DOI : https://doi.org/10.56452/7-2-547

ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

Multi-Class Support Vector Machines for Text Classification Using a Class-Incremental Learning

Divya Kapil
Asst. Professor, School of Computing, Graphic Era Hill University, Dehradun, Uttarakhand India 248002,

Abstract:

The challenging topic of multiclass picture classification has been extensively studied in the past. Methods based on decomposition are often used to deal with it. These techniques often involve breaking down a complex issue into simpler subproblems that may then be tackled using any number of tried-and-true learning algorithms that would not otherwise be applicable. This study examines the effectiveness of decomposition-based approaches and offers some suggestions for enhancing the meta-learning stage. This study presents four strategies for improving multiclass classification ensembles. The first one shows how decomposition-based approaches for multiclass issues may benefit by using a combination of experts scheme to minimise the number of operations required during training. Bayes' theorem underpins the second method for merging learner-based outcomes. Training complexity is reduced in relation to the number of classifiers when the Bayes rule is combined with arbitrary decompositions. Decomposing the original problem into smaller ones and assembling the output of the base learners together with that of a multiclass classifier are two additional strategies recommended for improving the final classification accuracy. Finally, four datasets with varied degrees of categorization complexity are used to test the effectiveness of the suggested unique meta-learning strategies. The suggested approaches consistently show a significant increase in accuracy over classic Text categorization methods.

Keywords: ensemble learning; mixture of experts; decomposition-based methods;

Introduction

Several different techniques, all belonging to the ensemble learning family of methods, attempt to solve the multiclass classification problem by using various algorithms and methods for combining data from several networks. To find the best binary classifier, several methods start with a decomposition strategy [1,2] that uses sophisticated criteria to divide the data. The issue of transferability becomes more pressing as more data is collected or as more networks are combined. It is important not to ignore the use from ensemble approaches [3,4,5] in simultaneous training [6] to aggregate the predictions of several methods or network into a single forecast. Some studies [7] conclude that a generalizer may be used to integrate expert viewpoints and coordinate the allocation of information necessary to make a final conclusion[8].The goal of this research was to investigate the field of deep learning for the purpose of classifying images into several categories. We present a set of classification methods that utilise ensembles using deep neural networks to address the multiclass classification problem and increase accuracy. In this article, we analyse and rank many methods of meta-learning[9]. Each multiclass problem may be partitioned into two binary ones, allowing us to more precisely identify class boundaries on the basis of shared characteristics.Typically, we consider an ensemble to be the primary multiclass classifier supplemented with class-specific OvR components. The unique use of this approach is in combining many multicenter datasets with varying concentrations on what is needed into a single ensembles scheme[10].

To create an ensemble-based design, many models (or "classifiers") are combined. Multiple classifier systems, sometimes known as ensemble systems [11], are one common name for such setups. This section elaborates on some scenarios where ensemble-based approaches make mathematical sense. The steps used before a life-or-death medical diagnosis are similar to those in ensemble learning[12]. People often get second opinions from specialists before making major healthcare decisions, check out product reviews before buying (particularly expensive) airline tickets, and research prospective employers by checking out their references[13]. The ultimate verdict in each case will be reached by balancing the opinions of many knowledgeable people[14,15]. The major goal is to eliminate the potential for needless medical procedures, faulty products, and incompetent labour.

Voting Ensembles

Machine learning ensembles like the voting ensemble (also known as the "majority voting ensemble") pool the forecasts of many models into a single prediction. It's a method for making a model perform better, and it may be used to make a model perform better than the rest of the models in an ensemble. The ensemble voting procedure takes into account the results of many different models. It works well for both regression and classification problems. Estimating the typical sample predictions is what this means in regression. The sum of the votes for each possible label is then used to determine the final category.

Figure 1 depicts the several pre-trained, fine-tuned models that will serve as the basic learners.

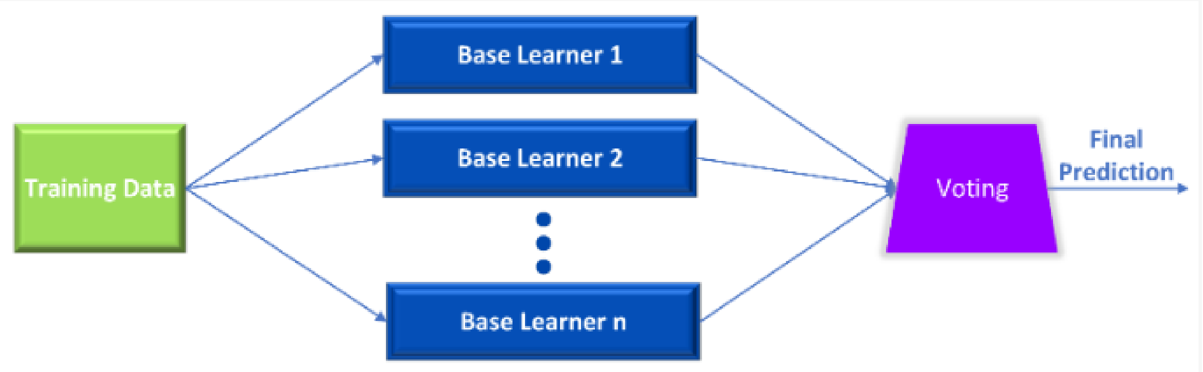


Figure 1. Architecture of Machine Based Learning

Literature Review

C M Suneera et.al (2020) Classifying texts into meaningful categories by using labels that have already been established is known as text classification. It finds practical use in fields as diverse as engineering, medicine, the biological sciences, the social sciences, the humanities, business, and government. Text categorization issues using labels have recently been amenable to machine learning and deep learning techniques. In this study, we compare the results of several popular machine learning and deep learning text categorization methods. To do this, we chose to analyse the performance of six machine learning algorithms utilising three distinct vectorization approaches and five deep learning algorithms. The 20 newsgroups dataset is used for all trials. Logistic Regression was shown to be superior to the other ML methods, and the Bi-channel Convolutional Neural Network model outperformed the other deep learning models in a number of fascinating ways.

Shovan Chowdhury et.al.,(2020) In this study, we use and compare several Machine Learning (ML) methods for identifying scholarly literature. The ultimate goal is to get useful information from abstracts that have been previously published. To do this, we use ML methods to categorise articles in the literature into the three broad categories of science, business, and social science. Support Vector Machines, Naive Bayes, K-Nearest Neighbour, and Decision Tree are the ML approaches used here. Methodology and methods for text recognition using these ML approaches are also supplied, in addition to a discussion of the ML algorithms used. The results of a comparison using four distinct performance metrics indicate that, with the exception of the Decision Tree method, the offered ML approaches with the extensive pre-processing algorithms are effective at categorising articles using just the abstract text.

Kaushika Pal et.al.,(2020) Since Indic languages were morphologically rich because too much information is fused in words, text classification in these languages presents fundamental issues in terms of reaching excellent accuracy. In this work, we show the results of an experiment we conducted to categorise Hindi poetry into three categories: Shringar, Karuna, and Veera. Morphologically rich languages make it harder to categorise emotions since poem content expresses mood and has feelings linked. In the current experiment, 122 manually-collected web documents were processed, and 122 documents containing only meaningful data were generated; then, features were extracted from the processed documents using the Bag of Words Model, and the resulting numerical representation of the features was passed into the Training model. Five different machine-learning classification methods—including two variants of each—are used for classification.

Text Mining

Information or knowledge of high quality may be mined from the massive amounts of unstructured data using a method called text mining. Text mining is a subset of data mining. Text, charts, and multimedia content are all examples of unstructured data that may be found in files and papers. Unstructured data does not conform to a standard format or representation. Because of the many anomalies and unknowns it creates, it is challenging for machines to comprehend. Data mining, machine learning, information retrieval, statistics, natural language processing, and computational linguistics are just some of the many disciplines that contribute to text mining. Text mining focuses on natural language text that is available in the form of semi-structured or unstructured data. Characters in a text document unite to create words, which may then be rearranged to make sentences and phrases. All of these grammatical features work together to represent the aforementioned classes of things, ideas, and interpretations. There are several actions to do in order to extract the useful data effectively. Text data collection, pre-processing of obtained data, text transformation, feature selection, pattern analysis, pattern extraction, result storage, and result interpretation are the main stages here. The text mining process steps are depicted in Figure 2.

DOI : https://doi.org/10.56452/7-2-547

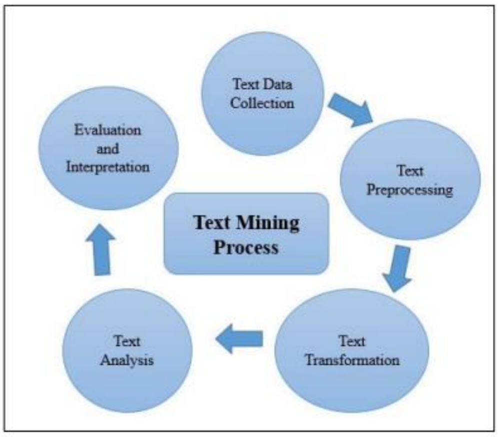


Figure 2 Text Mining Process

Text Clustering

Clustering documents according to their shared characteristics is called document clustering. The same unstructured format is likewise handled by the clusters that include the unstructured format. The massive amount of text documents has benefited greatly from its usage for efficient navigation, organisation, extraction, summarization, and retrieval. The primary goal of document clustering is to improve the precision of a document-based search engine. Clustering papers is an effective method for finding similar documents inside large collections. Clustering is currently used to search through a repository of documents and standardise the results returned by a search engine in response to a user's query. The relevant characteristics are extracted by text clustering, and the features are then represented in meaningful ways. Text mining systems portray documents as multi-dimensional high-semantic-valued documents. Automatic topic extraction, document organisation, and information retrieval are just a few of the many uses for document clustering. While much research has been done in the field of text clustering, there is always need for improvement in this area. The primary goal of this effort is to categorise or cluster the massive amount of text documents according to their subject matter. In this study, we present a new SCPSO technique for document clustering. The goal of this study is to use spectral clustering in tandem with swarm optimisation to handle large quantities of text texts. The randomization process is performed on the seed population while taking into account both global and local optimisation functions. The SCPSO algorithm incorporates key processes including similarity creation, swarm optimisation, and the clustering technique. The global convergence, computational complexity, and objective function handling will all be improved. Performances of the SCPSO method are compared to those of K-means, Bisecting K-means, Spherical K-means, the Expectation Maximisation algorithm, the Artificial Bee Colony, and Particle Swarm Optimisation. Various datasets, including benchmark datasets like Reuters and 20Newsgroup and TDT2 and BBC and actual datasets gathered from PCs, are used in the trials. Clustering Accuracy, Normalised Mutual Information, and the Adjusted Rand Index are employed as performance indicators. The suggested algorithm outperformed state-of-the-art methods in experimental tests, demonstrating its superiority. The proposed methodology is shown in Figure 3

DOI : https://doi.org/10.56452/7-2-547

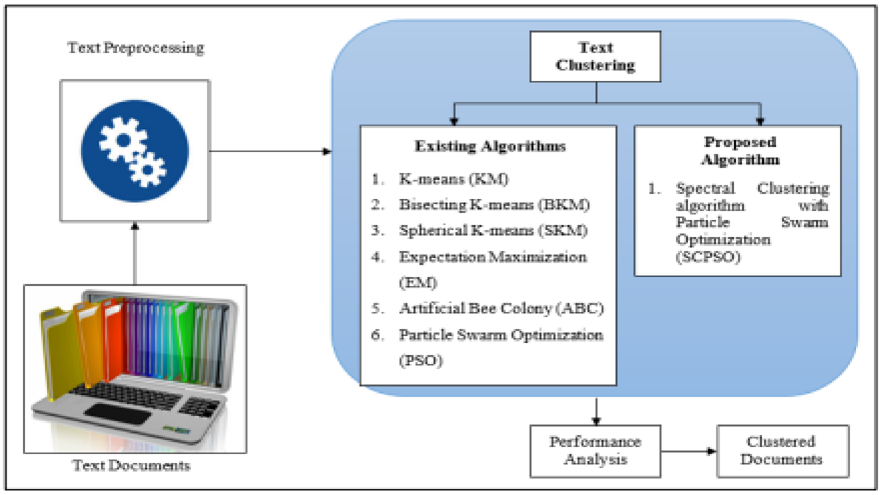
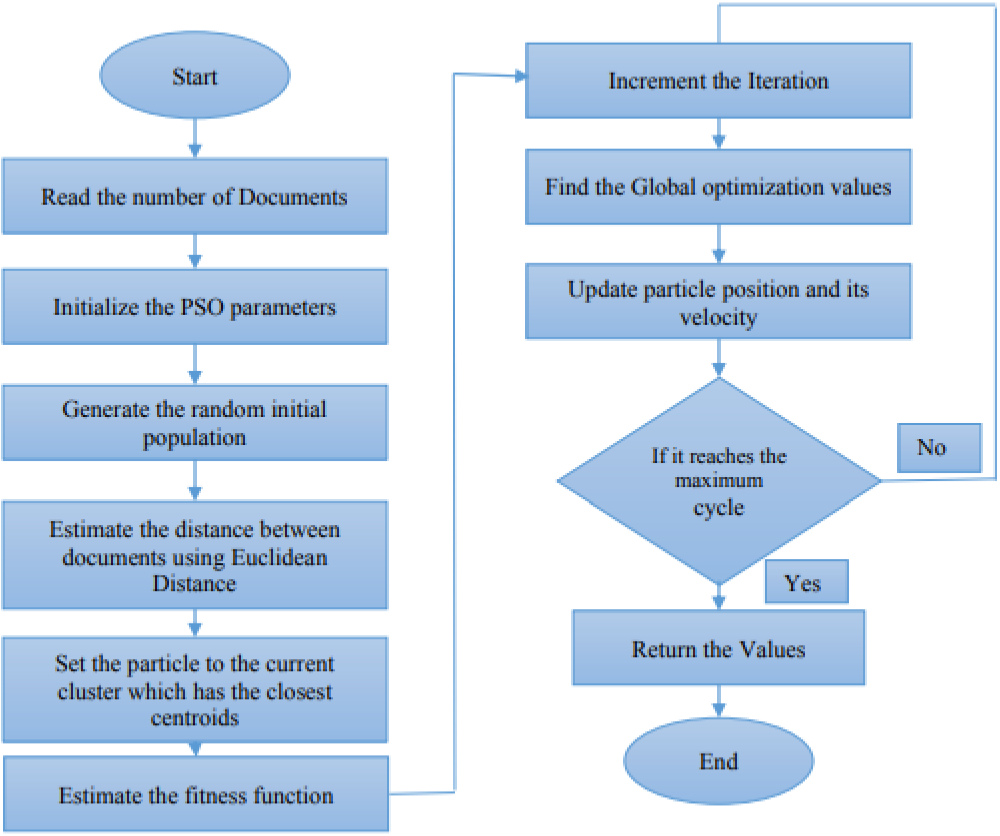

Text Preprocessing
Text Clustering
Existing Algorithms
1. K-means (KM)
2. Bisecting K-means (BKM)
3. Spherical K-means (SKM)
4. Expectation Maximization (EM)
5. Artificial Bee Colony (ABC)
6. Particle Swarm Optimization (PSO)
Proposed Algorithm
1. Spectral Clustering algorithm with Particle Swarm Optimization (SCPSO)
Text Documents
Performance Analysis
Clustered Documents

Figure 3. Methodology – Text Clustering

Particle Swarm Optimization

In this process, we search for the best possible answer. Particles, which are possible solutions, travel around the issue space at high speeds thanks to the constant iteration of the current optimal particle. Particles track their fitness-valued counterparts in a global problem space. The pbest is the name given to this particular number. This result is referred to as global best or gbest once the particle is ahead of all of the population as its neighbours. Each particle's speed is adjusted incrementally as it moves from the initial state to the pbest and best positions in the particle swarm optimisation concept.


Start
Read the number of Documents
Initialize the PSO parameters
Generate the random initial population
Estimate the distance between documents using Euclidean Distance
Set the particle to the current cluster which has the closest centroids
Estimate the fitness function
Increment the Iteration
Find the Global optimization values
Update particle position and its velocity
If it reaches the maximum cycle
No
Yes
Return the Values
End

Flow Chart 1: PSO Document Clustering

Datasets This stage receives as input the results from Text Transformation. The aforementioned steps were taken in the following document sets. The necessary paperwork for this stage is included in Table 1.

Table 1 Dataset Description

S.No	Dataset	Number of Documents from AAFNG phase
1	Reuters	16122
2	20 Newsgroup	15160
3	TDT2	55680
4	BBC	1662
5	Real	5020

Table 2 Performance Comparison on Text Clustering Techniques

Dataset	Performance	KM	BKM	SKM	EM	ABC	PSO	SCPSO
Reuters	ACC	0.524	0.538	0.612	0.679	0.708	0.765	0.842
	NMI	0.542	0.551	0.645	0.713	0.729	0.804	0.884
	ARI	0.519	0.526	0.607	0.648	0.699	0.721	0.811
20Newsgroup	ACC	0.534	0.601	0.675	0.714	0.754	0.728	0.832
	NMI	0.558	0.618	0.705	0.738	0.762	0.755	0.861
	ARI	0.526	0.599	0.628	0.694	0.721	0.713	0.806
TDT2	ACC	0.587	0.619	0.674	0.699	0.729	0.762	0.85
	NMI	0.591	0.634	0.709	0.732	0.75	0.806	0.894
	ARI	0.574	0.608	0.637	0.681	0.718	0.738	0.82
BBC	ACC	0.598	0.635	0.619	0.712	0.734	0.754	0.851
	NMI	0.614	0.647	0.638	0.728	0.766	0.771	0.874
	ARI	0.583	0.611	0.603	0.694	0.724	0.741	0.822
Real	ACC	0.601	0.651	0.662	0.684	0.719	0.759	0.837
	NMI	0.624	0.637	0.701	0.729	0.736	0.799	0.838
	ARI	0.6	0.628	0.612	0.675	0.707	0.734	0.819

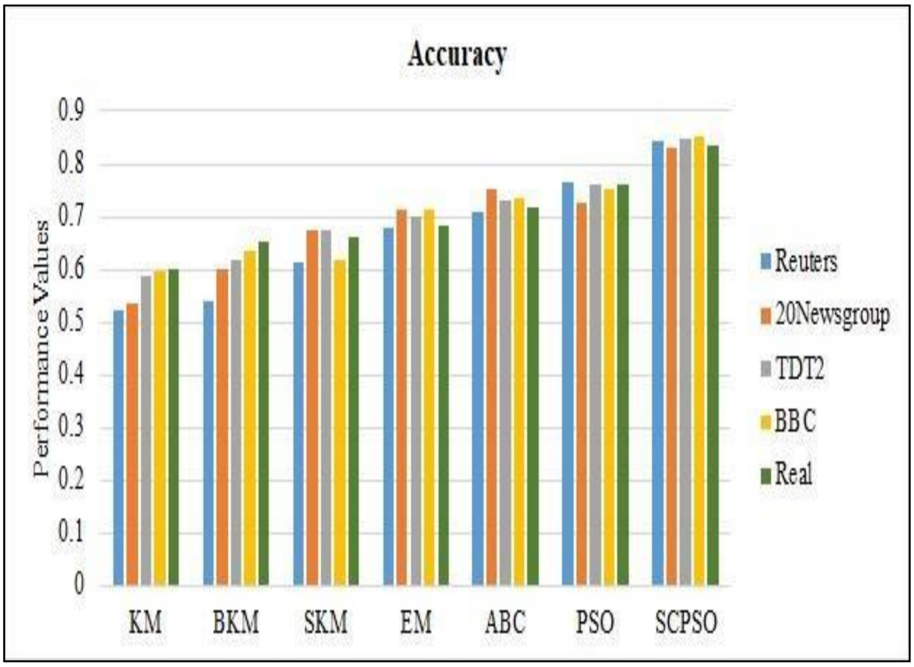


Figure 4 Text Clustering Accuracy Comparison

Conclusion

Researchers in the fields of text mining and information retrieval are now grappling with the challenge of document grouping and categorization. The end objective of this effort is to evaluate evolutionary algorithms and find strategies to boost their performance so that the best possible solution may be reached. In order to increase the precision of document clustering, a novel approach called Spectral Clustering with Particle Swarm Optimisation (SCPSO) was introduced in this study. When compared to conventional clustering techniques, the spectral clustering algorithm fares very well. The suggested SCPSO technique improves the accuracy of text clustering more so than spectral clustering. Even when dealing with a large number of documents, the suggested method performs optimally. The number of repetitions, parameter choice, particle setup, etc. all have a role in the execution's computing complexity and duration.

References

1. M. Suneera and J. Prakash, "Performance Analysis of Machine Learning and Deep Learning Models for Text Classification," *2020 IEEE 17th India Council International Conference (INDICON)*, New Delhi, India, 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342208.
2. S. Chowdhury and M. P. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques," *2020 Intermountain Engineering, Technology and Computing (IETC)*, Orem, UT, USA, 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249211.
3. K. Pal and B. V. Patel, "Automatic Multiclass Document Classification of Hindi Poems using Machine Learning Techniques," *2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154001.
4. K. E. Aydın and S. Baday, "Machine Learning for Web Content Classification," *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Istanbul, Turkey, 2020, pp. 1-7, doi: 10.1109/ASYU50717.2020.9259833.

DOI : https://doi.org/10.56452/7-2-547

5. I. D. Buldin and N. S. Ivanov, "Text Classification of Illegal Activities on Onion Sites," *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, St. Petersburg and Moscow, Russia, 2020, pp. 245-247, doi: 10.1109/EIConRus49466.2020.9039341.
6. X. Yang and X. Liu, "Convolutional Recurrent neural network with attention mechanism based improved skip-gram algorithm for text sentiment classification," *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*, Changsha, China, 2020, pp. 410-414, doi: 10.1109/ICISCE50968.2020.00092.
7. S. Kurniawan and I. Budi, "Indonesian Tweets Hate Speech Target Classification using Machine Learning," *2020 Fifth International Conference on Informatics and Computing (ICIC)*, Gorontalo, Indonesia, 2020, pp. 1-5, doi: 10.1109/ICIC50835.2020.9288515.
8. Y. Zheng, "An Exploration on Text Classification with Classical Machine Learning Algorithm," *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, Taiyuan, China, 2019, pp. 81-85, doi: 10.1109/MLBDBI48998.2019.00023.
9. S. A. Aboalnaser, "Machine Learning Algorithms in Arabic Text Classification: A Review," *2019 12th International Conference on Developments in eSystems Engineering (DeSE)*, Kazan, Russia, 2019, pp. 290-295, doi: 10.1109/DeSE.2019.00061.
10. W. Wang, G. He and X. Liu, "Text Multi-classification Based on Word Embedding and Multi-Grained Cascade Forest," *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, Chengdu, China, 2019, pp. 13-17, doi: 10.1109/ICCC47050.2019.9064153.
11. R. Keeling *et al*., "Empirical Comparisons of CNN with Other Learning Algorithms for Text Classification in Legal Document Review," *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 2038-2042, doi: 10.1109/BigData47090.2019.9006248.
12. Venkatesh and K. V. Ranjitha, "Classification and Optimization Scheme for Text Data using Machine Learning Naïve Bayes Classifier," *2018 IEEE World Symposium on Communication Engineering (WSCE)*, Singapore, 2018, pp. 33-36, doi: 10.1109/WSCE.2018.8690536.
13. Abdhullah-Al-Mamun and S. Akhter, "Social media bullying detection using machine learning on Bangla text," *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, Dhaka, Bangladesh, 2018, pp. 385-388, doi: 10.1109/ICECE.2018.8636797.
14. A. Mohasseb, M. Bader-El-Den, H. Liu and M. Cocea, "Domain specific syntax based approach for text classification in machine learning context," *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, Ningbo, China, 2017, pp. 658-663, doi: 10.1109/ICMLC.2017.8108983.
15. M. U. Salur, S. Tokat and İ. B. Aydilek, "Text classification on mahout with Naïve-Bayes machine learning algorithm," *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Turkey, 2017, pp. 1-5, doi: 10.1109/IDAP.2017.8090328.