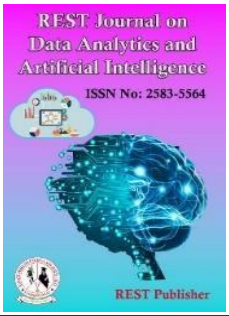
REST Journal on Data Analytics and Artificial Intelligence
Vol: 2(3), September 2023
REST Publisher; ISSN: 2583-5564

Website: http: //restpublisher.com/journals/jdaai/

DOI: https://doi.org/10.46632/jdaai/2/3/8



An Analysis of Text Mining in Big Data

*** [1,2] Uthaman.V**
Adaikalamatha College, Thanjavur, Tamil Nadu, India.
Annai Vailankanni Arts and Science College, Thanjavur, Tamil Nadu, India.
*Corresponding Author Email: **uthamanv@gmail.com**

Abstract: *The practice of extracting hidden predictive information from a database and structuring it for later use is known as data mining. Web mining, text mining, sequence mining, graph mining, temporal data mining, spatial data mining, distributed data mining, and multimedia mining are some of the diverse domains in data mining. Financial data analysis, the retail and telecommunications sectors, science and engineering, and intrusion detection and prevention are a few fields where data mining is used. Worldwide, big data analytics is a burgeoning research field in computer science and many other fields. We covered text mining methods and their uses in this paper. Text mining is becoming a significant topic of study. Text mining is the process of automatically finding new, previously undiscovered information using a computer.*
Keywords: *Big Data, Data Mining, Text Mining, Natural Language Processing.*

1. INTRODUCTION

Big Data: Big data is a greater volume of unstructured and structured data that is too large that it is hard to handle with traditional techniques and tools. Big data analytics means that to examine the data produced by big data sources to upgrade their businesses or studies. The primary goal of big data analytics manages the too large volume, velocity, and variety using various techniques based on intelligence (called Machine Learning). Big data analytics is applied to analyze the data presented by various big data platforms. The term big data which describes extremely large data sets is widely being used among different researchers all over the world. Traditional relational databases are not capable of handling big data.

Big Data Analytics: The term big data which describes extremely large datasets is widely being used among different researchers all over the world. Traditional relational databases are not capable of handling big data. An enormous quantity of data sets arrives from several sources like sensors, transaction applications, the web, and social media, etc. The big data phenomenon can be comprehended clearly by knowing the differences associated with them such as Volume, Velocity, Variety, Veracity, and Value. Big data analytics is a budding research area that deals with the collection, storage, and analysis of immense data sets to trace unknown patterns and other key information. Big data analytics helps us to recognize the data that are an integral component of future business decisions. Big data analytics can be abundantly found in domains such as the banking and insurance sector, healthcare, education, social media and entertainment industry, bioinformatics applications, spatial applications, agriculture, etc. It is a herculean task to handle big data using conventional data processing applications. Thus, to discover hidden data patterns, trends, and associations, intelligent machine learning methods can be adapted. The objective of the current research paper is to discuss various machine learning algorithms used by data scientists for analyzing and modeling big data.

Data mining: Data mining is the process of improving and arranging the large data set analysis to identify the rule patterns and create the relationship to solve the problems through the data analysis. Data mining is the continuous process of obtaining the records 'from the training session data through heuristics detection. Data Mining is the extraction of most hidden predictive information from large heterogeneous data warehouses. It is a powerful technology with great potential to motivate the companies that deal on the most vital information in their data warehouses. The tools used for data mining predict the future trends and behaviors, allowing the businesses to build proactive and knowledge-driven decisions. The most automated and forthcoming analysis offered by data mining is in motion away from the analysis of what went before events provide by showing tools typical of decision support systems. The tools can able to perform the business questions that traditionally were

too time consuming to obtain the solutions.

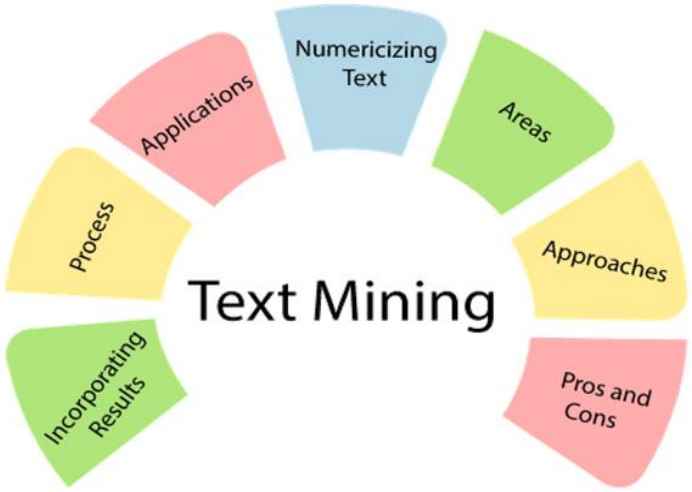


FIGURE 1. Process of Text Mining

Text Mining: Text Mining is also called as text data mining, which is used to find interesting information from large database. The data mining tools can be designed to handle structured data from database. But the text mining can be work with unstructured data or semi-structured data. Text Mining is to analyze large quantities of natural language text and it detects lexical patterns to extract useful information. Text Mining is useful for organization because most of the information is in text format. The following steps can be included in text mining.

- It converts the unstructured text into structured data
- Identify the patterns from structured data
- Analyze the patterns using Text Mining techniques
- Extract the useful information from the text

The applications of Text Mining are protein interaction, drug discovery, predictive toxicology, identification of recent product potentialities, detection of links between lifestyle and states of health, competitive intelligence, and lots of additional.

Literature Survey: **Yuefeng Li et al [13]:** A Text mining and classification method has been used for term-based approaches. The problems of polysemy and synonymy are the major issues. There was a hypothesis that pattern-based methods should outperform best compare to the term-based ones in describing user preferences. A large-scale pattern remains a hard problem in text mining. The state-of-the-art term-based methods and the pattern-based methods in the proposed model performs efficiently. In this work, a clustering algorithm is used. Relevance feature discovery based on both positive and negative feedback for text mining models.

Jian ma et al [4]: The author focused on the problem by classifying text documents axiomatically, for the most part in English. When working with non-English language texts it leads to the forbiddance. Ontology-based text mining approach has been used. It's efficient and effective for clustering research proposals encapsulated with English and Chinese texts using a SOM algorithm. This method can be expanded to help in searching for a better match between proposals and reviewers.

Chien-Liang Liu et al [2]: The paper concluded that the information about the movie rating is based on the result of sentiment classification. The feature-based summarizations are used to generate condensed descriptions of movie reviews. The author designed a latent semantic analysis (LSA) to establish product features. It is a way to reduce the size of the summary from LSA. They account for both accuracy of sentiment classification and the response time of a system to design the system by using a clustering algorithm. OpenNLP2 tool is used for implementation.

Yue Hu et al [19]: PPSGen is a new system that was proposed to solicitation of the presentation slides generated and can be used as drafts. It helps them to prepare the formal slides in a faster way for the proprietor. PPSGen system can bring out slides with better quality suggested by the author. The system was developed by the Hierarchical agglomeration algorithm. Tools are a Microsoft Power- Point and OpenOffice. A 200 combo of papers and slides are taken as tests set from the web demonstrate for evaluation process. PPSGen is comparably better than the baseline methods that were evident by the user study.

Xiuzhen Zhang et al [10]: The problem faced by all the reputation systems is concentrated by the author. However, the reputation scores are universally high for sellers. It is a situation requiring great effort for promising buyers to select trustworthy sellers. The author proposed CommTrust for trust evaluation by feedback comments

through mining. A multidimensional trust model is used for computation jobs. Data sets are collected from eBay, amazon. This technique used a Lexical-LDA algorithm. CommTrust can effectively address the good reputation problem issue and rank sellers are finally by showing definitely through extensive experiments on eBay and Amazon data.

Dnyanesh G. Rajpathak et al [9]: The challenging task is In-time augmentation of the D-matrix through the finding of new symptoms and failure modes. The proposed strategy is to construct the fault diagnosis ontology to abide by concepts and relationships frequently observed in the fault diagnosis domain. The needed artifacts and their dependencies from the unstructured repair verbatim text were found out by the ontology. Real-life data collected from the automobile domain. Text mining algorithms are used. To establish automatically the D-matrices by the unstructured repair verbatim data that was mined done by the ontology-based text mining composed while faulting diagnosis. Graph and graph comparison algorithms have to be generated for each D-matrix.

Jehoshua Eliashberg et al [11]: To forecast the box office performance of a movie at the crenulation point, it's suitable only if it holds the script and production cost. They extract textual features in three levels particularly genre and content, semantics, and bag-of- words from scripts using domain knowledge of screenwriting, input given by human, and natural language processing techniques. A kernel-based approach is to assess box office performance. Data set are collected from 300 movie shooting scripts. The proposed methodology predicts box office income more exactly 29 percent is reduced mean squared error (MSE) compared to benchmark methods.

Donald E. **Brown et al [17]:** Rail accidents present image of a valuable safety point for the transportation industry in many countries. The Federal Railroad Administration needs the railroads muddled in accidents to submit reports. The report has to be cuddled with default field entries and narratives. A combination of techniques is to automatically discover accident characteristics that can inform a better understanding of the patron to the accidents. Forest algorithm has been used. Text mining looks at ways to extract features from text that takes advantage of language characteristics particular to the rail transport industry.

Luís Filipe da Cruz Nassif et al [6]: In forensic analysis that was computerized with millions of files is usually examined. Unstructured text was found in most of the files performing analyzing process is highly challenging revealed by computer examiners. Document clustering algorithms for the analysis of computers on forensic department seized in police an investigation which was suggested by the author. Variety of mixture of parameters that leads to prompt of 16 different algorithms consider for evaluation. K-means, K-medoids, Single, Complete and Average Link, CSPA are the clustering algorithm are used. Clustering algorithms motivate to induce clusters formed by either relevant or irrelevant document which is used to enhance the expert examiner's job.

Charu C. Aggarwal et al [5]: Author focused on the Use of Side Information for Mining Text Data. an effective clustering approach was done by the classical partitioning algorithm with probabilistic models which was designed by the author. Dataset used is CORA, DBLP-four-area data set and IMDB. Running time and number of clusters are used as a parameter for analyzing purpose. The results can evident that the usage of side-information can improve the quality of text clustering and classification to sustain a high level of efficiency.

2. METHODOLOGY

Text Mining Techniques

Data Mining: DM is the process of identifying patterns in large sets of data. The aim is to uncover previously unknown, useful knowledge. When used in text mining, DM is applied to the facts generated by the information extraction phase. We put the results of our DM process into another database that can be queried by the end-user via a suitable graphical interface. The data generated by such queries can also be represented visually. Data mining can be loosely described as looking for patterns in data. It can be more fully characterized as the extraction of hidden, previously unknown, and useful information from data. Data mining tools can predict behaviors and future trends, allowing businesses to make positive, knowledge-based decisions. Data mining tools can answer business questions that have traditionally been too time consuming to resolve. They search databases for hidden and unknown patterns, finding critical information that experts may miss because it lies outside their expectations. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Information Retrieval: IR systems find the documents in a collection that match a user's query. The most well-known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. In order to achieve this goal statistical measures and methods are used for the automatic processing of text data and comparison to the given question. Information retrieval in the broader sense deals with the entire range of information processing, from data retrieval to knowledge retrieval. Although information retrieval is a relatively old research area where first attempts for automatic indexing were made in 1975, it gained increased attention with the rise of the World Wide Web and the need for sophisticated search engines. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally-intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of

documents for analysis. For example, if we are interested in mining information only about protein interactions, we might restrict our analysis to documents that contain the name of a protein, or some form of the verb 'to interact' or one of its synonyms.

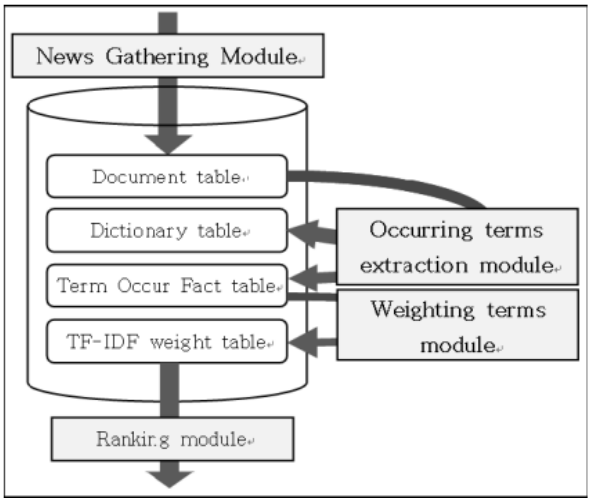


FIGURE 2. Keyword extraction in Text Mining

Information Extraction: Information Extraction A starting point for computers to analyze unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process called pattern matching. The software infers the relationships between all the identified people, places, and time to provide the user with meaningful information. This technology can be very useful when dealing with large volumes of text. Traditional data mining assumes that the information to be "mined" is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural language documents rather than structured databases. Since IE addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by an IE module can be provided to the KDD module for further mining of knowledge. After mining knowledge from extracted data, it can predict information missed by the previous extraction using discovered rules.

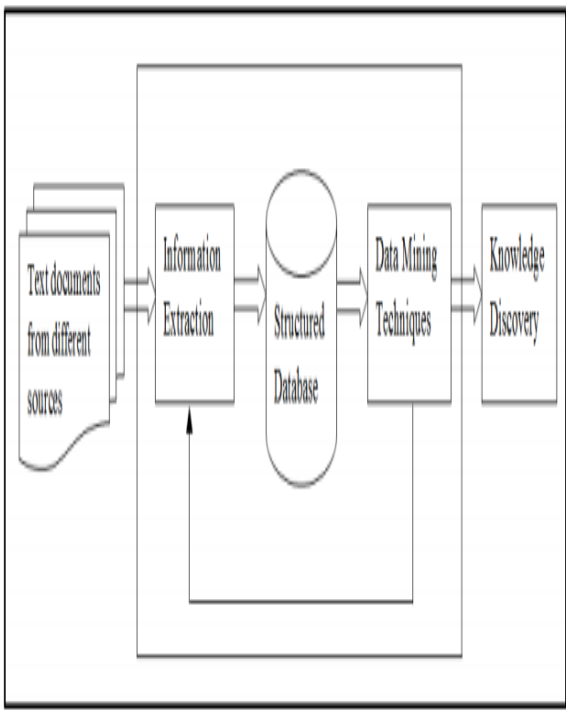


FIGURE 3. Information Extraction

Natural language processing: NLP is one of the oldest and most difficult problems in the field of artificial intelligence. Text and data mining approach has been used in Natural language processing to overcome the difficulties with the codes, keywords, and search techniques involved in knowledge or pattern discovery. The output of this mining process helps the analyst to discover the trends and new occurrences in the data available.

Many TM algorithms have been developed to maintain the text sources of the bilingual corpus or multi-lingual corpus. The general goal of NLP is to achieve a better understanding of natural language through the use of computers. Others include also the employment of simple and durable techniques for the fast processing of text, as they are presented. The range of the assigned techniques reaches from the simple manipulation of strings to the automatic processing of natural language inquiries. In addition, linguistic analysis techniques are used among other things for the processing of text. It is the analysis of human language so that computers can understand natural languages as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence. The role of NLP in text mining is to provide the systems in the information extraction phase with linguistic data that they need to perform their task. Often this is done by annotating documents with information like sentence boundaries, part of- speech tags, parsing results, which can then be read by the information extraction tools. Natural Language processing, Text mining and Machine learning methods can be applied for mining of interesting data available online for example gene ontology or protein function mapping using certain tools.

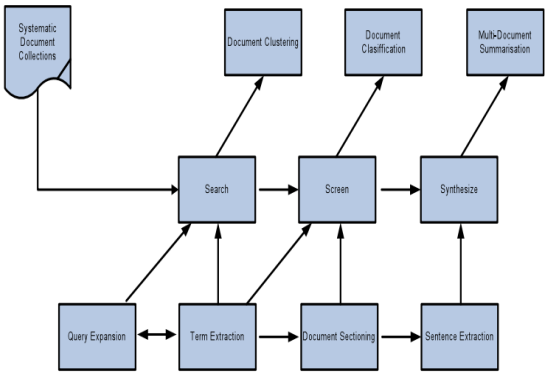


FIGURE 4. Work flow

3. CONCLUSION

The practice of extracting hidden predictive information from a database and structuring it for later use is known as data mining. Web mining, text mining, sequence mining, graph mining, temporal data mining, spatial data mining, distributed data mining, and multimedia mining are some of the diverse domains in data mining. Financial data analysis, the retail and telecommunications sectors, science and engineering, and intrusion detection and prevention are a few fields where data mining is used. Worldwide, big data analytics is a burgeoning research field in computer science and many other fields. We covered text mining methods and their uses in this paper. Text mining is becoming a significant topic of study. Text mining is the process of automatically finding new, previously undiscovered information using a computer.

REFERENCES

[1]. Luis Tari, Phan Huy Tu,Jorg Hakenberg, Yi Chen, Tran Cao Son, Graciela Gonzalez, And Chitta Baral,"Incremental Information Extraction Using Relational Databases", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2019.

[2]. Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, And Emery Jou," Movie Rating and Review Summarizationin Mobile Environment", IEEE Transactions on Systems, Man, And Cybernetics—Part C: Applications and Reviews, Vol. 42, No. 3, May 2019.

[3]. Fuzhen Zhuang, Ping Luo, Zhiyong Shen, Qing He, Yuhong Xiong, Zhongzhi Shi, And Hui Xiong, "Mining Distinction and Commonality Across Multiple Domains Using Generative Model for Text Classification" IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 11, November 2019.

[4]. Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu," An OntologyBased Text-Mining Method to Cluster Proposals for Research Project Selection", IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2019.

[5]. Charu C. Aggarwal, Yuchen Zhao, And Philip S. Yu, "On the Use of Side Information for Mining Text Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 6, June 2020.

[6]. Luís Filipe Da Cruz Nassif and Eduardo Raul Hruschka," Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection" IEEE Transactions on Information Forensics and Security, Vol. 8, No. 1, January 2019.

[7]. Bo Chen, Wai Lam, Ivor W. Tsang, And Tak-Lam Wong, "Discovering Low-Rank Shared Concept Space

for Adapting Text Mining Models" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 6, June 2019.

[8]. Francisco Moraes Oliveira-Neto, Lee D. Han, And Myong Kee Jeong." An Online Self-Learning Algorithm for License Plate Matching", IEEE Transactions on Intelligent Transportation Systems, Vol. 14, No. 4, December 2019.

[9]. Dnyanesh G. Rajpathak and Satnam Singh," An Ontology-Based Text Mining Method to Develop D-Matrix from Unstructured Text", IEEE Transactions on Systems, Man, And Cybernetics: Systems, Vol. 44, No. 7, July 2020.

[10]. Xiuzhen Zhang, Lishan Cui, And Yan Wang, "Commtrust: Computing Multi-Dimensional Trust by Mining E-Commerce Feedback Comments", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 7, July 2020.

[11]. Jehoshua Eliashberg, Sam K. Hui, And Z. John Zhang," Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 11, November 2020.

[12]. Riccardo Scandariato, James Walden, Aram Hovsepyan, And Wouter Joosen," Predicting Vulnerable Software Components Via Text Mining", IEEE Transactions on Software Engineering, Vol. 40, No. 10, October 2020.

[13]. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, And Moch Arif Bijaksana," Relevance Feature Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 6, June 2021.

[14]. Kamal Taha," Extracting Various Classes of Data from Biological Text Using the Concept of Existence Dependency", IEEE Journal Of Biomedical And Health Informatics, Vol. 19, No. 6, November 2021.

[15]. Shuhui Jiang, Xueming Qian, Jialie Shen, Yun Fu and Tao Mei, "Author Topic Model-Based Collaborative Filtering For Personalized Poi Recommendations", IEEE Transactions On Multimedia, Vol. 17, No. 6, June 2021.

[16]. Beichen Wang, Xiaodong Chen, Hiroshi Mamitsuka, And Shanfeng Zhu," Bmexpert: Mining Medline for Finding Experts In Biomedical Domains Based On Language Model", I IEEE /Acm Transactions On Computational Biology And Bioinformatics, Vol. 12, No. 6, November/December 2021.

[17]. Donald E. Brown," Text Mining the Contributors to Rail Accidents", IEEE Transactions on Intelligent Transportation Systems, Vol. 17, No. 2, February 2021.

[18]. Silvana V. Aciar, Gabriela I. Aciar, Cesar Alberto Collazos, And Carina Soledad González," User Recommender System Based on Knowledge, Availability, And Reputation from Interactions In Forums", IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje, Vol. 11, No. 1, February 2019.

[19]. Yue Hu And Xiaojun Wan," Ppsgen: Learning-Based Presentation Slides Generation for Academic Papers", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 4, April 2021.

[20]. Ning Zhong, Yuefeng Li, And Sheng-Tang Wu," Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2021.