Social Media Analytics during Pandemic for Covid19 using Topic Modeling

Vrishali Chakkarwar
Computer Science and Engineering Department
Government College of Engineering Aurangabad, Maharashtra, India
vrush.a143@gmail.com

Sharvari Tamane
Information Technology Department
Jawaharlal Nehru Engineering College
Aurangabad, Maharashtra, India
sharvaree73@yahoo.com

Abstract*— The enttte wottd ts factng the Covtd19 pandemtc. Thts pandemtc has vattous consequences on the pottttcat, cuttutat, economtcat and soctat ttfe of the communtty. Loctdown has affected the psychotogtcat tmpact on soctety. Thts ts teftected tn vattous soctat medta sttes. In such a phase soctat medta anatyttcs fot twtttet data can be usefut fot undetstandtng pubttc optnton. In thts papet, we have apptted the Latent Dtttchtet Attocatton Atgottthm as a toptc modettng atgottthm. Toptc modettng ftnds the matn theme that petvades the tatge data set. Twtttet medta ts constdeted as the most poputat mtctobtoggtng ptatfotm, hence data duttng thts pandemtc ts exttacted ftom twtttet. Natutat tanguage ptocesstng Techntques apptted as pteptocesstng and then toptc modettng apptted whtch has gtven sattsfactoty tesutts tn tetms of petptextty as a petfotmance measute. Toptc exttacted gtves an tdea of the tmpact of Covtd19 on soctety thtough thett optnton on twtttet. Thts can be hetpfut fot mattng futute pottctes by pottcymatets. *c

ceycords—ctopiccmodeling,cnaturalclanguagecpreprocessing,c socialcmediacanalytics,ctcittercc

I. INTRODUCTION

Today as entire world is facing COVID19 pandemic. This public health emergency is due to coronavirus. Currently more than 2ss countries, more than 2ssssss confirmed cases are found in entire world. Coronavirus or COVID 19 is an infectious disease caused due to coronavirus. Currently there is no specific treatment for coronavirus, best way to prevent from virus is stop spread of disease or avoid the expose to coronavirus. Covid 19 symptoms are dry cough, mild fever and in latter stage it developed as SARS Severe Acute Respiratory Syndrome. During such COVID 19 Outbreak, it has become necessary to do Social Media Analysis for understanding opinion of masses to make better policy decisions and understand different issues happening in society. Social media analytic is a method of collecting data from websites or social media like Facebook, twitter, Instagram to understand opinion of public.

Microblogging sites like twitter are now a days are very popular on which people post their opinion regarding current issues, complains, different topics. They express their feedbacks regarding some products they use. Many manufacturing companies use their analytics on such posts to find their customers feedback regarding their product [1]. This is very challenging task to summarize the overall blogs. Here we are trying to build a model that extracts the tweets which is in unstructured text form and process it to understand the social media opinions regarding some issue.

In this paper, information retrieval model using topic modeling that analyzes the effect of "COVID 19 Outbreak in India is studied.

II. LITRATURE SURVEY

Kursuncu et al. have reviewed predictive analytics of twitter data which can be used for emotion analysis, sentiment analysis for various domains like US elections, public opinion for gun reforms, twitter analysis for Healthcare, Topical Analysis, Age prediction for twitter user, sales and stock prediction. It is also mentioned, the use of tweet analysis for network analysis and information propagation through network. Many machine learning algorithms like Random Forest, Naïve Bayes, Neural network can be used for classification and Term Frequency Inverse Document Frequency TFIDF, Latent semantic Analysis (LSA), Latent Dirichlet Allocation are used for deriving textual Feature Representation [2].

Colace et al. 2019, proposed method for classifying job vacancies using topic modeling and forecasting job market trends. LDA model is used for classification of text [3].

Cybersecurity is very important to protect their citizen and cyberspace. Cybersecurity policy is text document. It is important to understand cybersecurity policies of all countries, so that we can design sufficient policies to protect own countries cyberspace. This is possible by comparing National Cyber Security policies of different countries by investigating similarity and differences between them. This is done by using clustering and Topic Modeling. Topic Modeling gives the abstract idea of text documents. Results obtained are used by policy makers for understand and analysis of text documents for developing national policies and strategies [4].

III. PROPOSED SYSTEM

To perform social media analytics on twitter data we propose a model which automatically extracts the theme of dataset. This gives brief idea of public opinion during this pandemic. The model used here consists of different phases like data extraction, preprocessing, feature extraction and topic modeling in this research work. Fig. 1.1 shows the detail phases of social media analytics using topic modeling.

A. Different steps performed during Research

Fig. 1.1 Flow Diagram showing Social Media Data Analytics

B. Data Extraction from Twitter

It is a Microblogging or social networking site on which people interact with each other using tweets regarding everything in their day to day life. Tweets are nothing but short messages. Basic tweet contains username, hashtag and retweet. Twitter is a very useful and popular platform to understand the social opinion. In this work, twitter API is used to extract tweets. Duration of Pandemic March and April are considered. 3500 tweets are extracted for query keywords like 'Covid Outbreak in India', 'Corona', & 'Covid 19' are used.

C. Preprocessing

Twitter data is basically unstructured text document. Tweets contain many things like hash tag, URLs, emoji, alphanumeric characters. This data needs to be preprocessed by using natural language processing techniques. NLTK libraries are used for this purpose. First, we tokenize the tweet data into smallest unit which is a token in natural language processing tasks [2]. Each word is changed to lowercase, as many post URLs relating to some subjects, these URLs need to be eliminated, stop-word like 'the', 'of', 'for' have less semantic importance that's why these to be removed. We need to remove all tokens that comprise just of non-alphanumeric characters (this incorporates all emojis) and every single short token (<3 characters). At that point, we sift through all non-English and exceptionally short tweets (<3 tokens). Here retweets are not considered since they serve more as an underwriting as opposed to a unique content. Fig. 1.2 shows a simple example of tweet, its tokenization and stop word removal.

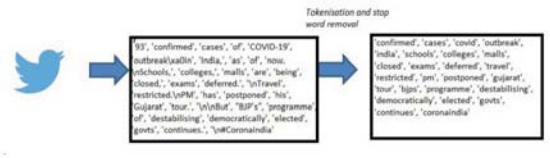

Fig. 1.2 Preprocessing steps for a typical Tweet

D. Feature Extraction

Any text mining algorithm cannot be applied to data which is in the form of unstructured text. To extract the features from s from raw text, text data should be represented in vector form. This vector is called as Bag of Words Model (BOW).

Term Frequency is the number of times the term occurs in document divided by total number of terms in document.

$$Tf_{i,j} = n_{i,j} / \sum_k n_{i,j} \quad (1)$$

Inverse Document Frequency- The log of ratio of number of documents by the total number of documents that contain the word w. The inverse document frequency computes the weights of infrequent words in the text corpus.

$$Idf(w) = \log (N / df_t) \quad (2)$$

After preprocessing the text corpus. It is necessary to represent the corpus of text in vector form so that different machine learning algorithms can be applied to it. Initially term frequency is computed which is the value showing number of times word occur in that document. This is referred as BoW Bag of words documents. Here our document is simple Tweet. Another type of feature extraction is using TFIDF vector.

A TFIDF is a vector which contains high value for most frequent term in a document and low value to least occurring term. If the term occures in all the documents, idf computed will be zero. TF-IDF is the product of term-frequency and inverse document frequency. The word which has higher tfidf score, the more significance is given to that word and vice versa [6].

$$Tf\text{-}idf(w) = tf(w) * idf(w) \quad (3)$$

E. Topic Modeling

Due to digitization in every field, huge information produced and deposited in the form of texts like news, blogs, publications, web pages, e-books, social media data like tweets. Hence, organizations, researchers, decision makers are in search of methods to search, organize, synthesize and understand this huge unstructured data.

David Blei (2012) researched that topic modeling method is a statistical method to recognize latent topics that are integral in text documents and gives improved interpretation of unstructured text data with topic labels.

Topic Modeling is a statistical machine learning algorithm that finds the theme of documents. Topics are set of commonly occurring words which represents the subject of documents [7]. Topic modeling techniques are widely used as text mining and Information Retrieval techniques. The Latent Dirichlet Allocation LDA technique works as a generative model on topic proportion on each document and models entire word corpus i.e. collection of documents. LDA model is constructed on Bag of word BOW representation of a document where each document is represented as unordered collection of words. BOW is vector representation of text documents.

Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic method, which has been mostly applied in text mining and topic extraction. LDA considers each document as a group of words and aims to discover latent topics from a distribution over words [8], [9] which is best described as shown in Fig.1.3.

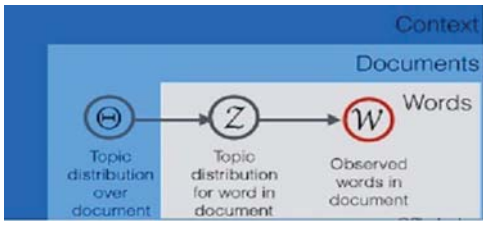

Fig. 1.3 Latent Dirichlet Allocation

Since LDA is an unsupervised learning algorithm, there is no need for manual labeling of each document. Document Term Matrix in the form of Bag of words (BOW) is generated and LDA algorithm is applied to perform text analysis of twitter data using associated words [10], [11]. Each identified topic can be associated most frequent discussions on Corona Outbreak linked on the occurrence of words in each document [12], [13].
In order to clarify results, we applied topic modeling algorithm to given dataset for many iterations 5,10,15 and 20 topics, meticulous observation and analysis indicated that 10 topics could represent coherent result and is acceptable for interpretation and analysis of the latent topics.

IV. RESULTS

After applying topic modeling to twitter data, the most probable group of words in the form of topic for entire word corpus is generated. As we know during pandemic of Covid 19 every country needs to design different policies to stop spreading of virus. Many changes are imposed on public, transportation, education, government and businesses. This has various effects on community. This can be reflected nowadays on social media network.
Social media data gives an idea of society responses regarding pandemic phase. In this work, we have worked on tweets regarding Corona Outbreak in India during March and April 2020. Results shown in following Table I gives a brief outline. A label is applied to every topic during that phase based on top words for a particular topic. Out of 10 words many words are repetitive as we extract topics based on probability distribution like discussion is related to corona or ministry policies. But some words make the topics noticeable. In every topic ministry word is repeated as most of tweets are opinion of society regarding government policies. Apart from this in every topic there are some specific words which show the theme of topics.

Topic 0 shows words indicating communal discussions, topic 1 shows corona effect in Iran as there was news regarding deputy health minister in Iran was infected due to corona. This can be most discussed topic over twitter during that period. Similarly, topic 2 related with traveling closure. Topic 3 regarding multiplex shutdown. Topic 4 is regarding handling coronavirus, and remaining topics shows wash hands, outbreak in Europe, New Delhi family discrimination case, ministry policies, and university closures. If we relate the events happened during that time period and the topics retrieved using topic modeling during the same period are similar. This indicates topic modeling applied to twitter data gives a brief idea of society replies on current situations during pandemic in India.

TABLE I Most frequent words per topic for top 10 topics

Topic	Topic Words	Topic Label
0	Deliberately, exacerbate, ministry, coronavirus, Hindus, kill, urges, statesca, extremists, accusing	Communal discussions
1	Time, fact checking, right, due, ministry, video, coronavirus, last, iran, acuteencephalitissyndrome	Outbreak in Iran
2	Close, conditions, passenger, travelers, ministry, coronavirus, group, family, help, coronavirus outbreak	Close Travelling
3	Coronavirus, right, looking, shutdowns, multiplex, ministry, office of ut, uncounted, globally	Multiplex shutdowns
4	Releasing, required, deep, fatal, know, ministry, shutdowns, miss, handling, corona, virus	Missing Deep knowledge of handling Coronavirus
5	Coronavirus, intl, seen, fight, ministry, wash, wa, shutdowns, citizen, battle	Appeal to citizens Wash Hands
6	Slight, remain, offering, utensils, right, coronavirus, universities, force, services, effectiveness	Universities remain closed
7	Coronaoutbreak, Bangladesh, crore, stayhomeindia, europe, ministry, right, coronavirusinindia, pools, gaming	Outbreak in Europe and Bangladesh
8	Coronavirus, don't, family, unprepared, newdelhi, covidrelated, spaces, widespread, discrimination, light	New Delhi Family Discrimination Case
9	Steered, st, lives, coronavirus, govts, shows, right, seems, conditions, ministry	Ministry conditions

V. PERFORMANCE MEASURE

In this work, perplexity is used as performance measure to find the effectiveness of topic modeling. The perplexity usually measures the generalization ability. This means lower the perplexity indicated the better generalization ability which suggests the high performance ability [15].
In this work we used two types of feature vectors. In this research we have implemented two topic modeling algorithms one using simple bag of words and other using TFIDF vector. Table II indicates perplexity values obtained during experimentation.

Table II Performance indication using perplexity

Feature Vector	Perplexity
TFIDF	-8.89
BOW	-7.03

This indicates TFIDF LDA model has given better performance.

VI. CONCLUSION

In this paper, topic modeling applied to twitter dataset which is considered as unstructured text corpus. This work have generated very useful topics which gives idea about public opinions during pandemic. We implemented simple Bag of word and TFIDF model for extracting topics from twitter dataset during this pandemic phase. Topic modeling has shown very promising results for overview of social media data analysis. Performance can be measured by using perplexity. Still both the models have given satisfactory results. Such systems can be used by policy makers and administrative staff for opinion analysis during Covid 19 pandemic.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Department of Computer Science, Columbia University, New York, NY 10027 USA I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[2] Kursuncu U., Gaur M., Lokala U., Thirunarayan K., Sheth A., Arpinar I.B., "Predictive Analysis on Twitter: Techniques and Applications", In: Agrawal N., Dokoohaki N., Tokdemir S., (eds) Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining. Lecture Notes in Social Networks. Springer, Cham

[3] F. Colace, M. De Santo, M. Lombardi, F. Mercorio, M. Mezzanzanica and F. Pascale, "Towards Labour Market Intelligence through Topic Modelling", Proceedings of the 52nd Hawaii International Conference on Systems Sciences, 2019.

[4] Kolini, Farzan and Janczewski, Lech," Clustering and Topic Modelling : A New Approach for Analysis of National Cyber Security Strategies" 2017, PACIS 2017 Proceedings-126.

[5] Carchiolo V., Longheu A., Malgeri M., "Using Twitter Data and Seniment Analysis to Study Diseases Dynamics", 2015, Lecture Notes in Computer Science, vol 9267. Springer, Cham, Online ISBN 978-3-319-22741-2

[6] Li-Ping Jing, Hou-Kuan Huang and Hong-Bo Shi, "Improved Feature Selection Approach TFIDF in Text Mining," Proceedings. International Conference on Machine Learning and Cybernetics, Beijing, China, 2002, pp. 944-946 vol.2, doi: 10.1109/ICMLC.2002.1174522.

[7] David M. Blei , Andrew, Michael I. Jordan,"Latent Dirichlet Allocation"Journal of Machine Learning Research, vol. 3, pp 993-1022,(2003)

[8] James O' Neill, Cecile Robin, Leona O' Brien,PaulBuitelaar,"An Analysis of Topic Modeling for Legislative Texts, In: Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2017), London, UK.(2017)

[9] Shiliang Sun , Chen Luo, Junyu Chen, "A Review of Natural Language Processing Techniques for Opinion Mining Systems", Information Fusion,Volume 36, July 2017,pp10-15Elsevier (2017)

[10] J.W. Uys, N.D. du Preez, E.W. Uys,"Leveraging Unstructured Information Using Topic Modeling, ", PICMET 2008 Proceedings,,pp 955-961, 27-31 July, Cape Town, South Africa (c),(2008)

[11] NamukKo, ByeongkiJeong, Sungchul Choi and Janghyeok Yoon, "Identifying Product Opportunities Using Social Media Mining: Application of Topic Modeling and Chance Discovery Theory", 2169-3536 © 2017 IEEE, DOI10.1109/ACCESS.(2017).

[12] Jen-Tzung Chien,"Hierarchical Theme and Topic Modeling", IEEE Transactions on Neural Networks and Learning System,Volume: 27, Issue: 3, pp 565 – 578, (2016)

[13] Chakkarwar V., Tamane S.C. (2020) "Quick insight of Research Literature using Topic Modeling", In: Zhang Y D., Mandal J. So-In C.,Thakur N.(eds) *Smart Trends in Computing and Communications Smart Innovation, System and Technologies,* Vol 165.,Springer, Singapore

[14] Hanqi Wang, Fei Wu, Weiming Lu, Yi Yang, Xi Li, Xuelong Li, *Fellow, IEEE*, and Yueting Zhuang, "Identifying Objective and Subjective words via topic modeling", IEEE transactions on Neural Network and Learning System, Digital Object Identifier 10.1109/TNNLS.2016.2626379

[15] Phand S. and Chakkarwar V. A., "Enhanced Sentiment Classification Using Geo Location Tweets," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, 2018, pp. 881-886, doi: 10.1109/ICICCT.2018.8473048.