Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

A Review of Text Mining

Sanitha P S
Dept. of Computer Science and Engineering, Federal Institute of Science and Technology,

Abstract - The technology of text Mining can discover, retrieve and extract information from a text corpus, which is usually too complicated for manual work. In order to better comprehend complicated written analytical processing systems, text mining combines technologies including natural language processing, artificial intelligence, information retrieval, and data mining. Text mining was first intended to help national security and intelligence agencies find terrorist activity and other security threats. Text mining made use of text analysis components and technology from other fields, including computer science, management science, machine learning, and statistics, to enhance its performance. Today, text mining's precision and capacity for managing complicated issues have continuously increased.

Key words: Text mining Techniques.

1. INTRODUCTION

The process of converting unstructured text into a structured format with the purpose of identifying significant patterns and fresh insights is known as text mining, also known as text data mining. The organizations are investing heavily to discover a solution that can analyse customer and competitor data to increase competitiveness due to the growing level of competition in business and shifting customer perceptions. E-commerce websites, social networking platforms, published articles, surveys, and many more are the main sources of data. Because a bigger portion of the created data is unstructured, firms find it difficult and expensive to examine it using human resources. Analytical tools have expanded as a result of this problem and the exponential expansion in data output. In addition to handling enormous amounts of text data, it also aids in decision-making. The use of text mining software enables users to extract relevant information from the vast amount of available data. Other frequent applications of text mining include selecting job candidates based on the language used in their resumes, filtering spam emails, categorising website content, flagging insurance claims that might be fraudulent, examining corporate documents as part of electronic discovery procedures, and analysing descriptions of medical symptoms to aid in diagnose.

Text Mining Process:
The text mining process incorporates the following steps to

extract the data from the document.

Text transformation:
Controlling the capitalization of the text is done via a text transformation approach. The two main methods of document representation are provided here.

- Bag of words
- Vector Space

Text Pre-processing:
Information retrieval, Natural Language Processing (NLP), and Text Mining all require preprocessing as a substantial effort (IR). Data pre-processing is used in the text mining industry to glean knowledge and information from unstructured text data. In order to satisfy the user's request, information retrieval (IR) involves selecting which documents from a collection should be retrieved. **Feature selection:**
A big element of data mining is features selection. Choosing features is the process of limiting the amount of processing input or locating the most important information sources. Variable selection is another name for the feature selection.
Data Mining: The text mining process now combines with the traditional process at this step. The structural database uses traditional data mining techniques.
Evaluate: Afterward, it evaluates the results. Once the result is evaluated, the result abandon.

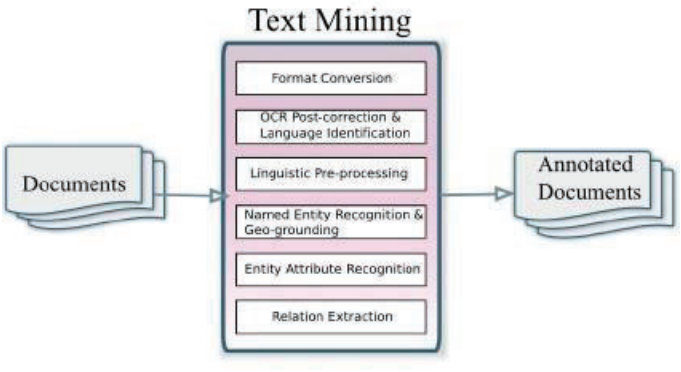

Figure1: Architecture of text mining

The main goal of this review paper is to review the numerous Text Mining -related research and studies.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

2. LITERATURE REVIEW

A journal article titled "Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text" discusses the use of text mining as a method of predicting "proactive personality." Five machine learning techniques were used to create categorization, including Support Vector Machine (SVM), XGBoost, K-Nearest-Neighbors (KNN), Naive Bayes (NB), and Logistic Regression (LR). Four short-answer questions were set to reflect proactive personality. Participants were asked to respond to all of the questions in roughly 60 words, according to their actual thoughts and circumstances, after reading the questions. Data preprocessing is essential for improved classification performance in text mining studies. Preprocessing the data comes next, preprocessing of the participants' written responses. The next phase is Term Frequency-Inverse Document Frequency (TF-IDF), which is a popular method for extracting text features from text. It's time for feature selection now. The very last phase is model training. Finally done "Classification of Proactive Personality". Finding the correct topics from the biomedical documents is a difficult issue, as discussed in the journal article "Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering." Redundancy also has a detrimental effect on the effectiveness of text mining in biomedical text documents. As a result, the exponential rise of unstructured information necessitates topic modelling techniques using machine learning that can identify specific subjects. In this research, a topic modelling method for text mining using a hybrid inverse document frequency and fuzzy kmeans clustering algorithm is proposed. The suggested method reduces redundancy and extracts relevant subjects from biomedical text sources. Topic Modeling is a popular method that discovers the hidden theme and structure in unorganized biomedical text documents the topic modelling technique eliminates the detrimental effects of word repetition in biomedical papers and outperforms RedLDA and LDA for redundant corpora in terms of performance. The proposed topic modelling strategy offers a new method for text mining over biological information and raises the classification accuracy for biomedical datasets. The article "A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining" talks about the BERN tool, which is used for neural biomedical named entity recognition and multi-type normalisation. High-performance BioBERT named entity recognition models are used by the BERN to both identify already recognised entities and discover new ones. The categories of overlapping entities are also identified using probability-based decision methods. In order to give each identified entity a unique identifier, different named entity normalisation models are implemented into BERN. For tagging items in PubMed articles or raw text, the BERN offers a Web service. For text mining tasks including finding new named entities, retrieving information, responding to queries, and extracting relationships, researchers can use the BERN Web service. A journal article titled "Assessment of Career Adaptability: Combining Text Mining and Item Response Theory Method" in this journal assesses career adaptability by combining text mining and item response theory (IRT), using responses to questionnaire items as an objective measure and college students' self-reported career adaptability as a subjective measure. Under a Bayesian paradigm, the two are combined. Text categorization findings were used as prior knowledge when calculating IRT capacity parameters to examine if adding prior information can increase accuracy. The validity of text categorization and IRT, together with model measurement, were also explored. In this study, titled "Noise-Robust Wagon Text Extraction Based on Defect-Restore Generative Adversarial Network," Wagon text extraction primarily relies on the tedious, time-consuming, boring, and error- prone manual identification of pertinent information. Create a two-stage wagon text extraction method based on transfer learning and defect-restore generative adversarial networks to overcome this issue (GAN).

3. METHODOLOGIES

Text mining can be used to analyse human personality and predict careers. There is discussion of the various methods for predicting careers and human personalities in [1][4] articles. In [1] In order to assess proactive personality, four short-answer questions were created. After reading the questions, participants were asked to answer each in around 60 words, based on their actual thoughts and situations. In text mining investigations, data pre-processing is crucial for enhancing classification performance. The preparation of the data and the written responses from the participants follows. The following step is Term Frequency-Inverse Document Frequency (TF- IDF), a well-liked technique for extracting text features from text. Now is the time to choose your features. Model training is the final stage. "Classification of Proactive Personality" is finally finished. For classification, many machine learning techniques are utilised.

- SUPPORT VECTOR MACHINE (SVM)

The hyperplane with the greatest separation was sought for as a proportional classification border, and this was the initial goal of SVM. It has exceptional generalisation performance and is based on the statistical learning theory's structural risk minimization (SRM) premise. In order to effectively handle the nonlinear classification problem, kernel functions such linear kernel, polynomial kernel, radial basis kernel (RBF), fourier kernel, and spline kernel have recently been added to support vector machines (SVM).

- XGBOOST

Data scientists have employed the scalable end-to-end tree boosting method known as XGBoost, which is distinguished by quick computation and good performance, to achieve cutting-edge outcomes in numerous machine learning competitions. Scalability in all situations is the key to XGBoost's success. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

method scales to billions of samples in distributed or memory-limited situations and performs more than ten times quicker than currently used popular alternatives on a single machine.

• K NEAREST NEIGHBORS (KNN)

KNN, a classifier introduced by Cover and Hart in 1968, has demonstrated remarkable performance with huge sample sizes. In very mild circumstances, the error rate of KNN can achieve Bayes optimization. Since KNN is an instance-based approach, no such 'model' was trained; instead, classification was determined by comparing instances in the training data with cases. Because of this, KNN is particularly sensitive to the number of features; irrelative features can significantly affect prediction accuracy.

• NAIVE BAYES A straightforward but widely used statistical classifier is called Naive Bayes. Decisions in text mining are based on the presence or absence of specific traits. This indicates that each feature was given a probability of belonging to a particular class based on training data. Following the calculation of all probabilities, a choice may be made depending on the presence of features in the testing set. The word "naive" denotes the autonomous treatment of each aspect. In other words, it will be assumed that all features will appear independently and that the frequency of features in the testing set will not be considered.

• LOGISTIC REGRESSION

By creating a regression function, one can utilise logistic regression, a generalised linear regression, to achieve classification or prediction. The binary classification problem is the focus of the logistic regression model, which can also handle multi-classification issues. The additional role played by social media text in forecasting people's personalities is remarkable, and it is logical to believe that this kind of influence is useful for predicting other attributes in addition to proactive personality. Text mining technology demonstrated considerable promise in predicting people's proactivity, particularly for identifying people with low proactivity, and this can be very helpful for career education practice in high school and college. The highest specificity and accuracy were 0.842 and 0.969, respectively. In this text mining, support vector machines and logistic regression displayed consistent results. In [4] the self-reported career flexibility of college students is used as a subjective measure in this work, and responses to questionnaire items are used as an objective measure, together with text mining and item response theory (IRT). Under a Bayesian paradigm, the two are combined. Text categorization findings were used as prior knowledge when calculating IRT capacity parameters to examine if adding prior information can increase accuracy. The validity of text categorization and IRT, together with model measurement, was also explored. Here, use three techniques to evaluate the latent features of the individuals' career adaptability: Three methods of combining textual analysis and IRT in a Bayesian context are

(1) IRT modelling on the 35-item questionnaire

(2) text classification of the self-narratives

(3) text classification.

Using a text-IRT combination method, suggest a way to test career adaptability that combines subjective and objective measurement. By contrasting three analytical techniques— text classification, IRT, and text-IRT combination methods—we compare the three samples of validation sets of 300, 600, and 900 people to test the suggested method. The outcome demonstrates that, especially in a small sample, the text-IRT combination method could deliver the best results for career adaptability prediction. The best strategy for large samples is text categorization, particularly when identifying people with low career flexibility.

1). In this instance, the text classification approach had the maximum sensitivity in 300-person samples, while the text- IRT method had the best prediction effect, high reliability, and special advantages in accuracy.

2). The text classification technique provided the best predictive impact in 600-person samples. The outcome was largely positive and offered particular benefits for diagnosing low career adaptability. But this must be chosen in accordance with actual requirements. The text-IRT technique is more suitable if the accuracy requirement is high and sensitivity can be compromised.

3). When accuracy, sensitivity, and specificity must be considered, the text-IRT technique is more suited for 900 subjects. Text classification works well when identifying individuals with low career flexibility.

4). The sensitivity of IRT and text-IRT approaches, as well as the accuracy, specificity, and the negative predictive values of text classification, were all influenced by sample size. In [2][3] articles, the use of text mining in biomedical research is discussed. Researchers can spend less time and effort finding and extracting meaningful information from the large amount of biomedical literature by using quick and accurate text mining methods. In [2] finding the precise themes from the biological documents is difficult when topic modelling. Additionally, redundancy in biomedical text documents has a detrimental effect on text mining quality. The exponential rise of unstructured materials necessitates machine learning topic modelling algorithms that can identify certain themes. In this research, we suggested a hybrid inverse document frequency and machine learning fuzzy k-means clustering algorithm topic modelling method for text mining. The suggested method reduces redundancy and extracts specific subjects from the biological text documents. This approach assesses the number of topics in biomedical documents and can be used with discrete and continuous data. a topic modelling technique that extracts latent semantic subjects from biological documents has been

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

suggested. In contrast to RedLDA and LDA for redundant corpora, the suggested topic modelling technique outperforms both and eliminates the detrimental effects of word redundancy in biomedical documents. The proposed topic modelling strategy offers a new method for text mining over biological information and raises the classification accuracy for biomedical datasets. With different numbers of topics, the suggested topic modelling technique produces improved clustering results. Additionally, experimental findings demonstrate that the suggested topic modelling technique's time performance is steady as the number of topics is raised. In [3] Text mining technologies like tmTool and ezTag use outdated named entity recognition methods that are ineffective at reliably identifying new things. Additionally, conventional text mining methods do not consider overlapping entities, which are typically shown in results of named entity recognition using several types. We suggest BERN, a tool for multitype normalisation and neural biological named entity recognition. High-performance BioBERT named entity recognition models are used by the BERN to both recognise and find new entities. Additionally, decision rules based on probability are created to identify the kinds of entities that overlap. Using Word Piece embeddings, BERN's BioBERT NER models identify already-known entities and find new ones. The out-of- vocabulary issue affects Pyysalo et al word's embeddings. The embeddings cannot give a word in a text a rich representation if the word is not included in their lexicon. The Word Piece embeddings, on the other hand, are a method for breaking a word up into smaller units (i.e., sub- word units) and expressing each unit. As a result, it is possible to use the Word Piece embeddings to extract features from uncommon or unfamiliar words, which is particularly beneficial for finding new entities. The RESTful Web service of BERN was implemented using Python and Node.js. BERN run BioBERT NER models which are pre- trained with TensorFlow3, on our server to recognize incoming biomedical text such as PubMed articles and raw text. The server specifications are as follows: • Operating system: Ubuntu 18.04.2 LTS

- CPU: Intel Xeon E5-2687W v3
- RAM size: 128 gigabytes (GB)
- GPU: NVIDIA Titan X (Pascal) with 12 GB of memory
- Hard disk drive size: 2 terabytes

In the ‘‘Text’’ tab of BERN Web service, researchers can obtain NER+NEN results of submitted raw text in Pub Annotation JSON format, and see the visualized results under the text window. Also, as the BERN demonstration shows, entities are highlighted in their entity type color. When the mouse cursor is placed on an entity name, its entity type and entity ID are displayed in a tooltip. Using BioBERT NER models, BERN detects recognised entities and finds new entities. In terms of F1-score on genes/proteins, diseases, drugs/chemicals, and species, the BioBERT models surpass NER models of current Web- based text mining technologies. In [5] Wagon text extraction primarily relies on the tedious, time-consuming, boring, and errorprone manual identification of pertinent information. We create a two-stage wagon text extraction system based on transfer learning and defect-restore generative adversarial networks to address this issue (GAN). The proposed technique significantly outperforms the state-of- the- art with an accuracy of 97.76% on 2682 real-world test sub-images. Evaluate CTPN, EAST, and TextBoxes++ as three current architectures for text detection using the dataset. On four open datasets, TextBoxes++ outperformed rival methods, but on the wagon photos, it only manages an F-score of 0.3385. For the new feature-space distribution, these architectures need to be rebuilt.t is shown that the performance of the refined models outperforms the similar pretrained models by a large margin. For example, the pre- trained CTPN model only receives an F-score of 0.5367 while the optimised CTPN model receives an F-score of 0.9107. Compare the proposed method to three current scene text recognition architectures, including CRNN, CRNN with Attention, and Attention-based OCR, in order to show its efficacy. The proposed defect-restore GAN model outperforms all previous text recognition techniques with an accuracy of 97.76%, proving the distinguishability of the derived features.

4. CONCLUSION

The process of converting unstructured text into a structured format with the purpose of identifying significant patterns and fresh insights is known as text mining, also known as text data mining. Text mining combines technologies including natural language processing, artificial intelligence, information retrieval, and data mining. Text mining was first intended to help national security and intelligence agencies find terrorist activity and other security threats. Text mining made use of text analysis components and technology from other fields, including computer science, management science, machine learning, and statistics, to enhance its performance. Today, text mining's precision and capacity for managing complicated issues have continuously increased. This review paper focuses on text mining approaches that are applied in various contexts, such as human-related career and personality prediction, as well as text mining's utility in mining or extracting text from challenging biological papers. The technology utilised in article [4] is superior to that in article [1] when it comes to text mining for human-related personality and career prediction. The article [3] uses better technology than article [2] when comparing the technologies utilised in text mining for biomedical publications. Frequent applications of text mining include selecting job candidates based on the language used in their resumes, filtering spam emails, categorising website content, flagging insurance claims that might be fraudulent, examining corporate documents as part of electronic discovery procedures, and analysing descriptions of medical symptoms to aid in diagnoses. All this fields use text mining.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

5. REFERENCES

[1] Peng Wang, Yun Yan, Yingdong Si, Gancheng Zhu, Xiangping Zhan, Jun Wang, And Runsheng Pan "Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text".

[2] Junaid Rashid, Syed Muhammad Adnan Shah, Aun Irtaza, Toqeer Mahmood, Muhammad Wasif Nisar, Muhammad Shafiq, And Akber Gardezi "Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering".

[3] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, And Jaewoo Kang "A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining".

[4] Lihui Zhang, Gancheng Zhu, Shujie Zhang, Xiangping Zhan, Jun Wang, Weixuan Meng, Xin Fang, And Peng Wang "Assessment of Career Adaptability: Combining Text Mining and Item Response Theory Method".

[5] Meng Lei, Yi Zhou, Li Zhou, Jiannan Zheng, Ming Li, And Liang Zou" Noise-Robust Wagon Text Extraction Based on Defect-Restore Generative Adversarial Network"