**Classification of News Articles Using NLP Techniques and Machine Learning Algorithms: Comparative Study**

Ali M. Ahmed Al-Sabaawi, Hanan S. Mohamad Zakariya, Raheed F. Putrus Bizo,Shahad A. Abdalhaleem Al-MashhadaniSoftware Dept., Faculty of Information Technology, Nineveh University, Mosul, Iraq

[ali.mohsin@uoninevah.edu.iq](mailto:ali.mohsin@uoninevah.edu.iq)

**Abstract.** News primarily aids people in recognizing the events that have occurred in their surroundings. Since news is essential as a venue for social interaction, daily newspapers, whether online or in print, emphasize news. In contemporary times, the majority of individuals access news through online platforms. However, not all readers engage with every news category; instead, each reader tends to exhibit preferences for specific categories of interest. Consequently, the news category is crucial. Getting all the news from a single news website is really challenging. Occasionally, it might not have the category the user wants. To overcome these challenges, news articles are systematically categorized using advanced classification techniques. Accordingly, this paper presents multiple classification methods to categorize news articles into distinct categories, including sports, economy, politics, and business. The dataset undergoes preprocessing by eliminating punctuation, spaces, symbols, and stop words. Subsequently, feature extraction is performed using two techniques: Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BOW). Following this, three machine learning algorithms—Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN)—are employed to classify the news articles into their respective categories. A standard dataset called AG news classification is used in this study. The experimental results revealed that the neural network is outperformed in most metrics.

**Keywords:** Classification, Natural Language Processing, machine learning, SVM, RF, NN.

1. **Introduction**

News plays a vital role in educating, entertaining, and advertising, with its main purpose being to keep individuals updated on events that could affect their lives. It addresses critical topics across politics, business, sports, technology, and health, ensuring people have access to essential information. Breaking news, often uncategorized, is delivered promptly due to its high priority. People’s preferences for news vary; some focus on technology, others on business, and some read for general knowledge, exam preparation, or leisure. News websites usually organize content into categories, though some unclassified news is also present. Moreover, categorizing news enables users to access their preferred news in real-time, helping them save time [1].

Unclassified news creates confusion for readers, making it difficult to locate similar content without proper categorization of news articles [1]. This challenge has prompted researchers to explore various methods for classifying news. Manual classification is impractical, especially with large datasets, as it becomes overly complex [2]. Earlier research has employed machine learning for text classification, enabling news to be sorted into relevant domains [3-6]. The study [4] compared widely-used machine learning techniques, including Naive Bayes, SVM, and Neural Networks, to address the challenge of automatically categorizing Nepali news. It proposed machine learning strategies for this task and conducted a thorough analysis of the algorithms' performance. Through experiments with different optimization parameters and the grid search method, the study determined that SVM performed better than Multilayer Perception, Neural Networks, and Naive Bayes, achieving an accuracy rate of 74.65%. Ref [7] explored an automated system for classifying Bangla news documents, highlighting the model's high accuracy. Additionally, the study emphasized that using a stemmer and removing stop words are essential for improving accuracy in Bangla news document classification. Ref [8] focused on multi-label classification for Arabic articles, evaluating three classifiers: random forest, decision tree, and k-nearest neighbors (KNN). The results revealed that the decision tree classifier performed better than both random forest and KNN. Their experiments consistently showed the decision tree's superior performance in this task. Study [9] used Support Vector Machine (SVM) to classify Indonesian news articles. The study aimed to utilize a one-pass clustering approach for news categorization, using a dataset of 160 articles. The experiments achieved an average classification accuracy of 85% with the SVM model.

Upon reviewing prior research, several limitations can be identified. Firstly, earlier studies predominantly relied on either the news title or description for classification purposes, potentially overlooking the comprehensive use of both elements. Secondly, these studies primarily utilized the Term Frequency-Inverse Document Frequency (TF-IDF) method for feature extraction, neglecting alternative techniques that could enhance accuracy [10, 11]. Lastly, the datasets employed in these studies were often insufficient in size, which may have hindered the precision and reliability of the classification results [12]. Therefore, this study addresses these limitations by incorporating both the description and title of news articles for classification. Additionally, two feature extraction methods TF-IDF and BoW are employed to comprehensively extract and analyze text features, aiming to enhance the accuracy and robustness of the classification process [13, 14].

**2. Statement of the Problem**

Our research's goal is to use machine learning to automatically classify news articles according to their domain. Fig. 1 illustrates our methodology.



**Fig. 1.** Methodology

**2.1 Data Cleaning and Preprocessing**

We cleaned and prepared our data for processing by following these steps, duplicates removal, tokenization, lowercasing, punctuation, special characters removal and vectorization. Fig. 2 shows the steps of data cleaning and pre-processing.



**Fig. 2.** data cleaning and pre-processing

The duplicates removal removes duplicated words from each article, the lowercasing converts all the letters in the article to lowercases, punctuation and special characters removal remove any character that is not an English letter, the tokenization is a process of breaking the articles into tokens and the vectorization convert the textual data into numerical representations.

**2.2 Word Vectorization**

The process of generating new features from unprocessed data in order to enhance the prediction expertise of the learning algorithm. It transforming raw data into meaningful input that improve the performance of a machine learning model. It involves selecting, creating, and modifying data to enhance predictive accuracy, determining whether a model is good or bad. Performed using TF-IDF and BoW.

1. **TF-IDF (Term Frequency (TF) — Inverse Document Frequency (IDF):** is a numerical statistic method that reflects the importance of a word in a documentIt considers two primary factors: a word's frequency within a document (TF) and its frequency throughout the corpus of documents (IDF). Generally speaking, we give each word a score that represents its significance within the document and corpus.

The Term Frequency (TF) indicates how often it occurs in a document. It is computed by taking the total number of words in a document and dividing that number by the number of times a phrase appears. The result is a value between 0 and 1.

The Inverse Document Frequency (IDF) is a measure of how important a term is across all documents. It is calculated by taking the logarithm of the total number of documents in the corpus divided by the number of documents in which the term appears. The result value is a number greater than or equal to 0.

The equations below denote the TF-IDF vectorization.

TF = (1)

IDF = log (2)

TF-IDF = TF x IDF (3)

1. **BoW (Bag of Words):** is a simple method and was widely used technique in natural language processing (NLP) for representing textual data.

It ignores context, word order, and grammar in instead of treating a document as a collection of words. The term "Bag of Words" conveys the notion that the approach depicts a text as an unarranged "bag" or collection of words. It is referred to as a "bag of words" since it ignores any information describing the structure or sequence of words and just considers whether or not known terms appear in the document, not where they appear.

The Bag of Words method convert text into fixed-length vectors by counting how many times each word appears. (A data modeling technique). It involves two steps:

1. A vocabulary of known words.
2. A measure of the presence of known words.

**2.3 Classification using training algorithms**

To group the news articles in our research into significant classes, we use multi-class classification. The prediction model was developed to forecast news stories according to their domains of relevance. We used Support Vector Machine, Random Forest, and Neural Network classification algorithms to classify the news articles based on the relevant category. News stories are categorized by the prediction model according to their pertinent domains as world, sports, business and sci/tech.

1. **Support Vector Machine (SVM):** This technique works very well for handling multi-dimensional data, such as text representation vectors. Because it can classify text quickly and efficiently, it is also regarded as the most accurate classifier for this type of task. It works by mapping data to a high-dimensional feature space so that the data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong.

The equation for the linear hyperplane can be written as:

(4)

Where:

* is the normal vector to the hyperplane (the direction perpendicular to it).
* is a vector containing the features of a single data point.
* is the offset or bias term, representing the distance of the hyperplane from the origin along the normal vector *w*.

To find the hyperplane that maximizes the margin:

(5)

Subject to the constraint:

Where:

* is the class label (+1 or -1) for each training instance.
* is the feature vector for the i-th training instance.
* is the total number of training instances.

The condition   ensures that each data point is correctly classified and lies outside the margin.

The distance between a data point x and the decision boundary can be calculated as:

(6)

Where:

||w|| represents the Euclidean norm of the weight vector . Euclidean norm of the normal vector .

Since our dataset consists of 4 classes and SVM is primarily designed for binary classification, we employed the One-vs-One (OvO) approach.

The One-vs-One (OvO) method, a binary classifier is constructed for each potential pair of classes. This method is particularly advantageous when the number of classes is relatively small, as it allows for a focused comparison between pairs of classes. Each classifier is responsible for distinguishing between two specific classes, effectively ignoring the others.

1. **Neural Network:** is an algorithm that describes how the human brain processes information. It predicts and propagates a result by analysing the values or data received in its input layer, simulating the human brain. The second layer receives the data from the input layer and forwards it to the subsequent hidden levels. The neurons or nodes in the second layer aggregate the data after identifying and filtering highly relevant patterns. The input weight is changed by assigning a weight to each input value. A logistical or sigmoid function defines and sums up these resulting values. The output of the preceding layer is examined and processed in later hidden layers before being passed on to the following layer. Then, in the output layer, values are recombined to achieve and propagate the result.

A Multilayer Perceptron (MLP) model, which is a type of Feedforward Neural Network (FNN), was used for classification and regression tasks. The network consists of two hidden layers, containing 256 and 128 neurons, respectively.

To compute the weighted sum of the inputs:

(7)

Where, is the input feature, is the corresponding weight and  is the bias term.

The Rectified Linear Unit (ReLU) activation function was used:

(8)

to calculate the loss binary cross-entropy loss function was used:

(9)

 indicates the actual label,  refers to the predicted label, and  is the number of samples.

The Adam optimizer was used to minimize the loss function:

First moment (mean) estimate:

(10)

Second moment (variance) estimate:

(11)

and ​ are the moment estimates, ,​ are indicate rates, and represents the gradient at time

**C. Random Forest:** is a collection of decision trees that are trained on different subsets of the data and features. Each tree makes a prediction based on its own rules, and the final output is the average or majority vote of all the trees. Effectively, it fits a number of decision tree classifiers on various subsamples of the dataset. When creating each tree, the algorithm randomly selects a subset of features or variables to split the data rather than using all available features at a time. This adds diversity to the trees. This algorithm is currently one of best performing algorithms for many classification problems.

Calculate Gini index to decide how nodes on a decision tree branch.:

(12)

represents the number of classes, ​ is the proportion of samples of class in node .

To calculate the weighted Gini index:

(13)

​ is the number of samples in the left node, ​ refers to the number of samples in the right node, and is the total number of samples.

To calculating the feature importance:

(14)

**3. Results and Discussions**

In this section, the evaluation of the three models is carried out. Three different classification algorithms are achieved, namely NN, SVM and RF. Four evaluation metrics are utilized to measure the performance of the proposed methods, namely precision, recall, F1-measure and time.

**3.1 Dataset**

AG News Classification Dataset is utilized in this research. It involves 4 classes: World, Sports, Business or Sci/Tech. Dataset has 120,000 records they're divided into two parts 80% for training consists of 96,000 samples and 20% for testing consists of 24,000 samples, evenly distributes across the four classes.

**3.2 Evaluation Measurements**

The evaluation is based on 5 metrics, are the precision, recall, F1-measure, accuracy and the time to calculate the efficiency of the algorithms that are utilized in this research. The time represents the timing for each algorithm to finish its training. The ratio of accurately categorized articles to all articles utilized for classification is known as the accuracy rate. The model's recall, sometimes referred to as its true positive rate, is its capacity to locate every pertinent instance within a dataset. It is calculated as , where TP is number of True Positive and FN is the number of False Negatives. In this analysis, it is calculated as News correctly classified as:

(15)

Precision is the percentage of pertinent cases that are truly pertinent. It is calculated as , where FP is number of False Positives. It is computed as follows:

(16)

Maximizing precision comes at the expense of recall, and vice versa. In order to get around this restriction, the F1 measure was implemented. It is computed using the following formula and represents the harmonic mean of precision and recall:

(17)

**3.3 Experimental results**

The table below provides a comprehensive comparative analysis of news classification performance. It is structured into two sections based on feature extraction techniques: TF-IDF and BOW. Each section includes three primary columns representing the classification algorithms employed. The evaluation metrics considered in this analysis are precision, recall, and F1-score, all of which are computed for each algorithm. Additionally, the average accuracy rate is derived by calculating the mean F1-score across all categories. Lastly, the training time for each algorithm is measured in minutes to assess computational efficiency.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TF-IDF | SVM | | | Neural Network | | | Random Forest | | |
| Topics | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| World | 0.92 | 0.90 | 0.91 | 0.92 | 0.90 | 0.90 | 0.90 | 0.88 | 0.89 |
| Sports | 0.95 | 0.97 | 0.96 | 0.93 | 0.95 | 0.95 | 0.90 | 0.96 | 0.93 |
| Business | 0.87 | 0.87 | 0.87 | 0.86 | 0.87 | 0.86 | 0.85 | 0.84 | 0.85 |
| Sci/Tech | 0.88 | 0.87 | 0.88 | 0.87 | 0.85 | 0.86 | 0.86 | 0.83 | 0.85 |
| Average | 0.90 | 0.90 | 0.90 | 0.89 | 0.89 | 0.89 | 0.87 | 0.87 | 0.87 |
| Average accuracy rate | | | 90.32% |  | | 89.49% |  | | 87.88% |
| Training Time | | | 17m |  | | 36m |  | | 4m |
|  | | | | | | | | | |
| Bow | SVM | | | Neural Network | | | Random Forest | | |
| Topics | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| World | 0.88 | 0.89 | 0.88 | 0.90 | 0.90 | 0.90 | 0.91 | 0.88 | 0.89 |
| Sports | 0.94 | 0.95 | 0.95 | 0.95 | 0.96 | 0.95 | 0.91 | 0.96 | 0.93 |
| Business | 0.85 | 0.84 | 0.85 | 0.86 | 0.86 | 0.86 | 0.84 | 0.85 | 0.84 |
| Sci/Tech | 0.86 | 0.85 | 0.85 | 0.87 | 0.86 | 0.87 | 0.86 | 0.84 | 0.85 |
| Average | 0.88 | 0.88 | 0.88 | 0.89 | 0.89 | 0.89 | 0.88 | 0.88 | 0.88 |
| Average accuracy rate | | | 88.29% |  | | 89.47% |  | | 88.11% |
| Training Time | | | 40m |  | | 24m |  | | 3m |

**Table 1.** Performance evaluation

For the first section (TF-IDF), SVM achieves the highest average precision, recall, and F1-score (0.90), followed closely by Neural Networks (0.89), while Random Forest lags behind (0.87). in addition, the Sports category consistently yields the highest classification scores across all models, with SVM achieving an F1-score of 0.96, the highest among all categories. On the other hand, the Business and Sci/Tech categories demonstrate slightly lower scores compared to World and Sports, indicating more challenges in distinguishing these topics effectively.

Furthermore, Random Forest is the fastest (4 minutes), making it computationally efficient. SVM takes 17 minutes, balancing between accuracy and computational cost. Neural Networks require 36 minutes, reflecting the higher computational expense associated with deep learning models.

According to the second section, the performance ranking remains relatively similar to TF-IDF, with Neural Networks (0.89) and SVM (0.88) performing slightly better than Random Forest (0.88) in terms of average precision, recall, and F1-score. Moreover, The Sports category continues to perform exceptionally across models, with Neural Networks achieving an F1-score of 0.95. The Business and Sci/Tech categories again report lower classification scores, reinforcing the notion that these topics may have overlapping features.

In terms of time, Random Forest remains the fastest (3 minutes), showing efficiency in handling BOW features. Neural Networks (24 minutes) perform slightly better than with TF-IDF but still demand high computational resources. SVM takes the longest time (40 minutes), likely due to the large-dimensional feature space created by the BOW model.

It can be observed from the results, TF-IDF appears to yield slightly better classification performance, particularly with SVM and Neural Networks. While, BOW is still competitive but tends to produce slightly lower accuracy rates, likely due to its inability to capture term importance as effectively as TF-IDF.

Hence, SVM performs best in terms of accuracy and robustness, making it a suitable choice when classification performance is the priority. Whereas, Neural Networks provide comparable accuracy but at a higher computational cost, making them less practical for real-time applications unless resources are abundant. Eventually, Random Forest is the most computationally efficient, providing decent classification performance with the lowest training time. The figures 1 and 2 demonstrate similar aspect as follows.

**Fig. 3.** TF-IDF comparison results

**Fig. 4.** BOW comparison results

In terms of the average accuracy rate and timing when we used TF-IDF, SVM algorithm was the best with 90.32% accuracy rate and 17m Training Time while Random Forest was the worst with 87.88% accuracy rate and 4m Training Time, and when we used BOW Neural Network algorithm was the best with 89.47% accuracy rate and 24m Training Time while SVM was the worst with 88.29% accuracy rate and 40m Training Time.

**4. Conclusions**

The study categorizes news articles into four distinct classes using three classification algorithms: SVM, Neural Networks, and Random Forest. The findings suggest that for real-world news classification applications, SVM with TF-IDF is the optimal choice when accuracy is the primary concern. Conversely, Random Forest with BOW is more suitable for scenarios requiring faster processing with moderate accuracy. While Neural Networks demonstrate strong classification performance, they demand substantial computational resources and extended training time, making them more appropriate for large-scale, high-resource environments. Furthermore, when accuracy is the top priority, SVM with TF-IDF remains the preferred approach. If computational efficiency is a critical factor, Random Forest with BOW serves as the best alternative. However, if deep learning solutions are viable, Neural Networks with TF-IDF offer robust classification capabilities. For future research, combining TF-IDF and BOW could be explored to leverage the advantages of both feature extraction methods. Additionally, a hybrid classification approach could be implemented to further enhance performance. Finally, incorporating an additional dataset would enable a more precise evaluation of model accuracy.

**References**

1. G. Kaur and K. Bajaj, "News classification and its techniques: a review," IOSR Journal of Computer Engineering (IOSR-JCE), vol. 18, pp. 22-26, 2016.
2. J. Agarwal, S. Christa, A. Pai H, M. A. Kumar and G. P. M. S, "Machine Learning Application for News Text Classification," 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2023, pp. 463-466, doi: 10.1109/Confluence56041.2023.10048856.
3. Y. V. Singh, P. Naithani, P. Ansari and P. Agnihotri, "News Classification System using Machine Learning Approach," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 186-188, doi: 10.1109/ICAC3N53548.2021.9725409.
4. T. B. Shahi and A. K. Pant, "Nepali news classification using Naïve Bayes, support vector machines and neural networks," in 2018 International Conference on Communication Information and Computing Technology (ICCICT), 2018, pp. 1-5.
5. M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," WSEAS transactions on computers, vol. 4, pp. 966-974, 2005.
6. V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," International Journal of Artificial Intelligence & Applications, vol. 3, p. 85, 2012.
7. A. N. Chy, M. H. Seddiqui, and S. Das, "Bangla news classification using naive Bayes classifier," in 16th Int'l Conf. Computer and Information Technology, 2014, pp. 366-371.
8. M. A. Shehab, O. Badarneh, M. Al-Ayyoub, and Y. Jararweh, "A supervised approach for multi-label classification of Arabic news articles," in 2016 7th International Conference on Computer Science and Information Technology (CSIT), 2016, pp. 1-6.
9. D. Y. Liliana, A. Hardianto, and M. Ridok, "Indonesian news classification using support vector machine," World Academy of Science, Engineering and Technology, vol. 57, pp. 767-770, 2011.
10. S. Kamrus, M. K. Alam, Md. A. Nabi, A. Fahim , and B. A. Faisal, "A comparative study of different text classification approaches for bangla news classification." In 2021 24th International Conference on Computer and Information Technology (ICCIT), pp. 1-6, 2021.
11. L. Chun-Ming, C. Mei-Hua, K. Endah, K. V. Vinod, and Y. Chao-Tung, "Fake news classification based on content level features," Applied Sciences 12, no. 3, pp. (2 - 5), 2022.
12. A. E. Aman, AG News Classification Dataset, 2020, Available: https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset. Accessed: 2024-10-24.
13. Y. Savaş, J. Dhanya, K. Can, and B. Ayşe, "Classification of "hot news" for financial forecast using NLP techniques," In 2018 IEEE International Conference on Big Data (Big Data), pp. 4719-4722, 2018.
14. N. Disayiram and R. A. H. M. Rupasingha. "A comparative study of classifying English news articles using machine learning algorithms," Trends in Electrical, Electronics, Computer Engineering Conference (TEECCON), pp. 50-55.