IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 6 - V7I6-1203)
Available online at: https://www.ijariit.com

Role of Machine Learning in text classification – An extensive review

Amisha Srivastava
amisha18@iiserb.ac.in
Indian Institute of Science Education and Research,
Bhopal, Madhya Pradesh

Kushagra Jain
kushagrajain.is19@rvce.edu.in
RV College of Engineering,
Bengaluru, Karnataka

ABSTRACT

Cyberspace has elevated business insights and created a virtual space to store all forms of information online. Due to the rapid development in the online world, the usage of digital documents has increased because it is comfortable for the users to share, update or keep track of the records in one place without losing data. However, maintaining massive data does not suit optimal decision-making and is extremely expensive for storage, processing, and collection. There is a gigantic possibility that human annotators make errors while classifying data because of distraction, monotony, fatigue, and failure to meet the requirements. Once the text classification method uses machine learning approaches, the process will execute with fewer mistakes and more accuracy. The main goal of this review paper is to highlight and explain the role of different machine learning methodologies in text classification. Concurrently, this paper describes the challenges faced by other machine learning techniques and text representation. Furthermore, this review paper will provide an extensive survey on how various machine learning techniques such as Neural Networks, Naive Bayes, Logistic Regression, Random Forest, Decision Trees, and Support Vector Machine (SVM) - are implemented in Text classification.

Keywords—*Text classification, Support Vector Machine (SVM), Neural Networks, Naive Bayes, Logistic Regression, Random forest (RL), Decision Trees, Machine Learning Techniques, Sentiment Analysis*

1. INTRODUCTION

In recent years, the usage and storing of documents in online storage over the internet have expanded. Various decentralized record management systems and blockchain techniques were implemented in [1] to handle electronic medical records (EMR's) in the medical field. The Support Vector Machine (SVM) is used in classification [2] to find the binary classifier accuracy and maximize the limit of selected features. The classification problems in [2] include sentiment classification and cancer classification used for quick analysis and accurate diagnosis in medical data. Text classification is a method to extract the complementary information displayed according to document filtering, automatic metadata generation & classification, document filtering, word-sense disambiguation (WSD), Library catalogues, and other general applications.

1.1 Importance of text classification using machine learning approaches

The Text Mining method allows deriving high-quality data from standard text. Text classification is one of the significant functions of text mining that helps categorize text documents according to the predefined categories. This paper discusses the importance of text classification using machine learning approaches. The basic model of how text classification works is shown in fig 1.

- The Machine Learning Techniques and Data Mining allows to automatically discover and arrange the patterns from the E-documents (Electronic Documents).
- The text classification & machine learning (ML) techniques are used to analyze, comprehend, arrange, and sort the nature of the data.
- The Machine Learning (ML) approach allows to automatically analyze the data such as legal documents, patients record, surveys, social media, comments, chatbots, email that is hard to read and time-consuming.

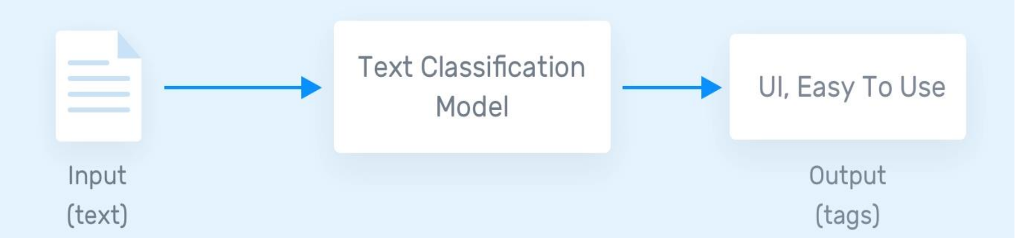

Fig. 1: Text Classification Model

- The ML approaches scrutinize the critical situations in Real-time analyzes and take prompt action.
- There is a massive possibility that human annotators make errors while classifying data because of distraction, monotony, fatigue, and failure to meet the requirements. Once the text classification method uses machine learning approaches, the process will execute with fewer mistakes and more accuracy.

1.2 Contributions of the review study

The contribution of the review paper focused on the classification techniques, selection methods, hybrid techniques, and statistical techniques used in the existing machine learning methodologies on text classification. Moreover, we compared the different machine learning methods and stated their advantages, disadvantages, and future ideas of the review study.

1.3 Organization of the paper

The remaining review paper has been organized as follows. Section II discusses the theoretical background of the text classification systems. Section III represents the papers reviewed from the techniques used, performance measures, merits, and demerits discussed. Section IV summarizes the research gaps presented in the field, and finally, Section V represents the paper's findings and suggestions for future direction.

2. THEORETICAL BACKGROUND

Social Media platforms likely allow the user to use emotions for transferring information to express their thoughts in the digital world. Moreover, the information exhibits on the internet through different mediums such as blogs, forums, social media platforms, documents, and much more. The different processes on various machine learning approaches with the existing algorithms suit the users' basic needs in the online world.

The Sentiment Analysis process (Arabic Dataset) used in [3], for natural language processing (NLP), Semantic Analysis in [5], and text mining (TM) to test English dataset; classify the opinions on positive, negative, and neutral expressions. In [3], machine learning algorithms such as Decision trees and Naive Bayes are used for classification methods. The algorithms have been implemented with additional techniques using the python language, demonstrated sentiment analysis, and eventually helped for better results. However, the existing algorithm in [3] faces issues when it comes to feeding the dataset. It is a time-consuming process and requires extensive work on the lexicon side. The SVM model in [12] can run millions of records in a short period. However, the existing model (SVM with Hadoop MAP) results in low accuracy on prediction and costs more to implement the model.

The test classification methods used to evaluate the personality test in [16], and the classification methods include Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbors to predict the results. Furthermore, the proven results have shown that Naive Bayes surpasses 60% accuracy compared to other classification methods. However, the method used in [16] resulted in less accuracy when compared to the previous method. Combining the three machine learning methods (MNB, KNN, SVM) procures 65% compared to the individual machine learning algorithm. Table 2.1 demonstrates the comparison of various machine learning techniques and the accuracy of the models.

Table 1: Accuracy of Machine Learning methods

Method	Accuracy
Multinomial Naive Bayes (MNB)	63%
K-Nearest Neighbors (KNN)	60%
Support Vector Machine (SVM)	61%
Combined (MNB, KNN, SVM)	65%

3. RELATED WORK

Text mining is a multidisciplinary field that relies on information processing, machine learning, computational linguistics, data mining, and statistical parameters. The Term Frequency - Inverse Document Frequency (TF-IDF) & singular value decomposition (SVD) dimensionality reduction methods are implemented in [6] using the k-means algorithm to cluster topics in the identical category. The implemented methodology stated in [6] does not improve accuracy, but the dimension minimized to 100 from 9,636 & 4,613. However, the accuracy should be improvised in other clustering documents too.

The various tokenization tools are used in [9] to analyze the performance of finding keywords in different languages. However, some tokenization tools only read and consider the limited texts. Moreover, it resulted that the NLP Dotnet tokenizer provides the best outcome compared to other tokenizer tools. Furthermore, the tokenization method is applicable to limited languages, so researchers should find more ways to use tokenization in other languages.

The support vector machine algorithm (SVM) is used in [11] to examine the information and recognize patterns in biomedical named entity recognition (NER) and the experimental results 84.24% in precision rate, and 80.76% in recall rate in GENIE corpus, which is higher compared to the previous research methodology.

The Chain Augmented Naive Bayes (CAN) techniques [15] are used in different text classification challenges. The CAN technique uses enhanced smoothing methods, and the model can work at word level or character level and provide the ability to handle Eastern and Western Languages. However, the proposed model offers limited labelled training data, which is not enough to analyze the performance and future machine learning techniques - implemented to improve the performance in text classification.

In today's modern world, personality plays an important role both on the internet and in the physical world. Social media has become an excellent medium for communication and extends the horizon of the online circle. The text classification allows the users to identify personalities based on the texts shared or written on a social platform like Twitter. In [16], the hybrid classification methods (KNN, SVM, MNB), used to make this personality test more accurate, have procured 65% rather than calculating the classification models individually. Check table 2.1 for the accuracy test on different classification models.

In [21], different machine learning methods are used to increase text classification performance and accuracy. However, sometimes due to high dimensionality, it is challenging to establish the classifier model. Moreover, removing the irrelevant data from data becomes difficult with the text classification methods. Various filter-based selection methods, machine learning methods, and transformation approaches are exploited for text classification to overcome these difficulties.

4. RESEARCH GAPS

The research gaps are identified based on other machine learning approaches used in text classification.

In [17], the Chinese web text collection - used to download the information automatically from the search engines. The Naive Bayes classification algorithm - implemented to identify the information which aligned in the same category. The experiment used word segmentation tools, training data, categorization datasets, and general evaluation approaches. However, the classification should be applied effectively in future work.

In [21], text classification becomes challenging because of irrelevant features, high dimensionality, and noisy features. Machine learning methodologies are used in text classification to increase accuracy and improve data quality. Moreover, Decision Tree algorithms are suitable only for fewer features in text classifiers and various reduction methods and feature selection techniques, applied in the different supervised and unsupervised machine learning methodologies to enhance the accuracy of the classifier. However, future research should focus on hybrid machine learning methods that help increase the performance of the classification method.

In [22], Multi-class classification techniques used various deep learning (DL) techniques include Deep Neural Network (DNN), Recurrence Neural Network (RNN), and Convolutional Neural Network (CNN) experimented with the massive dataset with 0.15 million labelled sentences and procured 84% accuracy in classifying and in event extraction of Urdu Language. The comparison of Deep Learning models - CNN, DNN, and RNN using one-hot-encoding has been shown in figure 2.

In [23], Single-layer Multi-size Filters Convolutional Neural Network (SMFCNN), proposed to evaluate the classification and differentiate the performance of Machine Learning (ML) techniques on different datasets. The performance has been evaluated in different ways. The experimental results using the SMFCNN approach achieved 95.4 % accuracy on the medium dataset, 91.8% accuracy on the large dataset, and 93.3% accuracy on the small dataset. However, the Urdu dataset used in the experiment is not publicly visible. Here as well, there is a need to implement hybrid text categorization (TC) techniques of Machine Learning (ML) & Deep Learning (DL) in the future.

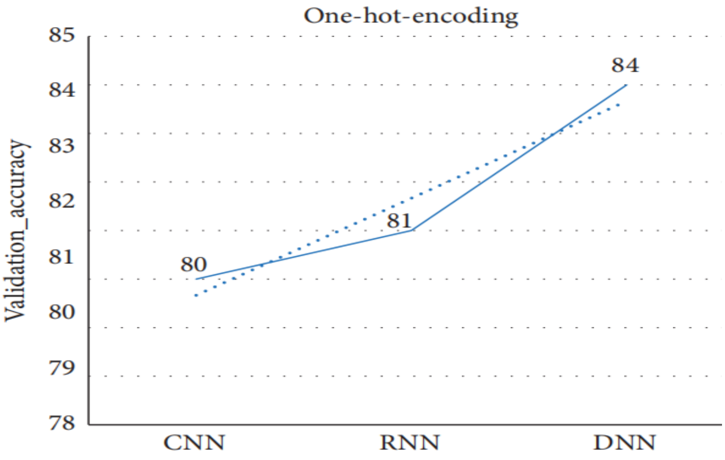

Fig. 2: CNN, DNN, and RNN using one-hot-encoding

In [24], the proposed approach, bidirectional long short-term memory (BiLSTM), is used in complex architectures to evaluate the efficiency in sentiment analysis and achieved 80.5% on MR datasets, 85.78% on SST2 datasets, and 90.58% on IMDb datasets on accuracy. But these approaches have not yet been implemented in real-time applications of sentiment analysis, so future work should be concentrated on the same.

5. CONCLUSIONS

Multiple machine learning algorithms are used in text classification to enhance the performance, automated parameters setting, and data accuracy. Unfortunately, many existing methods are not up to the level for calculating classification performance on different training sets, using different datasets in other languages. The problems in text mining should be appropriately addressed in future research works. The main goal of this review paper is to highlight and explain the role of different machine learning methodologies in text classification. Concurrently, this paper describes the challenges faced by other machine learning techniques and text representation.

Furthermore, this review paper provides an extensive survey on how various machine learning techniques such as Support Vector Machine, Naive Bayes, Logistic Regression, Random Forest, Decision Trees, and Neural Networks - are implemented in Text classification. Also, it elaborately explains the role of text classification using different machine learning and deep learning techniques in numerous fields and the challenges faced by the researchers. To conclude, researchers should propose hybrid machine learning algorithms to evaluate the performance & enhance the accuracy of the data on the world wide web.

REFERENCES

[1] A. Azaria, A. Ekblaw and T. Vieira, "Medrec: Using blockchain for medical data access and permission management," 2016 *2nd International Conference on Open and Big Data (OBD)*, IEEE, 2016.

[2] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *Eur. J. Oper. Res.*, vol. 265 (3), pp.993–1004, 2018.

[3] L. Al-Horaibi and M.B. Khan, S"entiment analysis of Arabic tweets using text mining techniques," First International Workshop on Pattern Recognition, International Society for Optics and Photonics, 2016.

[4] M. Bilal, H. Israr, M. Shahid and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naı¨ve Bayesian, Decision Tree and KNN classification techniques," *J. King Saud Univ.-Computer Informat. Sci.*, vol. 28 (3), pp.330–344, 2016.

[5] P.V. Ngoc, C. V. T. Ngoc, T. V. T. Ngoc and D. N. Duy, "A C4. 5 algorithms for English emotional classification," *Evolving Syst.* vol.10 (3), pp.425–451, 2019.

[6] A.I. Kadhim, Y.-N. Cheah and N.H. Ahamed, "Text document preprocessing and dimension reduction techniques for text document clustering," 2014 4th *International Conference on Artificial Intelligence with Applications in Engineering and Technology*, IEEE, 2014.

[7] E. Vellingiriraj, M. Balamurugan and P. Balasubramanie, "Information extraction and text mining of Ancient Vattezhuthu characters in historical documents using image zoning," 2016 *International Conference on Asian Language Processing* (IALP), IEEE, 2016.

[8] Z. Yao and C. Ze-wen, "Research on the construction and filter method of stop-word list in text preprocessing," 2011 *Fourth International Conference on Intelligent Computation Technology and Automation*, IEEE, 2011.

[9] S. Vijayarani and R. Janani, "Text mining: open source tokenization tools-an analysis," *Adv. Comput. Intell.: Int. J.*, vol.3(1), pp.37–47, 2016.

[10] A.K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manage.* vol.50(1), pp.104–112, 2014.

[11] Z. Ju, J. Wang and F. Zhu, "Named entity recognition from biomedical text using SVM," 2011 *5th international Conference on Bioinformatics and Biomedical Engineering*, IEEE, 2011.

[12] V.N. Phu, V.T.N. Chau and V.T.N. Tran, "SVM for English semantic classification in parallel environment," *Int. J. Speech Technol.*, vol.20(3), pp.487–508, 2017.

[13] N. Li and D.D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decis. Support Syst.*, vol.48(2), pp.354–368, 2010.

[14] C. Silva and B. Ribeiro, "On text-based mining with active learning and background knowledge using svm," *Soft. Comput.*, vol.11(6), pp.519–530, 2007.

[15] F. Peng, D. Schuurmans and S. Wang, "Augmenting naive bayes classifiers with statistical language models," *Inf. Retrieval.*, vol.7 (3–4), pp.317–345, 2004.

[16] B.Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," *2015 International Conference on Data and Software Engineering* (ICoDSE), IEEE, 2015.

[17] Z. Gong and T. Yu, "Chinese web text classification system model based on Naive Bayes," 2010 *International Conference on EProduct E-Service and E-Entertainment*, IEEE, 2010.

[18] F. Peng, D. Schuurmans and S. Wang, "Language and task independent text categorization with simple language models," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003.

[19] W. Chen, H. Shahabi, A. Shirzadi, T. Li, C. Guo, H. Hong, M. Ma, M. Xi and B.B. Ahmad, "A novel ensemble approach of bivariate statistical-based logistic model tree classifier for landslide susceptibility assessment," Geocarto Int., vol.33(12), pp. 1398–1420, 2018.

[20] A.J. Dobson and A.G. Barnett, *An Introduction to Generalized Linear Models*, CRC Press, 2018.

[21] B. Agarwal, and N. Mittal, "Text classification using machine learning methods-a survey," *In Proceedings of the Second International Conference on Soft Computing for Problem Solving* (SocProS 2012), December 28-30, 2012 (pp. 701-709). Springer, New Delhi. 2014.

[22] D. Ali, M. M. S. Missen, and M. Husnain, "Et Event Classification from Text," *Scientif. Program.*, vol.2021, 2021.

[23]M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," *IEEE Access*, vol.8, pp.42689-42707, 2020.

[24]Z. Hameed, and B. Garcia-Zapirain, "Sentiment classification using a single-layered BiLSTM model," *Ieee Access,* vol.8, pp.73992-74001, 2020.