Design and Analysis of News Category Predictor

Arif Hussain
Department of Electrical Engineering
Sukkur IBA University
Sukkur, Pakistan
arif.hussain@iba-suk.edu.pk

Gulsher Ali
Department of Electrical Engineering
Sukkur IBA University
Sukkur, Pakistan
gulsher@iba-suk.edu.pk

Faheem Akhtar
Department of Computer Science
Sukkur IBA University
Sukkur, Pakistan
fahim.akhtar@iba-suk.edu.pk

Zahid Hussain Khand
Department of Computer Science
Sukkur IBA University
Sukkur, Pakistan
zahid@iba-suk.edu.pk

Asif Ali
Department of Computer Science
Sukkur IBA University
Sukkur, Pakistan
asifali@iba-suk.edu.pk

***Abstract*—Recent technological advancements have changed significantly the way news is produced, consumed, and disseminated. Frequent and on-spot news reporting has been enabled, which smartphones can access anywhere and anytime. News categorization or classification can significantly help in its proper and timely dissemination. This study evaluates and compares news category predictors' performance based on four supervised machine learning models. We choose a standard dataset of British Broadcasting Corporation (BBC) news consisting of five categories: business, sports, technology, politics, and entertainment. Four multi-class news category predictors have been developed and trained on the same dataset: Naïve Bayes, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Each category predictor's performance was evaluated by analyzing the confusion matrix and quantifying the test dataset's precision, recall, and overall accuracy. In the end, the performance of all category predictors was studied and compared. The results show that all category predictors have achieved satisfactory accuracy grades. However, the SVM model performed better than the four supervised learning models, categorizing news articles with 98.3% accuracy. In contrast, the lowest accuracy was obtained by the KNN model. However, the KNN model's performance can be enhanced by investigating the optimal number of neighbors (K) value.**

Keywords-category predictor; Naïve Bayes; random forest; KNN; SVM; accuracy

I. INTRODUCTION

Technology has a significant impact on society and has significantly changed the way people access information. News is a well-known and standard service. Recent technological advancements have considerably changed the way news is produced, consumed, and disseminated. It has enabled more frequent and on-spot news reporting that smartphones can access anywhere and anytime. Therefore, people now expect to receive news of their interest in real-time. The news sources are already flooded with colossal information. Therefore, it is essential to automatically classify the news in specific categories based on the information content to allow timely and efficient information dissemination. Automatic Document Classification (ATC) can be used to efficiently manage text-based information (i.e. news) [1-3]. It allows timely and efficient information retrieval in the search phase. ATC can assign a relevant category to a news from a predefined set of reference categories based on the text feature extraction by correctly understanding the meaning and context of words. The time required to categorize the news correctly is directly proportional to the quantity of the text. In the newspaper's archive, the comprehensive range of articles varies from business to technology, so it is inconceivable that humans could manage this abundant content of information in a reasonable time frame. Manual document classification is cumbersome and resource-exhaustive.

The news category predictor aims to recognize and categorize different articles based on content/information type. The automatic news classification plays a vital role in processing a massive amount of articles. It can classify and label the news articles by analyzing the content (i.e. extracting feature values) to quickly access where they are focused in, allowing efficient and speedy news dissemination. Additionally, news websites can also increase their visibility by developing a recommendation system that suggests/ recommends relevant news to attract more attention. Several

Corresponding author: Faheem Akhtar

studies have been carried out to study modeling and performance evaluation of news category predictors using machine learning (ML) algorithms over different datasets (which differ in languages and range of categories) [4-10]. In these studies, well-known machine algorithms, such as Naïve Bayes (NB), SVM, Random Forest (RF), etc. are used to model news category predictors. The findings/results show that the category predictor's performance can vary with the machine algorithm deployed and the dataset used to train the model. In contrast, ML is envisioned to solve problems in various related domains [11, 12]. For a given ML algorithm, prediction performance can vary significantly depending upon the dataset. To quote a few, the NB algorithm's precision in categorizing news articles is reported to be 0.92 in [3] and 0.88 in [5]. In these cases, different datasets were used to train the same ML model, and the prediction performance is different. In the past few years, much research is carried out using different ML algorithms in natural language processing (i.e. text/news classification [13-17]). However, the current review paper is focused on evaluating and comparing category predictors' performance based on well-known ML algorithms. A standard BBC dataset was chosen having news of five categories: business, sports, technology, politics, and entertainment. This is a balanced dataset and quite different from traditional datasets that usually contain biases.

The main contribution of this research is that and it is the first time a multi-class news category predictor was developed by training four well-known machine learning algorithms (i.e. NB, RF, KNN, and SVM) on the same dataset. Each category predictor's performance was evaluated by analyzing the confusion matrix and quantifying the test dataset's precision, recall, and overall accuracy. Finally, the performance of all category predictors was studied and compared.

II. METHODOLOGY AND DATASET

The ultimate aim of this study is to classify the news into specific categories and analyze the performance of the category predictor. Initially, data are collected and preprocessed, then the content of text document (D_j) is converted into useful features ($w1_j$... wk_j) by feature extraction algorithms such as unigrams. The extracted features are transformed into numeric data that act as inputs for machine learning algorithms or classifiers (NB and RF). Finally, the ML models are trained on these transformed features, and the performance is evaluated on the test dataset. The research methodology/work flowchart is given in Figure .

A. Dataset

In this study, a BBC-originated news data set was used, which is obtained from Kaggle. It consists of 1490 documents from the BBC news website corresponding to stories in five typical areas: business, sports, technology, politics, and entertainment. The dataset is almost balanced: it contains an approximately equal portion of each class (Figure 2). However, most samples (23.2%) belong to the sports category, whereas the tech category has the least. The distribution of classes plays an important role in classification, and balanced datasets result in better learning models. In this study, the dataset is broken into 1,192 (80%) records for training and 298 (20%) testing.

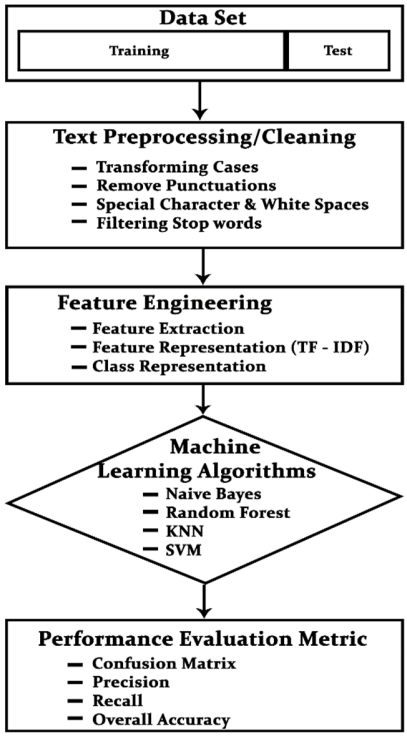

Fig. 1. Workflow.

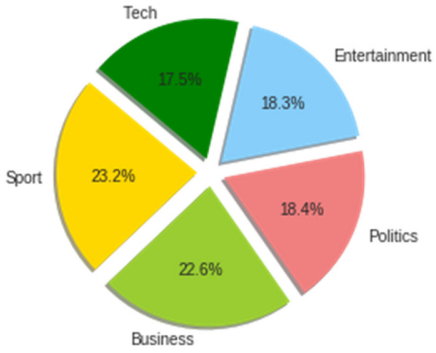

Fig. 2. Class distribution in the dataset.

B. Text Cleaning/Preprocessing

Text preprocessing or cleaning is a preliminary and crucial step of news classification, which reduces the required space and makes the classification more efficient [18]. Most of the times, the dataset is unstructured in combinations of useful and useless data. Unnecessary information such as stop words, punctuations, special characters, irrelevant sentences, quotations, and dates do not add any predictive power to the classifier/model. They only consume space and can distort the ML model. Therefore, before extracting any feature from the raw dataset, a cleaning process should be performed to minimize the distortions introduced to the model. In this paper, several steps have been followed to preprocess the news text.

1) Transforming Text

Transforming text in the same letter size (i.e. lower case) to eliminate homologous words that are different only in their case. For instance, the words "Fruit" and "fruit" are the very same in a real sense and should not be considered separately for prediction.

2) Removing Punctuations and Special Characters

Characters such as ?, !, ; and . are disposed of, this process simplifies computations in the next steps. Any special character and unnecessary whitespaces are also removed because they don't contribute to prediction power.

3) Filtering Stop Words

This technique is mainly used to remove unnecessary words or words with no specific meaning, such as "the", "an", "a", "what", etc. so that classifier cannot co-relate stop words and important class features. Furthermore, the most frequent or rarely used words do not contribute to the predictive power model. Therefore, they must be removed from the training set. In this study, we have downloaded a list of English stop words from the nltk library and then removed them from the dataset.

III. FEATURE ENIGINEERING/TEXT REPRESENTATION

The ML decision models (classifiers or regression algorithms) can only process and learn from numerical feature vectors. Therefore, the text features must be converted to a numeric representation [19, 20]. Feature engineering is a process to transform data (in this case, text) into numeric features that can act as inputs for ML algorithms. It involves two steps: feature extraction to extract the unique features/patterns and feature representation to represent each feature numerically. Bag of Words (BOW) and N-grams are commonly used techniques to create text features [21, 22]. BOW is the simplest feature extraction technique. It simply breaks apart the words in a document into individual word count statistics such as each word count/occurrence is used as a feature, but without considering the order. In contrast, N-grams is simply a sequence of *N* tokens (words) in the text. It not only finds the frequency of a word in the document but also considers the order and relationship between words. The N-grams can be Unigram (*N*=1), Bigram (*N*=2), or Trigram (*N*=3) depending upon the number of tokens taken into consideration. These extracted features values can be represented by different techniques such as Binary Representation (BR), Term Frequency (TF), Term Frequency Inverse Document Frequency (TF-IDF), and Normalized TF-IDF [23, 24]. In this study, the TF-IDF feature representation technique is used. TF-IDF technique disposes of the most common words and extracts only the most important feature words from the text [25]. The TF-IDF algorithm works on the principle that if a word (w_k) is more frequent in one document (j) and appears less frequently in a specific corpus, then it has a stronger ability to distinguish the category of texts, and it should be given more weight. The TF-IDF can be estimated by:

$$wk_j = tf_{kj} \times log\frac{N}{df_k} \quad (1)$$

where w_{kj} is the weight of the word k in the document j, N represent the total number of documents, tf_{kj} is the frequency of the word k in the document j and df_{kj} represents the number of documents containing the word (k).

IV. CLASS REPRESENTATION/ENCODING

News category prediction is multi-class classification. For instance, the dataset used in this study corresponds to five classes: business, sports, technology, politics, and entertainment. Each class is labeled to make it more understandable and often labeled in words. For ML models, label encoding is used to transform labels into numeric values. It can be done by a Label Encoder, which converts class labels into values between 0 and n-1, where n is the number of unique class labels. The actual and encoded labels of the dataset used in this study are given in Table I.

TABLE I. CLASS REPRESENTATION ENCODING

Labels	Encoded
Business	0
Entertainment	1
Politics	2
Sports	3
Tech	4

V. MACHINE LEARNING ALGORITHMS/CLASSIFIERS

A classifier is a ML model that maps input data to a proper category. In this study, NB, RF, KNN, and SVM algorithms are used to train a model that can classify news articles into categories.

A. Naïve Bayes

NB is a probabilistic classification algorithm based on Bayes' Theorem. It is simple, yet quite useful in a model, especially in text classification. The probability of any specific event is estimated by calculating its frequency in the past [26]. The fundamental NB assumption is that each feature is independent and unrelated to any other class feature. The Bayes theorem is:

$$p\,(C \mid f\,) = \frac{p\,(f \mid C) * p\,(C\,)}{p\,(f\,)} \quad (2)$$

$p\,(C \mid f\,)$ is the probability of occurrence of C given that event f has already occurred. The event f is termed as evidence, $p(C)$ is the prior probability of the class, $p\,(f \mid C)$ is termed as the likelihood and $p\,(C \mid f\,)$ is the posterior probability. In text classification, features can be numerous such as $f\,(f1, f2, f3, f4 \ldots\ldots fn)$ so by substituting f and expanding using the chain rule we get:

$$p\,(C \mid f_1, f_2\, n, \ldots\ldots fn\,) = \frac{p\,(f_1 \mid C\,) * p\,(f_2 \mid C\,) * \ldots . p\,(f_n \mid C\,) * p\,(C\,)}{p\,(f_1\,) * p\,(f_2\,) * \ldots p\,(f_n\,)} \quad (3)$$

Thus, we can find the category by finding the class with maximum probability.

B. Random Forest

Random Forest (RF) is a machine learning algorithm based on a set of trees classifiers [27]. The RF is an ensemble method used for classification that constructs several decision trees at training time and makes a final decision on majority voting. It uses bootstrap sampling in which data samples are sampled

independently and with the same distribution for all trees in the forest [26].

C. K-Nearest Neighbors

KNN is an intuitive supervised learning algorithm and an easy method to implement. It is used to classify objects based on their nearest examples in training sets space. The procedure to identify an object is classified by the majority vote of its neighbors like an object is assigned to a common class among its closest neighbors. The new vector classification is found by classes of its k-nearest neighbors where k is a positive integer. This algorithm is implemented using Euclidean distance metrics to detect the nearest neighbor [29-31]. The main challenge in KNN is to determine the optimal value of k. A higher value of k will increase the rise of over-learning so, it is necessary to take a valid value of k that reduces over-learning. The Euclidean distance metric $d(x, y)$ between two points is computed as:

$$d\ (x,y) = \sum_{i=1}^{N} \sqrt{(x_i^2) - (y_i^2)} \quad (4)$$

where N is the number of features like $x = \{ x_1 , x_2 , x_3 , \dots\dots , x_N \}$ and $y = \{ y_1 , y_2 , y_3 , \dots\dots , y_N \}$. The number of k-neighbors used to test a new vector varies from 1 to 10.

D. Support Vector Machine

The SVM is a kernel-based ML algorithm that can categorize input data input into specific classes or categories. SVM constructs a classifier that makes the decision boundary for every class and defines the hyper-plane to linearly or non-linearly separate them. The accuracy of categorization can be increased by increasing the hyper-plane margin that also enlarges the distance among classes. Hence, the farthest hyper-plane provides more immunity against noise. SVM is a kernel-based classifier that defines the process of mapping the training data set to develop its similarities to a linearly independent data set. The main reason to use mapping is to enhance the depth of the data set done by kernel function like some commonly used kernel are linear, RBF, and quadratic, etc. [32,33].

VI. PERFOMANCE EVALUATION METRICS

To precisely gauge the performance of the category predictor, there are different performance evaluation techniques and metrics such as Confusion Matrix, Accuracy, Precision, Recall or Sensitivity, and F1-Score. In this study, the Confusion Matrix is evaluated first, and Accuracy, Precision, and Recall are analyzed to get a true insight of the prediction performance. The Confusion Matrix is a table that is often used to quantify the performance of a category predictor or classification model on a set of test data for which the true/actual values are unknown. The confusion matrix summarizes the prediction performance by quantifying the correct and incorrect predictions (misclassification) broken down into each class. It gives a detailed insight into ways in which category predictor is confused while classifying the input data. Therefore, a confusion matrix is a good option for evaluating the performance of the multi-class category predictor.

VII. RESULTS AND DISCUSSION

The performance results of multi-class category predictors based on different supervised learning models are evaluated and compared in this section. This study's learning models are NB, RF, KNN, and SVM. The evaluation was done by observing each category predictor's prediction results by analyzing the Confusion Matrix and quantifying Precision, Recall, and overall Accuracy. This analysis was made on a test dataset consisting of 298 news samples. The Confusion Matrices of each category predictor are given in Figure 3.

Figure 3 shows the confusion matrix of NB (a), RF (b), KNN (c), and SVM (d). In general, every category predictor has achieved good accuracy. SVM based category predictor achieved the highest accuracy for the given dataset. The SVM based category predictor predicts the news category with an accuracy of 98.3%, and this can be verified from Figure 3(d), as it shows only five wrong predictions out of 298 samples. The NB model's performance was observed to be as good as SVM's with an accuracy of 97.3%. In the NB model, the most misclassified category was Technology with four incorrect predictions, while the most accurately classified category was sports having no wrong predictions. The detailed category/class wise analysis of Precision and Recall for SVM and NB is given in Table II and Table III, respectively.

TABLE II. PERFORMANCE EVALUTION OF CATEGORY PREDICTOR USING SVM

Category	Precision	Recall
Business	0.98	0.98
Entertainment	0.97	0.98
Politics	0.98	0.98
Sports	1.00	1.00
Tech	0.98	0.96

TABLE III. PERFORMANCE EVALUTION OF CATEGORY PREDICTOR USING NAÏVE BAYES

Category	Precision	Recall
Business	0.97	0.98
Entertainment	0.97	0.97
Politics	0.96	0.98
Sports	1.00	1.00
Tech	0.96	0.92

If we analyze the category predictor model based on RF algorithm, the prediction performance is satisfactory. However, there more incorrect predictions as compared to SVM and NB. The most misclassified category was business with seven inaccurate predictions, while the most accurately classified category was sports. RF model achieved an Accuracy of 94.9%, with 283 correct predictions out of 298 test samples. The detailed category/class wise analysis of Precision and Recall for the NB model is given in Table IV. KNN based category predictor achieved the lowest accuracy. It achieved an accuracy of 94.2% with 17 wrong predictions, as shown in Figure 3(c). The detailed category/class wise analysis of Precision and Recall for the KNN model is given in Table V. Although KNN has achieved the lowest accuracy grades as compared to SVM, NB, and RF, its performance is still consideredsatisfactory.

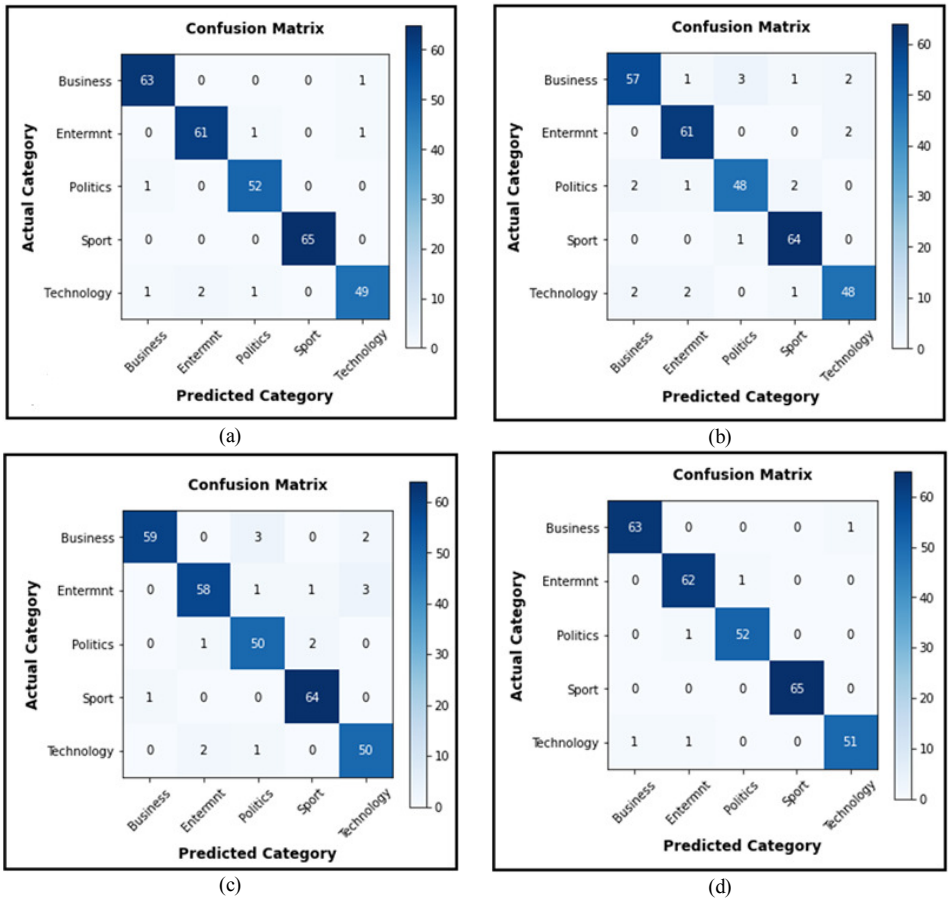

Fig. 3. Confusion Matrices of multi-class category predictors: (a) NB, (b) RF, (c) KNN, (d) SVM.

TABLE IV. PERFORMANCE EVALUTION OF CATEGORY PREDICTOR USING RANDOM FOREST

Category	Precision	Recall
Business	0.93	0.89
Entertainment	0.94	0.97
Politics	0.92	0.91
Sports	0.94	0.98
Tech	0.92	0.91

TABLE V. PERFORMANCE EVALUTION OF CATEGORY PREDICTOR USING KNN

Category	Precision	Recall
Business	0.92	0.89
Entertainment	0.95	0.92
Politics	0.91	0.94
Sports	0.96	0.98
Tech	0.91	0.94

VIII. CONCLUSION AND FUTURE WORK

This paper presents a comparative analysis of the multi-class category predictor's prediction performance. News category predictors were developed by deploying/training well-known machine learning algorithms (Naïve Bayes, Random Forest, K-Nearest Neighbor, and Support Vector Machine) on a BBC news dataset having five categories (business, sports, technology, politics, and entertainment). Later, using performance evaluation metrics, we analyzed the Confusion Matrix and quantified the test dataset's Precision, Recall, and overall Accuracy. As a result, the SVM model was proven to be the best among the four supervised learning models in correctly categorizing news articles with 98.3% accuracy. The lowest accuracy was obtained by the KNN model with K=5. However, the KNN model's performance can be enhanced by investigating the value of the optimal number of neighbors K. As future work, deep learning schemes will be introduced to further improve the classifier performance.

REFERENCES

[1] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Yogyakarta, Indonesia, Oct. 2014, doi: 10.1109/ICITEED.2014.7007894.

[2] G. Mujtaba, L. Shuib, R. G. Raj, R. Rajandram, and K. Shaikh, "Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study," *Journal of Forensic and Legal Medicine*, vol. 57, pp. 41–50, Jul. 2018, doi: 10.1016/j.jflm.2017.07.001.

[3] V. S. Padala, K. Gandhi, and D. V. Pushpalatha, "Machine learning: the new language for applications," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 4, pp. 411–421, Dec. 2019, doi: 10.11591/ijai.v8.i4.pp411-421.

[4] F. Miao, P. Zhang, L. Jin, and H. Wu, "Chinese News Text Classification Based on Machine Learning Algorithm," in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, China, Aug. 2018, vol. 02, pp. 48–51, doi: 10.1109/IHMSC.2018.10117.

[5] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," in *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, Coimbatore, India, Mar. 2016, pp. 112–116, doi: 10.1109/ICETECH.2016.7569223.

[6] G. L. Yovellia Londo, D. H. Kartawijaya, H. T. Ivariyani, Y. S. Purnomo W. P., A. P. Muhammad Rafi, and D. Ariyandi, "A Study of Text Classification for Indonesian News Article," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, Yogyakarta, Indonesia, Mar. 2019, pp. 205–208, doi: 10.1109/ICAIIT.2019.8834611.

[7] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, and O. R. Rusli, "News Article Text Classification in Indonesian Language," *Procedia Computer Science*, vol. 116, pp. 137–143, 2017, doi: 10.1016/j.procs.2017.10.039.

[8] A. N. Chy, Md. H. Seddiqui, and S. Das, "Bangla news classification using Naive Bayes classifier," in *16th Int'l Conf. Computer and Information Technology*, Khulna, Bangladesh, Mar. 2014, pp. 366–371, doi: 10.1109/ICCITechn.2014.6997369.

[9] I. Dilrukshi, K. De Zoysa, and A. Caldera, "Twitter news classification using SVM," in *2013 8th International Conference on Computer Science Education*, Colombo, Sri Lanka, Apr. 2013, pp. 287–291, doi: 10.1109/ICCSE.2013.6553926.

[10] H. Sawaf, J. Zaplo, and H. Ney, "Statistical classification methods for arabic news articles," presented at the Natural Language Processing in ACL2001, Toulouse, France, 2001.

[11] I. J. Mrema and M. A. Dida, "A Survey of Road Accident Reporting and Driver's Behavior Awareness Systems: The Case of Tanzania," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 6009–6015, Aug. 2020.

[12] M. Kiruthika and S. Bindu, "Classification of Electrical Power System Conditions with Convolutional Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 10, no. 3, pp. 5759–5768, Jun. 2020.

[13] N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, "Performance analysis of supervised learning models for product title classification," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 3, pp. 228–236, Dec. 2019, doi: 10.11591/ijai.v8.i3.pp228-236.

[14] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *Journal of Advances in Information Technology*, vol. 1, no. 1, pp. 4–20, Feb. 2010.

[15] V. Ashwin, "Twitter Tweet Classifier," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 5, no. 1, pp. 41–44, Mar. 2016, doi: 10.11591/ijai.v5.i1.pp41-44.

[16] M. A. Al-Hagery, "Extracting hidden patterns from dates' product data using a machine learning technique," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 3, pp. 205–214, Dec. 2019, doi: 10.11591/ijai.v8.i3.pp205-214.

[17] A. Ferrario and M. Naegelin, "The Art of Natural Language Processing: Classical, Modern and Contemporary Approaches to Text Document Classification," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3547887, Mar. 2020. doi: 10.2139/ssrn.3547887.

[18] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[19] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, no. null, pp. 1157–1182, Mar. 2003.

[20] M. Farhoodi, A. Yari, and A. Sayah, "N-gram based text classification for Persian newspaper corpus," in *The 7th International Conference on Digital Content, Multimedia Technology and its Applications*, Aug. 2011, pp. 55–59.

[21] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, Mar. 2002, doi: 10.1145/505282.505283.

[22] S. R. Basha, J. K. Rani, J. J. C. Prasad Yadav, and G. Ravi Kumar, "Impact of feature selection techniques in Text Classification: An Experimental study," *Journal of Mechanics of Continua and Mathematical Sciences*, no. 3 special, Sep. 2019, doi: 10.26782/jmcms.spl.3/2019.09.00004.

[23] F. Debole and F. Sebastiani, "Supervised Term Weighting for Automated Text Categorization," in *Text Mining and its Applications*, S. Sirmakessis, Ed. Berlin, Heidelberg: Springer, 2004, pp. 81–97.

[24] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[25] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," in *Proceedings of the first instructional conference on machine learning*, Accessed: Oct. 06, 2020. [Online]. Available: https://www.semanticscholar.org/paper/Using-TF-IDF-to-Determine-Word-Relevance-in-Queries-Ramos/b3bf6373ff41a115197cb5b30e57830c16130c2c.

[26] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Machine Learning: ECML-98*, 1998, pp. 4–15, doi: 10.1007/BFb0026666.

[27] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, May 1991, doi: 10.1109/21.97458.

[28] B. Xu, X. Guo, Y. Ye, and J. Cheng, "An Improved Random Forest Classifier for Text Categorization," *Journal of Computers*, vol. 7, no. 12, pp. 2913–2920, Dec. 2012, doi: 10.4304/jcp.7.12.2913-2920.

[29] P. Soucy and G. W. Mineau, "A simple KNN algorithm for text categorization," in *Proceedings 2001 IEEE International Conference on Data Mining*, San Jose, CA, Nov. 2001, pp. 647–648, doi: 10.1109/ICDM.2001.989592.

[30] R. Palaniappan, K. Sundaraj, and S. Sundaraj, "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals," *BMC Bioinformatics*, vol. 15, p. 223, Jun. 2014, doi: 10.1186/1471-2105-15-223.

[31] S. R. Basha and J. K. Rani, "A Comparative Approach of Dimensionality Reduction Techniques in Text Classification," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 4974–4979, Dec. 2019.

[32] A. K. Mourya, S. U. Ahsaan, and H. Kaur, "Performance and Evaluation of Different Kernels in Support Vector Machine for Text Mining," in *Advances in Intelligent Computing and Communication*, Singapore, 2020, pp. 264–271, doi: 10.1007/978-981-15-2774-6_33.

[33] J.-Y. Jiang, R.-J. Liou, and S.-J. Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 335–349, Mar. 2011, doi: 10.1109/TKDE.2010.122.