ISSN 2347 - 3983
Volume 9. No. 8, August 2021
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter26982021.pdf
https://doi.org/10.30534/ijeter/2021/26982021

A brief survey of Text Mining and its applications

Shruti Chandrayan[1], Priyanka Bamne[2]
Master of Engineering Student, Dept. OF Computer Engineering S.G.S.I.T.S, Indore India,
shruti20.chandrayan@gmail.com
Assistant Prof. Department of Computer Engineering, S.G.S.I.T.S, Indore, India,
bamnepriyanka574@gmail.com

ABSTRACT

The world came across the worst pandemic of all times in the year 2020 due to the outburst of Severe Acute Respiratory Syndrome Coronavirus-2 or Covid-19.
All the questions about this outbreak were piled up and research was fast growing [1]. A study showing that in precisely just six months, substantial databases have been swamped with research articles, news, notes, and editorial related to coronavirus. It estimates that 23,634 distinctly published articles have been indexed on Web of Science and Scopus between 1 January and 30 June 2020. Imagine the data that is with us today!! Approximately 200,000 scholarly articles have been published related to Covid-19. This tells us that there is a need for simplifying search results to get answers to high priority questions for users specifically scientists. Currently, document clustering tools are being used in many areas. A similar clustering tool can be made particularly for Covid-19 which will help scientists and researchers get answers to high priority questions about this pandemic. In this paper, we are discussing about the process of text mining, text categorization and, text clustering. Also, a comparison of the algorithms used for clustering particularly in text data.
Key words: Text Mining, Text Categorization, Text Clustering.

1. INTRODUCTION

Our lives are different since the pandemic hit in year 2020. People are more health conscious. People are focusing on health and fitness more than ever specially on how to increase immunity and be optimistic in this difficult situation. We studied about how there was a change in people's search pattern during this time and we came across that Covid-19 is a huge topic in search of interest.
We can see a detailed study of how searches through Google trends have changed the search pattern during the pandemic. Below are some charts for details about the search trends. (Source: Google Trends)

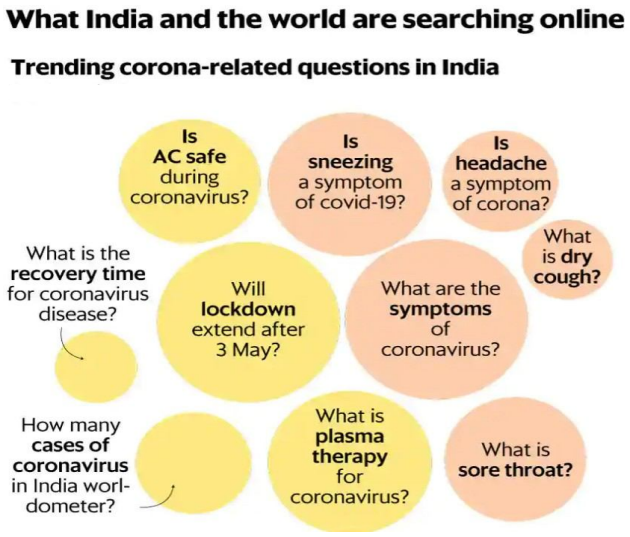

Figure 1: What are people searching online?

Figure 2: What are people searching online?

Through this study we observed that as the pandemic has spread, along with it fear and uncertainties have also prevailed.Fig. 1 and 2 tells us how people have turned to Dr. Google for answers to questions like "What are the symptoms of coronavirus?". Others have turned for questions like" H o w to", for example" How to use google meet?" and so on. Globally, search for coronavirus has

peaked five days after WHO declared it a pandemic. In India, it peaked just after the lockdown. Currently, the questions related to vaccine continues worldwide. We can see that there is a need for simplifying oursearch which will be helpful for scientists in getting answers from huge amount of data that has piled overtime since the pandemic. This task can be done using Natural language processing and machine learning techniques. Particularly text mining is a field which helps us in analyzing data which isin text form [2]. Text mining is a field that uses text data for discovering and mining knowledge [3]. Text data is a valuable source which includes news articles, blogs, product reviews and tweets. Researchers use text mining approaches to organize and deal with large sets of text data to determine various ideas like sentiment analysis and in fields such as bioinformatics [4].

It is necessary to understand the difference between data mining [5] and text mining. Both are concepts related to data analysis. The only distinct point is the type of data that we are dealing with. Data mining[34] is the process of analyzing vast data set to extract meaningful patterns [6]. It deals with numerical data. For example, a data set about finance, height, weight etc. Majorly it depends upon the statistical techniques and algorithms. Text mining is the process of discovering information from unstructured data by turning it into valuable structured information. It depends upon the lexical and syntax structure of the data. Text mining is formed on statistical techniques and linguistic techniques such as Natural Language Processing [7]. Example of text data set is a collection of twitter data, researcharticles, editorials etc. Text mining tools usually face many technical challenges due to diverse document formats liketext documents, emails, social media posts. Also, it requires dealing with multilingual texts, abbreviations, and different slang. These are the major challenges which make text mining different, difficult yet interesting to work.

With the rise of social networks and increased connectivity, huge interest has been recorded in this field. Fieldslike sentiment analysis, text clustering, text categorization, document summarization are in trend today. Text mining toolsare becoming more noteworthy. Text analysis as a whole is anemerging field of study.

Here, we will be introducing the above fields:

- SENTIMENT ANALYSIS Also known as opinion mining. As the name suggests, this field detects positive or negative sentiments in text. For example, during Covid- 19, this field can be used in a way to understand how many people were in depression through their social media posts, which will help in preventing suicidal nature.
  [8] We can predict mood during the pandemic through analyzing tweets. [9] Even humans struggle to analyze sentiments among themselves accurately, making sentiment analysis one of the hardest tasks in natural language processing but on the other hand it is a great field of research.
- TEXT CLUSTERING Also known as document clustering. This field uses machine learning and natural language processing to understand unstructured textual data. Google's search engine is the best example to under- stand text clustering. For example, people are constantly searching with keywords like immunity, symptoms, rate of spread, sore throat and so on since the previous year. Google gives us the pages that apply to that term, but how Google does that? It's because of text clustering.
- TEXT CATEGORIZATION Also known as Text classification. We have been using text classification to simplify things for a while now. This field is essential for text analysis. It is the process of labeling natural language texts with different categories in a prearranged set. [10] For example, this field can be used to classify news articles or research articles containing information about Covid-19.
- DOCUMENT SUMMARIZATION As the name suggests, text summarization creates a brief, precise, and vivid summary of a long text document. Automatic text summarization is a trendy field of interest in machine learning and natural language processing. Due to the shortage of time as well as busy schedules nowadays, people prefer reading summaries rather than going for the substantial amount of data circulating in the digital extent. Applying this technique in Covid-19 literature will facilitate the process of studying for information.

2. MOTIVATION

Analyzing text data requires more than using conventional machine learning models [11]. All the machine learning methods deal with numbers. So, text data is transformed into a form that these algorithms can handle. Also, text analysis can be used in various fields in a way to save our time in the busy and tight schedules. Current scenario requires a time saving search method with respect to the pandemic. Analyzing the text data related to Covid-19 will boost up the research process. Analyzing text using machine learning combined withvisualization will be useful for better understanding and good search results. This motivated us to study about text mining.

3. LITERATURE REVIEW

As discussed, earlier handling text data is quite different from normal data. Previous searches about handling text

data and how to analyze it using various algorithms has been outlined in this section.

In the late 1990's, researchers started to utilize text as data, which gave rise to text mining [12]. Text mining came into attention when people were trying to assemble or classify documents (Cutting, 1992). The objective is not so much about discovering about new patterns but cleaning the data stored in databases. Mining basically refers to clean a valuable item and in case of text mining it is separating valuable keywords from a mass of other words and using them to identify or extract knowledge. [13]

With time there has been an enormous need to design methods, tools and algorithms through which we can effectively process text applications. Few of such algorithms which are regular in text domain and focus on mining methods are introduced here. [14] Text mining deals with information from discrete written resources. The goal is to discover unknown information. The dissimilarity between regular data mining and text mining is that in text mining patterns are known by analyzing natural language text alternative to the data we usually come across with. [15]

As algorithms and methods came into practice, techniques to extract patterns came into existence. Computers were taught how to analyze, recognize and produce information through text. Technologies were put together by natural language processing. Technologies like information extraction, categorization, clustering, visualization and summarization [16] are widely used today. [17] [18]

The very first step towards mining data is preprocessing. Three steps that are performed in preprocessing are stop word removal, stemming and TF/IDF algorithms. Stop words are thewords that do not add meaning in a document. Example, a, an, the etc. Stemming is done to identify the foundation of a word. This step removes the suffixes, to reduce the number of words which conserves time and memory. TF/IDF algorithms reveals how important a word is to a document in a collection.[19]

Once our data is cleaned and we are rid of noises we can apply any suitable technique for text analysis. [20] In this paper, author has briefed various algorithms for text clustering. A fine clustering of text requires efficacious preprocessing and apt choice of algorithm for the task. Distance based methods are the best and most favored among all.

Another technique which plays a principal role in text domain is Topic modelling. A topic is basically a cluster of words that occur jointly. Topic modelling links words with homogeneous meaning. There are four methods considerable under this technique. These methods are Latent Semantic analysis (LSA) [21], Probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA). [22] We have focused on Latent Dirichlet Allocation in this paper since it is widely used with text data. [23] LDA first introduced by Blei, Ng and Jordan in 2003 [12], is one of the most admired method in topic modeling.

4. METHODOLOGY

Following the methodologies and techniques for text analysis, we can make the search easy and more feasible to get better results.

4.1 Preprocessing

As discussed above that preprocessing is a crucial step in text mining, natural language processing and information retrieval. Preprocessing in simple words is cleaning our data, making it ready for analysis. It includes tokenization, stop word removal, stemming. Let's discuss the above-mentioned steps!

1) Tokenization*:* Tokenization is the process of disintegrating a stream of text into words, phrases, symbols which are nothing but tokens. These tokens are input for advance processing such as parsing or mining. Another aim of this process is identifying the meaningful keywords. [24]

2) Stop word removal: There are many words in English language that are used frequently but have no meaning. Like connectors in a sentence (and, or, because, it, at etc.) These words are useless in text analysis. This process improves the system performance. Hence, removing these words is necessary.

3) Stemming*:* Stemming is the process of producing morphological or similar variants of a base word. For example, the words:" presentation"," presented"," presenting" can be reduced to a common presentation" present". Tokens are used as input in this step. [24]

Overall, these preprocessing techniques put an end to noise in our data and pinpoints the root word for existent words lastly, minimizing the size of text data.

4.2 Vectorization

Vectorization is the process of converting text data in a form which machine learning algorithms can understand. Since machine learning algorithms deal with mostly numbers, vectorization converts text into vectors and voila!! Now text are numbers. Few algorithms for this process are discussed below:

1) TF-IDF: TF-IDF stands for term frequency-inverse document frequency. This algorithm studies the relevance of key- words to document in corpus. It anticipates the keywords using which some specified documents can be identified or categorized. For instance, if a businessman hires an internee whose main task is to add on new tweets to his personal twitter handle on a daily basis. However, the internee is not able to take care of tags which a common problem due to which tweets

are not categorized well. TF-IDF algorithm is of great use here, as it can point out the tags automatically. It will reserve enough of time for the internee. This algorithm is a blend of two words- Term frequency (TF) and Inverse document frequency (IDF). TF is used to estimate how many times a term exists in a document. Example if we have a document "T1" containing 3000 words and the word "Covid" is present in the document exactly 6 times. The term frequency of the word "Covid" in the document "T1" will be TF = 6/3000 = 0.0002 IDF assigns weights according to the occurrence of words. Lower weight is assigned to recurring words and greater weight for rare words. For example, we have 20 documents and the term "Lockdown" is present in 15 of those documents, so the inverse document frequency can be calculated as IDF = log e (20/15) = 1.106 [25] There are other algorithms which can be used in place of TF-IDF. We have discussed few of them below:

2) Word2Vec*:* This is used for word embedding which are vector representations of a specific word. Word2Vec uses shallow neural network. Loosely speaking, it is a two-layer neural net that processes text by vectorizing words. The input is a text corpus and product is a set of vectors, feature vectors that represent words in that corpus. It also turns text into a numerical form that deep neural networks can understand. Word2vec is prominent for breaking documents and identifying content and subsets of content.

3) Bag of Words*:* This version is a general way of representing text data when modeling text with machine learning. A set of vectors is created containing the count of word occurring in the document. This model faces a problem when we come across new sentences. TFIDF model consists of information on more important words and less important as well. TFIDF performs better in machine learning models.

All of the above algorithms make our text data in vector form so that machine learning algorithms can easily use the data in form of numbers.

4.3 Text Clustering

The clustering problem is defined as locating groups of alike objects in data. Resemblance between the objects is measured using a similarity function. In text analysis, clustering is appropriate for arranging documents to magnify retrieval and support browsing. [26] Text clustering algorithms [27] are split up in a wide variety of types such as agglomerative clustering problems, partitioning algorithms and some standard algorithms such as EM algorithm. [20] k-means algorithm [28] is one of the most broadly used algorithm which is effortlessly simple and systematic [29]. Few algorithms are discussed below:

1)K-means*:* The k-means clustering technique is very simple. Let's begin with a description of the basic algorithm. K-means algorithm can be defined as a sequential algorithm where the dataset is partitioned into some pre-defined distinct non-overlapping clusters 'k'.Each data point belongs to only one class or group. Basic K-means Algorithm for finding K clusters: a) Selecting K points as the initial centroids. b) Assigning all points to the closest centroid. c) Recomputing the centroid of each cluster. d) Repeating step 3 and 4 until centroids don't change.

*2)*BIRCH*:* Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a hierarchical based clustering algorithm. This algorithm needs a termination condition, ratherthan the number of clusters k as input. This algorithm pre- serves as much information as possible by generating a com- pact summary of large dataset. BIRCH is structured clusteringfor huge databases and it browses the database at one go. Itis not sensitive with respect to noise, but it takes more time compared to K-means. [30]

3)CURE: Clustering Using Representatives is also a hierarchical based clustering technique. It selects well-scattered points from the cluster and then reduces them in the center of the cluster by a specified fraction.

4.4 Topic Modelling

Topic modelling methods are used for organizing, under- standing, searching and summarizing large text data spontaneously. The topics are present but invisible and are to be evaluated with respect to the link words in the corpus and their occurrence in the documents. Topic models uncover the veiled themes throughout the collection and elucidate the documents according to those themes. Automatic topic modelling is another exiting work [31]. Ultimately, a documentcoverage spread of topics is created and it provides a new means to search the data on the perspective of topics. [32]Few topic modelling algorithms are discussed below:

1) LSA: Latent Semantic Analysis, was developed for the task of Information Retrieval, which involves, selecting a few relevant documents which match a given query from a large database of documents. It was one of the primary algorithms used for topic modelling. However, it lacks interpretable em- beddings and representation is less efficient for large datasets.

2) PLSA*:* Probabilistic Latent Semantic Analysis, the key idea is to discover a probabilistic model with hidden topicsthat can bring about the data we notice in our document-term matrix. PLSA just connects a probabilistic treatment of topics and words on top of LSA. LDA which is discussed next is an improved and extended version of PLSA.

3) LDA*:* Latent Dirichlet Allocation (LDA) [33]. LDA is a Bayesian version of PLSA The word 'LATENT' means the model discovers hidden topics from the documents. 'DIRICH- LET' is a distribution which this algorithm assumes for distribution of topics in a document and distribution of words in a topic. 'ALLOCATION' indicates the distribution of topics in the document. LDA overall works best because it can generalize to new documents easily.

5. CONCLUSION

This paper is a discussion about how text mining and the relevant techniques can be used to develop advanced computational tools. The White House, for example, has requested data scientists to develop tools through which CORD-19 data set can be analyzed. We have provided a summary of some of the most elementary algorithms and techniques used broadly in text domain. We have also discussed how text mining techniques can be applied in the current COVID-19 situations.Even though, it is not possible to detail all separate methods and algorithms in their profound form regarding the limitsof this article, it should give a rough run-through of ongoing advancements in the field of text mining.

Text mining is crucial for scientific research given the very high capacity of research-based literature being generated every year. These wide-ranging records of online scientific articles are growing notably as a great deal of new articles are added on a regular basis. While this development has made access to the scientific information for authorized researchers easier, it has also become hard for them to recognize articles which are more relevant to their interests. Thus, processing and mining this enormous amount of text is a great attraction for researchers.

REFERENCES

[1] J. Teixeira da Silva, P. Tsigaris, and M. Erfanmanesh, "Publishing volumes in major databases related to covid-19," *Scientometrics*, Jan. 2020.

[2] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[3] C. Faloutsos and D. W. Oard, "A survey of information retrieval and filtering methods," tech. rep., 1998.

[4] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57–71, 2005.

[5] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and data Engineering*, vol. 8, no. 6, pp. 866–883, 1996.

[6] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *et al.*, "Knowledge discovery and data mining: Towards a unifying framework.," in *KDD*, vol. 96, pp. 82–88, 1996.

[7] E. D. Liddy, "Natural language processing," 2001.

[8] S. Sharma and S. Sharma, "Analyzing the depression and suicidal tendencies of people affected by covid-19's lockdown using sentiment analysis on social networking websites," *Journal of Statistics and Management Systems*, 12 2020.

[9] A. S. M. Venigalla, S. Chimalakonda, and D. Vagavolu, *Mood of India During Covid-19 - An Interactive Web Portal Based on Emotion Analysis of Twitter Data*. New York, NY, USA: Association for Computing Machinery, 2020.

[10] K. Chatsiou, "Text classification of covid-19 press briefings using bert and convolutional neural networks," *ArXiv*, vol. abs/2010.10267, 2020.

[11] M. Kantardzic, *Data mining: concepts, models, methods, and algo-rithms*. John Wiley & Sons, 2011.

[12] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (kdt).," in *KDD*, vol. 95, pp. 112–117, 1995.

[13] V. Kotu and B. Deshpande, "Chapter 9 - text mining," in *Data Science (Second Edition)* (V. Kotu and B. Deshpande, eds.), pp. 281–305, Morgan Kaufmann, second edition ed., 2019.

[14] A. Hotho, A. N¨urnberger, and G. Paaß, "A brief survey of text mining.,"
in Ldv Forum, vol. 20, pp. 19–62, Citeseer, 2005.

[15] M. Hearst, "What is text mining," SIMS, UC Berkeley, vol. 5, 2003.

[16] P. Bhatia, "A survey to automatic summarization techniques,"

[17] S. V. Gaikwad, A. Chaugule, and P. Patil, "Text mining methods and techniques," International Journal of Computer Applications, vol. 85, no. 17, 2014.

[18] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," arXiv preprint arXiv:1707.02919, 2017.

[19] S. Vijayarani, M. J. Ilamathi, M. Nithya, et al., "Preprocessing techniques for text mining-an overview," International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pp. 7–16, 2015.

[20] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in Mining text data, pp. 77–128, Springer, 2012.

[21] S. T. Dumais et al., "Latent semantic indexing (lsi): Trec-3 report," Nist Special Publication SP, pp. 219–219, 1995.

[22] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," Int. J. Adv. Comput. Sci. Appl.(IJACSA), vol. 6, no. 1, 2015.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation.," in NIPS, vol. 1, pp. 601–608, 2001.

[24] S. Kannan, V. Gurusamy, S. Vijayarani, J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining," International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pp. 7–16, 2014.

[25] S. Qaiser and R. Ali, "Text mining: use of tf-idf to examine the relevance of words to documents," International Journal of Computer Applications, vol. 181, no. 1, pp. 25–29, 2018.

[26] P. G. Anick and S. Vaithyanathan, "Exploiting clustering and phrases for context-based information retrieval," in Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '97, (New York, NY, USA), p. 314–323,
Association for Computing Machinery, 1997.
[27] A. K. Jain and R. C. Dubes, Algorithms for clustering data. Prentice- Hall, Inc., 1988.
[28] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering.," in ICML, vol. 98, pp. 91–99, Citeseer, 1998.
[29] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," 1997.
[30] M. Gupta and A. Rajavat, "Comparison of algorithms for document clustering,"in 2014 International Conference on Computational Intelligence and Communication Networks, pp. 541–545, IEEE, 2014.
[31] M. Allahyari and K. Kochut, "Automatic topic labeling using ontologybased topic models," in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 259–264, IEEE, 2015.
[32] Z. Tong and H. Zhang, "A text mining research based on lda topic modelling," in International Conference on Computer Science, Engineering and Information Technology, pp. 201–210, 2016.
[33] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.
[34] A. B. Annasaheb and V. K. Verma, "Data mining classification techniques: A recent survey," International Journal of Emerging Technologies in Engineering Research, vol. 4, no. 8, pp. 51–54, 2016.