International Journal of Applied Engineering & Technology

MACHINE LEARNING BASED TEXT CLASSIFICATION USING PROBABILISTIC NEURAL NETWORK

Dr. T. Thendral[1*] and Ms. Eben Angel Pauline D[2]
[1]Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts & Science for Women, Coimbatore
[2]Assistant Professor, Department of Information Technology, Women's Christian College, Chennai
[1]Orcid: 0000-0002-8302-458X, Vidwan ID: 313348

ABSTRACT –
The text assists data company executives in sorting through the clutter and obtaining the pertinent facts to make the best options on corporate strategy and expansion. Text data, on the other hand, cannot be handled manually due to its un-structured, abundance, and raw state. To extract the text data, an effective automated method was required. One of the difficult issues in Natural Language Processing (NLP) is text categorization. The main objective of this work is to text data classification using Machine Learning (ML) algorithms. This work proposed Probabilistic Neural Network (PNN) classification method to solve the text classification tasks in the field of NLP.

Keywords: Text Mining, Text Classification, Natural Language Processing (NLP), Machine Learning

I. INTRODUCTION

Language has an important role in communication and information sharing, two things that are essential for the progress of the human species. Because of the Internet, interpersonal communication has changed in terms of tone, tempo, and efficacy. The development of technology has significantly transformed communication, to the extent that it has become the primary means of everyday communication. The Internet is used for almost everything, which produce a ton of data every time we place a pizza order, buy a television, write a movie review, spend time with a buddy, or send a photo through instant messaging. According to industry estimates, just 21% of the data that is now available is in a structured manner in the modern world, which is the twenty-first century. As we chat, tweet, or send communications via Facebook, WhatsApp, Twitter, or text messages, data is being created. More than 3.7 billion individuals use the Internet; of these, 456,000 tweets are published on Twitter, 46,740 photographs are posted on Instagram, and 1.5 billion people use Facebook every day. Every minute, 510,000 comments are submitted, 293,000 statuses are changed, and 103,447,520 spam emails are sent.

It is crucial to become familiar with text analysis methods in order to obtain meaningful and qualitative research from the text data. Text analysis, sometimes referred to as text mining, is the process of extracting useful information from natural language text. Text mining often involves the organization of input text, the identification of patterns in the structured data, and the evaluation of the resulting output. The primary goal of this project is to use NLP to transform the text into structured data for the purpose of analysis. This is because text mining involves extracting valuable information from textual sources.

Text Classification also known as the Text Categorization. It is the process of allocating the documents, based on its contents, into one or more categories or no category at all. A category discusses the correlation between the entities and elements of information or knowledge. The process entails discerning the significant themes in a text by categorizing the material into a predetermined set of subjects. After the text has been classified, a computer program may generally consider the document as a "collection of words". It does not endeavor to analyze evidence as information extraction does.

The categorization technique is used to quantify the frequency of terms in a text, and based on these frequencies, it determines the prominent subjects addressed in the document. The selection of themes in a thesaurus is usually predetermined, and relationships between words are established by the use of broad phrases, narrow terms, synonyms, and related terms. It has a wide range of applications in several disciplines. Numerous companies and sectors provide customer service and are required to respond to inquiries on various subjects from their customers.

Text classification plays a significant role in multiple information management systems due the massive growth of unstructured data in the web. Text being an exceedingly rich source of information and the extraction of knowledge would be a time consuming and complex task. Hence, text classification is utilized for structuring the text document which enhances the decision making and the automation processes. Preprocessing and Dimensionality

Reduction are the vital steps in the text classification since the high dimensional text serves as the important challenge in the text classification. This in turn enhances the speed and the accuracy of the classification. Dimensionality Reduction tasks are categorized into two sub tasks. They are Feature Extraction and Feature Selection. Feature Selection and Extraction is used to improve the accuracy, scalability and the efficiency of text classification by constructing a vector space.

The classifier is constructed automatically by acquiring knowledge of the characteristics of the classes from a predetermined set of training texts. Text categorization is used in diverse domains such as spam filtering, email routing, topic tracking, emotion analysis, and web page categorization.

Text classification may be carried out by human or automated methods. In the manual technique, a human annotator determines the textual content and categorizes the documents according to their content. This strategy produces superior outcomes, but, its main disadvantage lies in its high cost and time-consuming nature.

The field of image recognition has seen great progress using machine learning. Modern image categorization methods make use of Neural Networks (NN) with several layers. These networks function effectively because they can acquire increasingly abstract hierarchical representations of the input. NN have been found to produce good outcomes in the setting of NLP. Particularly, phrases and words with a specific structure make up human sentences.

II. RELATED WORKS

An Approach for Concept-based Automatic Multi-Document Summarization using Machine Learning (Padma Priya & Duraiswamy 2012) Programmed content rundown is an old test yet the ebb and flow research course inclines towards developing patterns in biomedicine, training areas, messages and web journals. This is because of the way that there is data overburden in these regions, particularly on the World Wide Web. To study the psychological part of content rundown feel that utilization of machine learning is exceptionally useful. Our methodology will create the extensive exactness measures. In future it will attempt to actualize this view topic to enhance the execution of content rundown framework.

The search results by mining subtopics from several perspectives (Chieh-Jen Wang et al. 2012) In order to diversify search results using the mined subtopics, it presents two document rating algorithms together with six subtopic mining approaches. Subtopics are mined either directly from the queries themselves or indirectly from the retrieved documents in order to maintain less repetitive subtopics and decrease the number of missing subtopics. The suggested subtopic-based diversification algorithm takes all three of these aspects of the mined subtopics into account at once: richness, significance, and novelty. According to experimental findings, the subtopic-based diversification algorithm may balance variety and relevance to enhance search result diversification performance.

In 2015, Solane Duque and Mohd Nizam bin Omar Pre-processing involves eliminating redundant entries and merging or deleting inconsistent or noisy data. The characteristics of the dataset must also be transformed into numeric data during pre-processing and saved in a readable manner. As they provide a significant amount of blogs, real-time data, SMS, social networking websites, and chat apps are appealing sources of information for data mining.

According to Jansen et al. (2009), the primary benefit of Twitter is that users can express themselves concisely in 140 characters or less. Tools like "followers," "following," and "re-tweets" make this possible.

In order to develop a low dimensional but informative vector representation of both words and texts that enables the finding of semantic commonalities between them, the suggested technique employs a NN model, i.e., paragraph vectors (Le & Mikolov 2014).

Working topic segmentation employing a lexicon that correlated phrase deviations to parameters was used for topic recognition and tracking (Liu et al. 2015). To find latent themes in a collection of documents, we first utilize k-means clustering to extract the K cluster centre of the learned paragraph vectors. It utilizes the vector's cosine similarity as a distance measure for kmeans clustering.

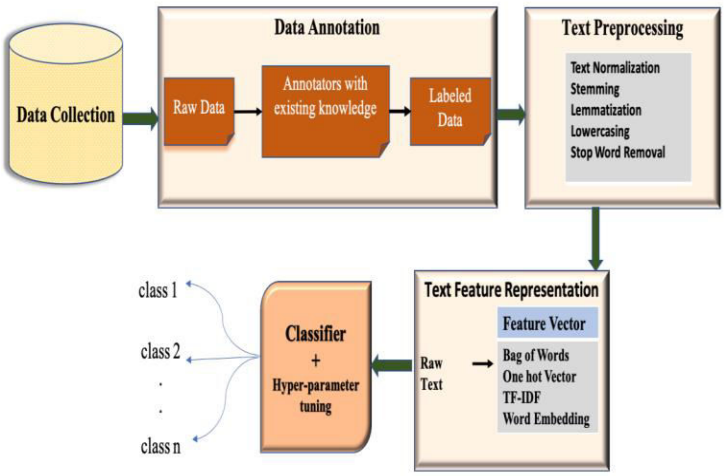
A subtopic-based Multi-records Summarization (SubTMS) technique receives probabilistic point model to find the subtopic data inside each sentence and uses an appropriate various leveled subtopic structure to depict both the entire reports assortment and all sentences in it. With the sentences spoke to as subtopic-vectors, it surveys the semantic separations of sentences from the reports assortment's principle subtopics and picks sentences which have short separation as the last rundown of the archives assortment (Gong S et al. 2010).

The content is spoken to by a differing set of potential markers of significance which don't target finding topicality. These pointers are joined, frequently utilizing AI procedures, to score the significance of each sentence. At last, a synopsis is created by choosing sentences in a covetous methodology, picking the sentences that will go in the rundown individually or all-inclusive enhancing the choice, picking the best arrangement of sentences to shape an outline (Nenkova, A & McKeown, K 2012).

Ammar Ismael, (2019) studied the automatic text classification techniques which belongs to the supervised machine learning. K-Nearest Neighbours (kNN), SVM, and NB were discussed. Better classification capability was offered by kNN among the three methods. It was used in combination with TF-IDF for better classification. Enhancement of specific terms was done by designing suitable weights using term weighting methods.

In their study, Muhammad Hassan and colleagues (2019) used the Lifelong Machine Learning (LML) technique to classify text. They achieved this by extracting and reusing knowledge blocks. It was used in a system that concurrently acquired knowledge from various areas. It was often used in the field of artificial intelligence for reinforcement, unsupervised, and supervised learning.

III. SYSTEM METHODOLOGY



A. Corpus Collection and Annotation

With more and more people using the internet every day, the amount of text data available to solve text classification problems has increased significantly. This has allowed the text classification problem to be viewed as an automatic text classification task, rather than limited to the relatively small amounts of data that humans can

process on their own. Text classification corpora are used to categorize natural language texts according to the content. For example, classifying a movie review based on positive or negative sentiment. This is usually done by annotating the corpus, which is labeling data to make it useable for the classifier. So, the corpora of movie reviews are manually labeled as the positive or negative sentiment. Later use those positive and negative reviews to learn a model and create a sentiment analyzer. As a result, the natural language dataset is referred to as corpora, and an annotated corpus is a single collection of data annotated with the precise definition.

B. Data Cleaning and Pre-processing

The text data generated over the internet is mostly in its natural human format of news articles, movie reviews, social media posts, tweets etc and in so many other forms and from many various resources. However, all this data has one flaw to it, i.e., the text is the most unstructured form of data available to us. Internet documents frequently include unnecessary noise, typos, and bits that don't provide any information, including HTML elements, repetitive phrases, scripts, etc. They need to be cleaned and pre-processed before feeding into the classifier. To pre-process the text simply means bringing the text into a predictable and analyzable form for the classification task. There exists a set of a variety of techniques to pre-process the text before feeding it to the classifier. Text normalization is the process of transforming a text into a canonical form. For instance, "good" and "gud" can be changed to "good," which is "good's" canonical form. Another illustration is the reduction of terms like "stop words," "stop-words," and "stop words" to their single word form. The need of text normalization increases for noisy texts like blog posts and comments on social media where Out-of-Vocabulary (OOV) is common.

The stem of a word is formed by removing its prefix or suffix. The stemming root could just be a canonical version of the root form and not a true root word. For instance, because endings were simply cut off, the phrases "trouble," "troubles," and "troubled" may really be changed to trouble instead of trouble. Stemming, which merges distinct words of a term together and addresses the sparsity difficulties as well, is frequently used as a vocabulary reduction strategy. Lemmatization, like stemming, reduces a word to its root form. The sole distinction is that lemmatization guarantees that the language-specific nature of the root term. The lemmatization produces genuine words rather than merely cutting stuff off. The term "better" would translate to "good," for instance.

C. Text Feature Representation

Textual data presents unique set of challenges. As machines and text classification algorithms are unable to read and comprehend language in any way like that of a person, dealing with text data is a challenge. These approaches work on numbers rather than the text. So, we need bunch of steps to encode the text into the numerical vector representations (features) that can be feed to the classifier so that they can churn out some great insights from text. These representations must be semantically significant as well as able to convey the fullest possible understanding of a word's linguistic meaning. The performance of the classifier as a whole can be significantly impacted by a carefully chosen informative input representation. There are only a few approaches for text encoding, and each has advantages and disadvantages.

Tokenization - The first thing that is usually done is to tokenize the text. It is the process of spilling a sentence into words/tokens in a way that a machine can process them, with a view to later training a classifier that can understand their meaning. Tokenization involves three steps: figuring out how each word fits into the overall sentence, breaking a long sentence into words and describing the input text's structure.

Character Encoding - Consider a word LISTEN, it's made up of a sequence of letters. These letters can be represented by numbers using an encoding scheme. A popular one called ASCII has these letters represented by the numbers. This bunch of numbers can then represent the word LISTEN. But the word SILENT has the same letters, and thus just the same numbers in a different order. So, it makes hard for us to understand the semantic meaning of a word just by the letters in it. So, it might be easier and informative to encode words then to encode simply the characters of words.

Word Encoding - Consider the sentence I love my Dog and then the sentence could be (001,002,003,004). Now if we take another sentence, I love my cat, then how would we encode it? I love my has already been given 1,2,3, so all we need to do is encode cat and give that a number 005. Now when we look at both sentences, they are 1,2,3,4 and 1,2,3,5 which show some form of similarity between them.

Bag of Words (BOW) – It creates a dictionary consisting of 'd' concepts, where 'd' represents the total number of unique words in the text corpus (Y. Zhang et al., 2010). Subsequently, it produces d-dimensional vectors for each textual phrase, where each dimension corresponds to the frequency of the respective word in the document. The approach is referred to as Bag-of-Words (BOW) because to its emphasis on the distinct words included in the text, rather than their sequential arrangement. The model's primary focus is on the presence of recognized words in the text content, rather than their specific location within the document.

One-hot encoding - It is another method of text encoding that is often employed. Every word in the document is used to create a dictionary in this case, and each word is converted into a column in the vector space. Then, each text is converted into a vector of 0s and 1s. Words are encoded with a 1 for presence and a 0 for absence. The one hot encoding retains the word order. A tensor serves as a representation for each document in the one hot encoding technique. The representation of the document corpus is a very large and very sparse number of document tensors, each of which is composed of potentially very long sequences of 0/1 vectors.

Encoding with an index - Index-based encoding is an alternative technique used to maintain the sequential arrangement of words in phrases. It works by assigning a distinct index or numerical value to every word. The first stage involves establishing a vocabulary that connects words with corresponding indices. Each phrase is represented by a set of indexes, with each number signifying a single word, using this vocabulary as a basis. It does, however, create an unnecessary numerical gap between paragraphs.

D. Text Classification

The expanding volume of information has made text categorization a crucial function in modern life. Today's real datasets have several labels, making text categorization more significant.

Naive Bayes (NB)

NB classifiers are linear classifiers recognized as easy yet very effective classifiers. The NB classifier probabilistic model is based on Bayes' theorem, and the naïve adjective is derived from the assumption that the attributes in a dataset are mutually independent. NB classifiers can outperform the stronger options, especially for tiny sample sizes.

Krill Herd (KH)

In the text classification, each document is assigned to their particular categories by analyzing the contents of the particular document. Likewise, the group of distinct krill search the similar food particles resolved by the least distance between each individual krill and the food particles. The krill herd algorithm for text classification consists of following steps. Initialization of Krill Herd parameters, Initialization of Krill Herd memory, Krill Herd Motion calculation, Krill Herd Genetic Operators, Krill position updation and Termination Criterion.

Proposed Probabilistic Neural Network (PNN)

This work proposed PNN for text classification. The PNN design is composed of several interconnected processing units or neurons organized in consecutive layers. The input layer unit just distributes information to the neurons in the pattern layer, without engaging in any computational processes.

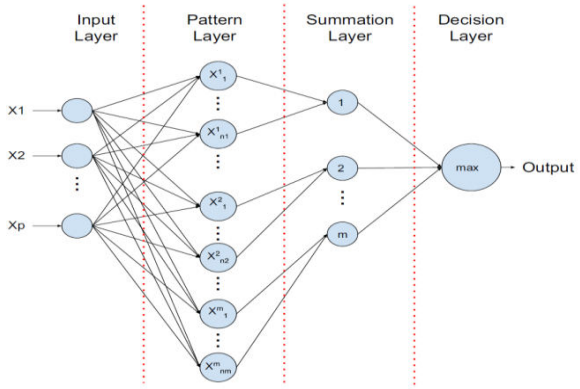


Figure 2: PNN Architecture

PNN Classification Algorithm

Input: N number of documents

Output: Classes with documents

Step 1. Using the neuron ranking algorithm, choose the training samples' most representative neuron for each class.

Step 2. Using each of the sample neurons that were chosen, build a probabilistic neural-network classifier. Find the classification error rate, which is determined by dividing the total number of training samples in each class by the number of misclassifications. Sort the training samples into the appropriate classes.

Step 3. Using the neuron ranking technique, choose an extra representative neuron for classes where the classification error rate criteria is not met.

$$p_i(\alpha) = \frac{1}{(2\pi)^{d/2}\sigma^d}\frac{1}{N_i}\sum_{j=1}^{N_i} \cdot \exp\left[-\frac{(\alpha-\alpha_{ij})^T(\alpha-\alpha_{ij})}{2\sigma^2}\right]$$

Navigate Step 2 is repeated up until the classification error rate requirements for all classes are met. Even if all of the training samples are used to create the pattern layer, if the training examples are subpar, the needed classification accuracy could not be achieved. A larger categorization error rate will be applied if this is the case.

IV. RESULT AND DISCUSSION

All the experiments are done in Python and running on windows 16.

A. Dataset Details

This work used BBC News dataset. It includes 2225 stories from five distinct domains—business, entertainment, politics, sport, and technology—is one of the most often used benchmarks in text categorization research.

B. Performance Factors

This research work used accuracy, precision, recall, and f-score measure have been used to evaluate the performance of the proposed model.

Table 1: Performance Analysis of ML Algorithms

ML Algorithms	Performance Measures			
	Precision	Recall	F-Measure	Accuracy
NB	0.659	0.621	0.639	0.751
KH	0.701	0.709	0.698	0.814
PNN	0.724	0.712	0.718	0.829

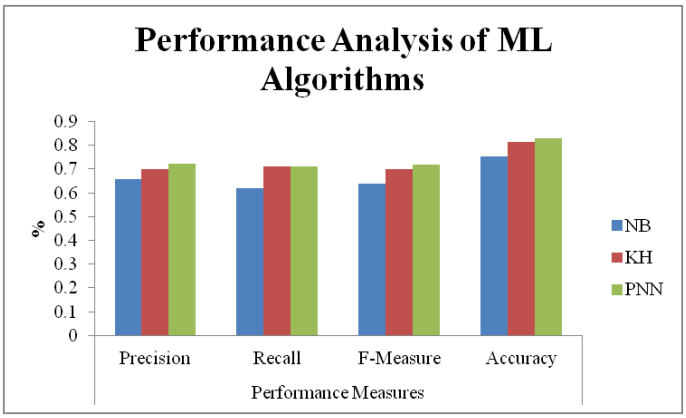


Figure 2: Comparative Analysis of ML Algorithms

Table 1 and Figure 2 represents the Performance Analysis of ML Algorithms. From the result observation, it is noticed that the proposed PNN algorithm gives higher accuracy, precision, recall, f-score measure than other ML algorithms.

	ArticleId	Text	Category	content_clean
0	1833	worldcom ex-boss launches defence lawyers defe...	business	worldcom ex-boss launches defence lawyers defe...
1	154	german business confidence slides german busin...	business	german business confidence slides german busin...
2	1101	bbc poll indicates economic gloom citizens in ...	business	bbc poll indicates economic gloom citizens maj...
3	1976	lifestyle governs mobile choice faster bett...	tech	lifestyle governs mobile choice faster bett...
4	917	enron bosses in $168m payout eighteen former e...	business	enron bosses $168m payout eighteen former enro...
5	1582	howard truanted to play snooker conservative...	politics	howard truanted play snooker conservative le...
6	651	wales silent on grand slam talk rhys williams ...	sport	wales silent grand slam talk rhys williams say...
7	1797	french honour for director parker british film...	entertainment	french honour director parker british film dir...
8	2034	car giant hit by mercedes slump a slump in pro...	business	car giant hit mercedes slump slump profitabili...
9	1866	fockers fuel festive film chart comedy meet th...	entertainment	fockers fuel festive film chart comedy meet fo...

Figure 3: Data cleaning and Stop-word removal

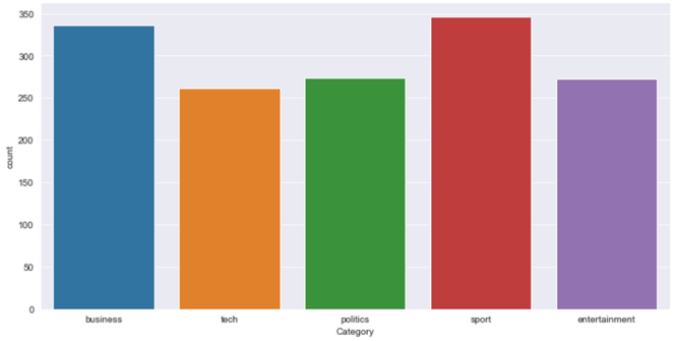


Figure 4: Output Label – Graphical representation

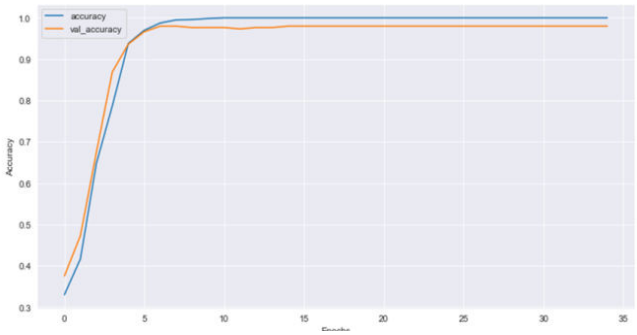


Figure 5: Training Accuracy – Graph

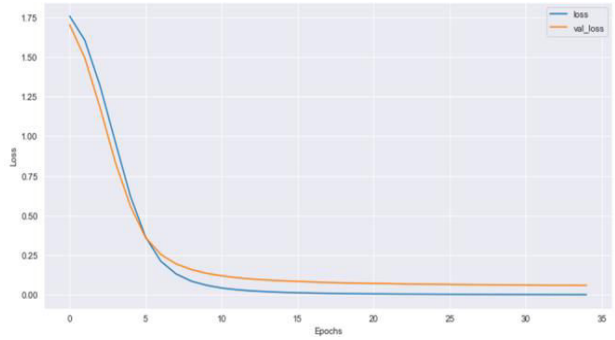


Figure 6: Training loss – Graph

V. CONCLUSION

The text assists data company executives in sorting through the clutter and obtaining the pertinent facts to make the best options on corporate strategy and expansion. Text data, on the other hand, cannot be handled manually due to its un-structured, abundance, and raw state. To extract the text data, an effective automated method was required. One of the difficult issues in NLP is text categorization. The main objective of this work is to text data classification using ML algorithms. This work proposed Probabilistic Neural Network (PNN) classification method to solve the text classification tasks in the field of NLP. From the result observation, it is noticed that the proposed PNN algorithm gives better performance than other ML algorithms.

VI. REFERENCES

[1] A. Korhonen, "Improving Literature-Based Discovery," Springer International Publishing Switzerland, 2015.

[2] Abdullah Wahbeh H, Mohammed A1-Kabi, “Comparative Assessment of the performance of three WEKA Text Classifiers Applied to Arabic Text”, Vol.21, No. 1, 15-28, 2012.

[3] Acharya, Anish, Rahul Goel, Angeliki Metallinou, and Inderjit Dhillon. "Online embedding compression for text classification using low rank matrix factorization." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6196-6203. 2019.

[4] Aghdam, Mehdi Hosseinzadeh, and Setareh Heidari. "Feature selection using particle swarm optimization in text categorization." Journal of Artificial Intelligence and Soft Computing Research 5, no. 4 (2015): 231-238.

[5] H. Yalcin, "Exploring Technology and Engineering Management Research Landscape," IEEE, 2019.

[6] Lai, Siwei, Liheng Xu, Kang Liu, and Jun Zhao. "Recurrent convolutional neural networks for text classification." In Twenty-ninth AAAI conference on artificial intelligence. 2015.

[7] Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang. "Recurrent neural network for text classification with multi-task learning." arXiv preprint arXiv:1605.05101 (2016).

[8] M. Yetisgen-Yildiz, "A new evaluation methodology for literature-based discovery systems," Journal of Biomedical Informatics, 2009.

[9] Marie-Sainte, S. L., & Alalyani, N. (2018). Firefly algorithm based feature selection for Arabic text classification. Journal of King Saud University-Computer and Information Sciences.

[10] McCallum K, et al, “Text classification from labeled and unlabeled documents using EM”. Machine Learning, 39(2/3), pp. 103–134, 2000.

[11] Menaka.S, Radha.N, “Text Classification using Keyword Extraction Technique”. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013 ISSN: 2277 128X

[12] Milward, "Linguamatics," 2020. [Online]. Available: https://www.linguamatics.com/what-text-mining-text-analytics-andnatural-language-processing. T. Mohler, "Lexalytics," 9 September 2019.

[13] Mrs. SayantaniGhosh et al., “A tutorial review on Text Mining Algorithms”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 4, June 2012.

[14] Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. "Text classification from labeled and unlabeled documents using EM." Machine learning 39, no. 2-3 (2000): 103-134.

[15] Pietramala, A., Policicchio, V. L., Rullo, P., & Sidhu, I. (2008, September). A genetic algorithm for text classification rule induction. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 188-203). Springer, Berlin, Heidelberg.

[16] R. Papka, J. Allan, Document classification using multiword features, in:Proceedings of the Seventh International Conference on Information and Knowledge Management Table of Contents, Bethesda, Maryland, United States, 1998, pp. 124–131.

[17] R. Talib, "Text Mining: Techniques, Applications and Issues," International Journal of Advanced Computer Science and Applications (IJACSA), 2016.

[18] S. Wu, "Literature Explorer: effective retrieval of scientific documents through nonparametric thematic topic detection," Springer, 2019.

[19] Vilar, Estevan. "Word embedding, neural networks and text classification: what is the state-of-the-art?." Junior Management Science 4, no. 1 (2019): 35-62.

[20] Wang, Fang, Zhongyuan Wang, Zhoujun Li, and Ji-Rong Wen. "Concept-based short text classification and ranking." In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1069-1078. ACM, 2014.

[21] Kadhim, Ammar Ismael. "Survey on supervised machine learning techniques for automatic text classification." Artificial Intelligence Review 52, no. 1 (2019): 273-292.

[22] Arif, Muhammad Hassan, Muhammad Iqbal, and Jianxin Li. "Extracting and reusing blocks of knowledge in learning classifier systems for text classification: a lifelong machine learning approach." Soft Computing (2019): 1-10.