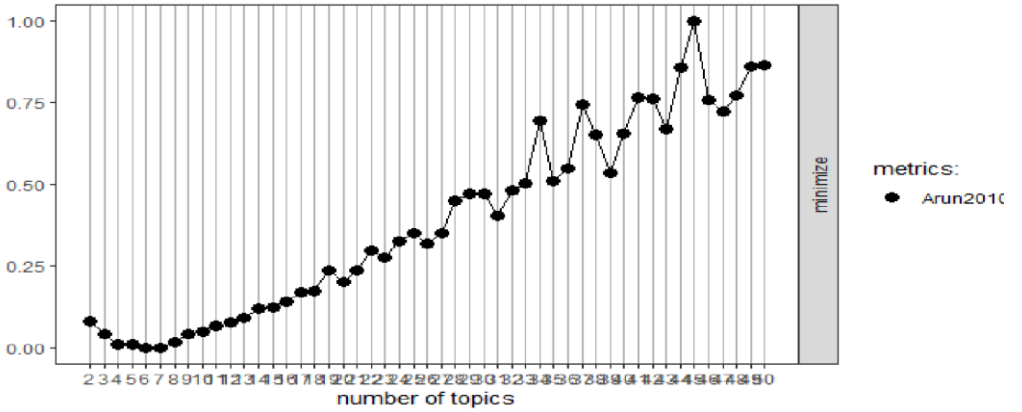
Figure 2. Methodology for proposed Approach to Cognitive semantic themes Detection in Covid-19 Dataset.



Latent Semantic Analysis

LSA is a two-step process, the first step is to create a term document matrix consists of document collection, where each row represents all terms in document collection and each column represents individual document in document collection (Deerwester et al.,1990), each cell in this matrix contains

Figure 3. Optimal number of topics selection.



the frequency with which the term of its row appears in the document denoted by its column. So, first step of latent semantic analysis is creating a term document matrix with term frequency as a basic weighting method for each term.

In second step single value decomposition is used on term document matrix, basically SVD is a dimension reduction technique (Papadimitriou et al.,2000), which decompose our m*n matrix (where m is number of terms and n is the number of documents) into a product of three matrices.

$$A = UWV^{T} \quad (1)$$

The component U, describes the original row entities in A -i.e. the term matrix (n*n), V matrix describes original column entities in A describes the document matrix (Gefen, et al.,2017). The third component W is a n*n diagonal matrix of singular values. The quality of factorization of LSA is that matrix (term document matrix) decomposed so perfectly that if we retain only the k greatest singular values in W and retain in U and V the column corresponding to those values then the product resulting matrices U, W, V is the best approximate of rank k.

Three Level weight

Three level weight is a concept inspired from Salton. (Salton et al.,1988; Buckley et al.,2004) where three factors are considered for assignment of appropriate weight to every single term. These three factors are

1. Total term count in corpus or document collection represented with term frequency.
2. Second factor is collection frequency factor that consider or separates relevant document from irrelevant documents. For. e.g. inverse document frequency is considered to increase the terms discriminating power in document collection.
3. Third factor is a way of considering the document length for analysis, a cosine normalization factor is incorporated to equalize the length of documents.

In this approach, we choose two three level semantic space, ATC – Augmented weighting at term level, log term frequency at document level and Cosine normalization at corpus level and, NPCNeutral at term level, probabilistic weighting at document level and Cosine normalization at

corpus level. Equation for Augmented(a), Term frequency(t), Probabilistic(P)and cosine normalization(C) are given below.

$$a = \frac{0.5 + 0.5 * tf_{t,d}}{\max(t\, tf_{t,d})} \quad (2)$$

$$t = idf = \log \frac{N}{df_t} \quad (3)$$

$$P\,prob() = \max[0, \log \frac{N - df_t}{df_t}] \quad (4)$$

$$c(cosine) = \frac{1}{\sqrt{w_{12} + w_{22} + \ldots w_{m2}}} \quad (5)$$

Algorithm 1: Contextual Hierarchical Clustering

1. Data set is prepared in .csv format from collection of text files.
2. Data is preprocessed using all necessary step like stemming, punctuation removal, stop-word removal, and all whitespaces etc. and corpus object is made
3. Corpus object of data set is converted into Term document matrix for further text processing.
4. Three level of weights are applied to document Term matrix as:

a) ATC – Augmented weighting at term level, log term frequency at document level and Cosine normalization at corpus level

b) NPC-Neutral at term level, probabilistic weighting at document level and Cosine normalization at corpus level.

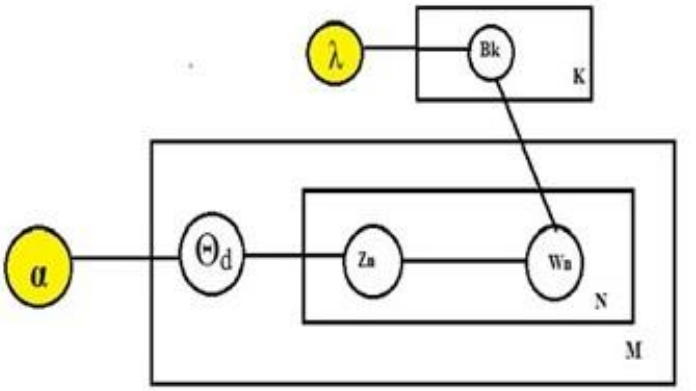
5. Latent Semantic Analysis function is applied to matrices generated in step 4. this will generate two latent semantic spaces known as ATC-Latent semantic space and NPC-Latent semantic space.
6. Using cosine similarity score for specific term, contextual hierarchical clusters are generated in both the semantic spaces.

Latent Dirichlet Allocation

Inspired from the very popular vector space assumption of text mining, known as ‘bag of words’ assumption, where the order of words in a document can be ignored. The theory of probabilistic language model, like latent Dirichlet allocation are founded on the assumption of exchangeability (Blei et al.,2003) It state that documents are exchangeable, and the order of documents can also be neglected. The idea of Latent Dirichlet Allocation (LDA) is based on the concept that one document exhibits multiple topics in different proportion, and each topic is defined as a distribution of fixed

set of words. For example, the document of Sports has vocabulary of sports as well as health and education. So, the document contains words related to all the three topics, and each topic has its fixed vocabulary that defines it. But how much proportion of these topics a document contains is a big challenge.

Figure 4: *Latent Dirichlet Allocation (LDA)Plate notation [***Griffiths** *et al.,2007]*



LDA formally cast the concept of semantic themes or topic detection through hidden variable model of documents. In these models the semantic themes in document collection are considered as hidden variables, and words in collection are observed variables, the process of learning the topic distributions in these document, and word distributions in topics is described through plate notation in figure 4.

Distribution of latent variables given the document

$$P(\vartheta,z \mid w,\alpha,)\beta = \frac{P(\vartheta,z,w \mid \alpha,)\beta}{P\, w(\mid \alpha\, \beta,)} \quad (6)$$

Proposed Intelligent Latent Dirichlet Allocation

It is extension of traditional topic modeling algorithms where the traditional theory of text mining algorithm is challenged. Traditional topic modeling works on the principal of 'bag of words' approach and also 'exchangeability', which state that the order of documents in a corpus does not matter, and in latter order of words in documents does not hold much weightage in text mining. Very few studies consider the importance of semantic order of words in text mining (Wallach. 2006).In this study an novel intelligent phrase refinement using two statistical measure known as point wise mutual information(PMI)(Gerlof Bouma,2009)and log frequency biased mutual dependency(LGMD) (Church and Hanks,1990) are applied to select only meaningful semantic phrases for topic modeling of Covid -19 dataset, at preprocessing level, semantic order between words are captured using these two exclusive statistical measures, and only those terms or phrases are considered in topic modeling, those crosses a basic threshold of these metric's statistical score.

Point wise Mutual Information (PMI)

It is a metric based on of how much the actual probability of a particular co-occurrence of events P (W1, W2) differs from what we would expect it to be on the basis of the probabilities of the

individual events. In PMI, it is assumed that rare events contain more information than frequent events. This means that the PMI of perfectly correlated words is higher when the combination is rarely occurring.
PMI can be interpreted as a measure of independence rather than as a measure of correlation.

$$I\,W(\ 1,W2) = \log2\ \frac{PW(\ 1,W2)}{PW(\ 1)PW(\ 2)} \tag{7}$$

Log Frequency Biased Mutual Dependency (LBMD)

Mutual dependency can be calculated in the phrases by subtracted from PMI the information that the whole event bears, which is self-information for any event X.

$$I\,X(\ ) = -\log P\,X(\ ) \tag{8}$$

So mutual dependency (MD) can be defined between two co-occurring word pair w1 and w2

$$DW(\ \ 1,W2) = I(W\,W1,\ \ 2) - I\,W(\ 1W2) \tag{9}$$

$$DW(\ \ 1,W2) = \log2\ \frac{P\,W(\ 1,W2)}{P\,W(1)P\,W(2)} \tag{10}$$

Mutual dependency will be maximized for perfectly dependent phrases or statistical confidence; it is suggested that slight bias towards frequency can be beneficial. So, a new measure known as log frequency biased MD can be defined as

$$D_{LF}(W1,W2) = DW\,W(1,\ \ \ 2) + \log2P\,W(\ 1W2) \tag{11}$$

In other words, it is combination of Mutual Dependency with T-score

Algorithm 2: Intelligent Topic Modeling

1. Data set is prepared in .csv format from collection of text files.
2. Data is preprocessed using all necessary step like stemming, punctuation removal, stop-word removal, and all whitespaces etc. and corpus object is made
3. Corpus object of data set is converted into Term document matrix for further text processing
4. Collocation function is applied to term document matrix to construct semantic phrases up to N-Gram.
3. Semantic collocation (phrases) are selected using

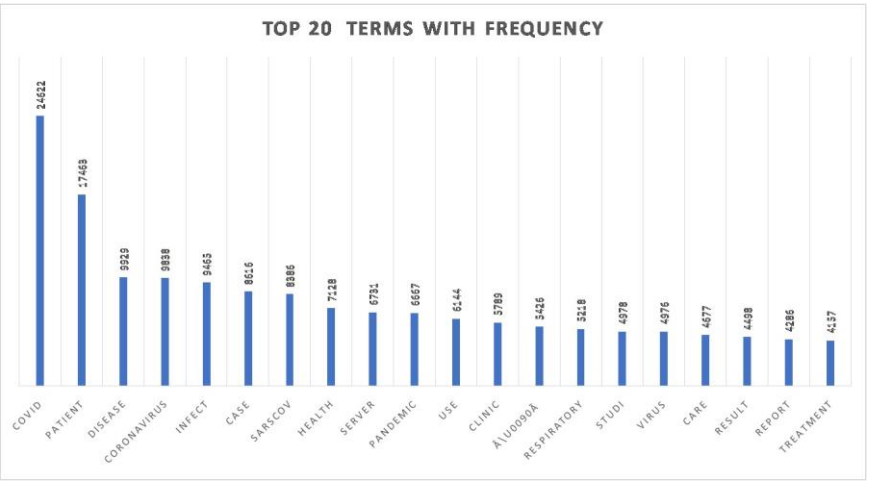
a) Point wise mutual information (PMI) score
b) Log Biased Mutual Dependency (LBMD) score

4. Semantic collocation matrix is constructed and latent dirichlet allocation is applied for intelligent topic modeling

Results and Analysis

In this paper to understand the literature of Covid-19, two topic modeling technique called Latent Dirichlet Allocation and Latent semantic analysis are used. Various details of dataset are analyzed.

Figure 5. Top 20 terms with frequency in dataset.



word Level Analysis

Before detailed data analysis through topic modeling techniques, authors explore the data set, by exploring the frequency of terms in dataset. We consider top 20, the most highly frequent terms used in various research papers in last decades. Figure 5, shown the terms with their frequencies.

Context Aware Hierarchical Clustering in three semantic spaces.

In this analysis of latent semantic analysis (LSA), using all the three proposed weight for latent semantic analysis, the three semantic spaces are constructed known as Latent semantic space-NNN, Latent semantic space-NTC, and Latent semantic space-ATC. Then in these semantic spaces context aware clustering for specific term is generated using heat-map. In this hierarchical clustering all the correlated term using cosine similarity measure are clustered together. These heat maps are constructed from 20 closest term to a specific term “antibody”. The pattern in the heat-map shows the association between rows and columns. Hierarchical clustering in heat-maps are formed based on distance and similarity between them. These contexts aware hierarchical clustering is shown in Figure 6(a-c). In

Figure 6(a). NNN-Latent Semantic Space

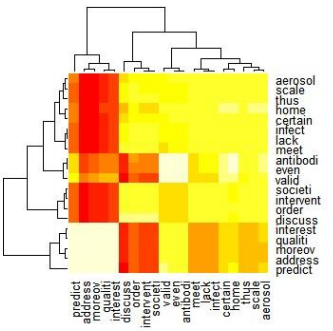


Figure 6(b). NPC-Latent Semantic space

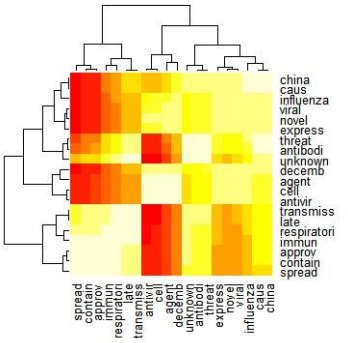


Figure 6(c). ATC-Latent Semantic Space

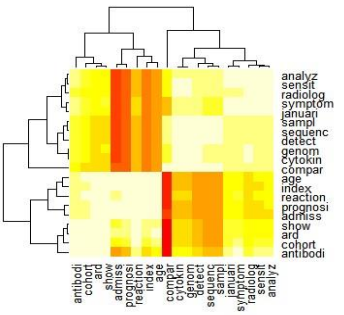
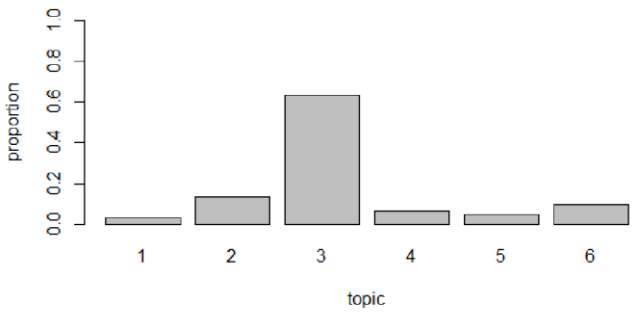


Table 1. Cosine similarity score of different terms in all the three latent semantic spaces

Chosen Terms	Neutral (NNN)-Latent Semantic Space	NPC-Latent Semantic space	ATC-Latent Semantic Space
	Similarity Score		
antibody, influence	0.48	**0.51**	0.51
suspect, immun	0.015	**0.70**	**0.39**
antivir, articl	0.081	**0.71**	**0.62**
diagnos, death	0.103	**0.71**	**0.79**
cov,contain	0.058	**0.64**	**0.59**
symptom,suspect	0.20	**0.95**	**0.90**
mortal,radiolog	0.09	**0.97**	**0.81**
cough,cardiovascular	0.063	**0.91**	**0.66**
chest,diabet	0.069	**0.96**	**0.83**
genom,organ	0.069	**0.60**	**0.64**

Figure 7 (a) Documents-Topic Proportion in LDA Topic Modeling



these heat maps we use two color schemes each represents semantic relatedness score between terms related to "antibody", Here red color shows similarity score between (0-0.4), light red (0.4-0.6), orange (0.6-0.7), yellow color shows 2nd highest semantic relatedness (0.8-0.9) and white color indicates 1, means terms are exactly same. In figure 6(a) all the terms related to "antibody" are clustered in Latent semantic space-NNN, and their correlation score to each other in the space are shown as heat-map. In figure 6(b) and 6(c) all terms clustered are shown in NTC and ATC latent semantic spaces. In all the three heat-maps, our main emphasis is to discover more number of terms with higher correlation score or high similarity score with term 'antibody'. Here latent semantic space-ATC has the highest number of terms with correlation score of around 1, and next is latent semantic space-NTC and in this scenario semantic space NNN has very few term with correlation score of 1. So we get terms with maximum correlation in proposed three level semantic spaces NTC and ATC.

Figure 6. Semantic space term correlation for "Antibody' in different weight spaces

Cosine Similarity score of term pair in three latent semantic spaces

In this experiment, authors have shown the effectiveness of three level weights in latent semantic analysis of covid -19 dataset. Here in table 1, cosine similarity score between different term pair are calculated in all the three semantic spaces. It is clearly visible in the results that for all term pair, in

Figure 7 (b) Documents-Topic Proportion in Intelligent-LDA Topic Modeling

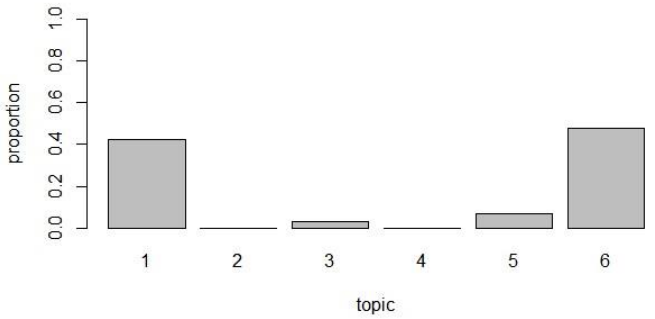
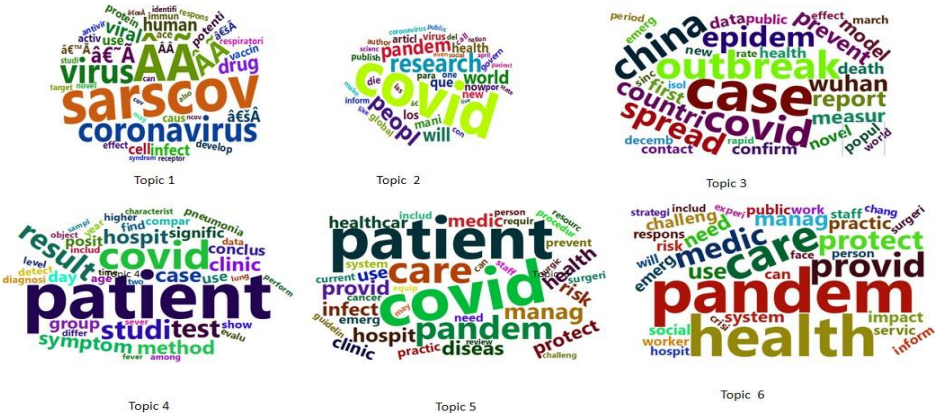


Figure 8. Six topics with top 15 terms as word clouds.



latent semantic space NPC and latent semantic space ATC the terms come closer contextually to each other, that shown more cosine similarity score.

Topic Modeling

In next phase of Covid-19, we use Latent Dirichlet allocation (LDA) algorithm for detecting semantic themes in the data set. The number of topics chosen is at 6 as shown in figure 2. After fixing the number of topics for the Covid-19 dataset at 6, we apply Latent Dirichlet allocation algorithm to the Document term matrix obtained from preprocessing of dataset. The six main semantic themes with nearby correlated terms are shown as word cloud in figure 8. The first topic Topic #1 is about severe acute respiratory Syndrome Coronavirus and it contains all terms related to this virus, including its genome structure like cell, protein, respiratory etc. Topic #2 contains information regarding Covid pandemic and its associated publication, third topic is about outbreak of pandemic in china, Wuhan and its spread in different parts of worlds, Topic #4 is about Covid-Symptoms, test, medicine, hospitals, precautions, diagnosis etc. Topic #5 is about Patients, clinical guidelines, healthcare infrastructure, management of pandemic etc. Topic #6 is about challenges of this pandemic and its effects like economy burst due to lockdown, medical care and other services like social work. So through this topic modeling technique and optimal parameter selection instead of random selection, we explore Covid -19 dataset of 20755 document in such an efficient way without overlapping of topics. All the researches are going very fast in multiple perspectives disciplines including biology, medicine, virology, bioinformatics and machine learning to help the humanity to handle this pandemic of Covid-19. These topic modeling results will be very helpful for intelligent information retrieval regarding the covid-19 for the overall benefits of humanity.

In next phase of analysis, we have done modeling using Intelligent LDA, in this phrase refinement using log based mutual dependency and point wise mutual information has been done. In this analysis we found more crisp topics in Covid-19 Data set, and perplexity of topic modeling with same number of parameters improved a lot, also this intelligent topic modeling provides more cohesive topics, because the model itself in the initial phase calculated very critical statistical measures, known as Point-wise mutual information(PMI) and log based mutual dependency(LBMD). After selection of quality phrases, the document term matrix is constructed for further topic modeling. In Figure 7(a)7(b), Document topic proportion in all the six topics are shown in both latent Dirichlet allocation topic modeling and in intelligent latent Dirichlet allocation topic modeling. Here intelligent phrase refinement shown an indication that 50% documents fits in topics 6 and 40% topic 1, then only 5% documents topic 5, and remaining 5% comprises the topic 2 and 4. Where in Latent Dirichlet allocation model, topic 3 contain 60% of documents, topic 2 contain 15% of documents, and topic 6 contain 10% of documents, topic 4 contains 5% of

documents and topic 5 contains 3% of documents, and topic 1 contain only 2% of documents. And when we calculated perplexity in both model at number of topics 6, the latent Dirichlet allocation has 822.32 and intelligent phrase refinement topic modeling has perplexity value is 445.2167. It is a great improvement of intelligent phrase refinement for topic modeling. For quality topic model the lower perplexity value is considered best.

CONCLUSION

When the entire world is suffering from Covid-19 pandemic, it is very important to understand the pandemic from multiple perspectives like virology, medicine, bioinformatics, economics, Artificial intelligence, Epidemiological models.

In this paper, authors proposed a novel semantic pattern detection approach based on contextual hierarchical clustering and intelligent topic modeling- to explore the Covid-19 literature, till May 23, 2020.The data set contain around 21323 documents. Data set is analyzed using latent semantic analysis with three level weights at term, document, and corpus level. These weights are known as NTC (Neutral, inverse document frequency, Cosine normalization) and APC (Augmented, Probabilistic, Cosine normalization). For evaluation of results, we compare these two semantic spaces results with NNN (Neutral at three level) or no weights except term frequency at document term matrix. In all the semantic spaces we compare the results using co-term similarity using cosine similarity score, which shows how close they appear contextually in all the three semantic spaces. In this proposed three level weighted NTC-latent semantic space and APC –latent semantic space shown significant improvement as compare to NNN-latent semantic space as shown in table 1. Also significant improvement in contextual hierarchical clustering using exploratory data analysis shown in Figure 6(a-c)

In next level of Covid-19 corpus analysis, authors used Latent Dirichlet allocation and intelligent latent Dirichlet allocation topic modeling techniques to find the topics in the dataset. Intelligent latent Dirichlet allocation has shown the lower perplexity values and more cohesive topics.

In the future, it is advisable that novel topic modeling techniques based on contextual semantics should be used in bioinformatics. Like genome sequence of SAR-Cov-19, can be explored using non-negative matrix factorization with its various variants for efficient pattern mining, to know the structure of SAR-COV viruses in more detailed way.

REfERENCES

Arun, Suresh, Madhavan, & Murthy. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg. doi:10.1007/978-3-642-13657-3_43

Biswas, K., & Sen, P. (2020). *Space-time dependence of corona virus (COVID-19) outbreak.* arXiv preprint arXiv:2003.03149.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information. *Proceedings of the Biennial GSCL Conference*, 31–40.

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1994). Automatic query expansion using SMART: TREC 3. *Proc. of the Third Text Retrievals Conference*, 69–80.

Burrell, C. J., Howard, C. R., & Murphy, F. A. (2016). Coronaviruses. In Fenner and White's Medical Virology. Academic Press.

Chakraborty, C. (2019). Advanced Classification Techniques for Healthcare Analysis. IGI Global. , 2019. doi:10.4018/978-1-5225-7796-6

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, *16*(1), 22–29.

Debole, F., & Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In *SAC '03: Proceedings ofthe 2003 ACM symposium on Applied computing*. New York, NY: ACM Press. doi:10.1145/952532.952688

Deerwester, S., Tumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *J. Soc. Inform. Sci*, *41*(6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AIDASI1>3.0.CO;2-9

Deng, Z.-H., Tang, S.-W., Yang, D.-Q., Zhang, M., Li, L.-Y., & Xie, K. Q. (2004), A comparative study on feature weight in text categorization. In AP Web (vol. 3007). Springer-Verlag Heidelberg.

Deveaud, SanJuan, & Bellot. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61–84. 10.3166/dn.17.1.61-84

Du, Z., Wang, L., Cauchemez, S., Xu, X., Wang, X., Cowling, B. J., & Meyers, L. A. (2020). Risk for transportation of coronavirus disease from Wuhan to other cities in China. *Emerging Infectious Diseases*, *26*(5), 1049–1052. doi:10.3201/eid2605.200146 PMID:32053479

Forster, P., Forster, L., Renfrew, C., & Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9241–9243. doi:10.1073/pnas.2004999117 PMID:32269081

Garg, L., Chukwu, E., Nasser, N., Chakraborty, C., & Garg, G. (2020). Anonymity preserving IoT-based COVID-19 and other infectious disease contact tracing model. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 159402–159414. doi:10.1109/ACCESS.2020.3020513

Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J. L., & Larsen, K. R. (2017). A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code: Studying Online Reviews and the Stack Exchange Community. CAIS, 41, 21.

Gralinski, L. E., & Menachery, V. D. (2020). Return of the Coronavirus: 2019-nCoV. *Viruses*, *12*(2), 135. doi:10.3390/v12020135 PMID:31991541

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(suppl 1), 5228–5235. doi:10.1073/pnas.0307752101 PMID:14872004 Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation. *Psychological Review*, *114*(2), 211–244. doi:10.1037/0033-295X.114.2.211 PMID:17500626

Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., Tan, K.-S., Wang, D.-Y., & Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. *Military Medical Research*, *7*(1), 1–10. doi:10.1186/s40779-020-00240-0 PMID:31928528

Gupta, A. K., Chakraborty, C., & Gupta, B. (2019). Monitoring of Epileptical Patients Using Cloud-Enabled Health-IoT System. *Traitement du Signal*, *36*(5), 425–431. doi:10.18280/ts.360507

Juan, C., Tian, X., Li, J., Zhang, Y., & Sheng, T. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing — 16th European Symposium on Artificial Neural Networks*, *72*, 7–9. doi:10.1016/j. neucom.2008.06.011

Kim, J. M., Chung, Y. S., Jo, H. J., Lee, N. J., Kim, M. S., Woo, S. H., Park, S., Kim, J. W., Kim, H. M., & Han, M. G. (2020). Identification of Coronavirus Isolated from a Patient in Korea with COVID-19. *Osong Public Health and Research Perspectives*, *11*(1), 3–7. doi:10.24171/j.phrp.2020.11.1.02 PMID:32149036

Li, M., Chen, J., & Deng, Y. (2020). *Scaling features in the spreading of COVID-19*. arXiv preprint arXiv:2002.09199.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., & Epstein, J. H. (2005). Bats are natural reservoirs of SARS-like Coronaviruses. *Science*, *310*(5748), 676–679. doi:10.1126/science.1118391 PMID:16195424

Lim, J., Jeon, S., Shin, H. Y., Kim, M. J., Seong, Y. M., Lee, W. J., Choe, K.-W., Kang, Y. M., Lee, B., & Park, S. J. (2020). Case of the index patient who caused tertiary transmission of coronavirus disease 2019 in Korea: The application of lopinavir/ritonavir for the treatment of COVID-19 pneumonia monitored by quantitative RT-PCR. *Journal of Korean Medical Science*, *35*(6), e79. doi:10.3346/jkms.2020.35.e79 PMID:32080993 Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2020). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN Computer Science*, *2*(1), 1–13. PMID:33263111

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, *61*(2), 217–235. doi:10.1006/jcss.2000.1711

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi:10.1016/0306-4573(88)90021-0

Sameni, R. (2020). *Mathematical modeling of epidemic diseases; a case study of the COVID-19 coronavirus.* arXiv preprint arXiv:2003.11371.

Wallach, H. M. (2006). Topic Modeling: Beyond Bag of-Words. *Proceedings of the 23rd International Conference on Machine Learning*, 977–984. doi:10.1145/1143844.1143967

Wang, Y., Hu, M., Li, Q., Zhang, X.-P., Zhai, G., & Yao, N. (2020). *Abnormal respiratory patterns classiðer may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner*. arXiv:2002.05534.

Weiss, S. R., & Leibowitz, J. L. (2011). Coronavirus pathogenesis. Advances in Virus Research, 81, 85-164. doi:10.1016/B978-0-12-385885-6.00009-2

Wuhan Municipal Health Commission. (n.d.). *Wuhan Municipal Health and Health Commission's Brieðng on the Current Pneumonia Epidemic Situation in Our City*. Available online: http://wjw.wuhan.gov.cn/front/web/ showDetail/2019123108989

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C. L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., & Shi, Z.-L. et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *588*(7836),
E6. Advance online publication. doi:10.1038/s41586-020-2951-z PMID:32015507

Pooja Kherwa is an assistant professor in Maharaja Surajmal Institute of Technology, New Delhi. She received her M. Tech in information Technology from Guru Govind Singh Indraprastha University; Dwarka, New Delhi in 2010.Currently she is pursuing her PhD from Guru Govind Singh Indraprastha University, Dwarka- New Delhi. Her research interest includes Topic Modeling, Sentiment Analysis, Machine Learning.

Poonam Bansal, PhD, is a Professor of Maharaja Surajmal Institute in Technology. . She has received her PhD from Guru Govind Singh Indraprastha University, Dwarka, New Delhi in 2010. Her areas of interest include Speech Recognition, Data Mining, Machine Learning.