## **Online News Article Classification Using Machine Learning Approaches**

Abinaya N, Jayadharshini P, Priyanka S, Keerthika S, Santhiya S

Assistant Professor, Department of Artificial Intelligence

Kongu Engineeering College, Perundurai

**Abstract:**

Multiple forms of information are fed into the internet news portals from numerous sources. In the majority of real-life settings, it is highly desired to categorize this information into the proper groups, and it is crucial to have a reliable mechanism in place for dividing news into these groups. The classification system is strengthened and improved using machine learning because there has been little research done on how to classify news headlines, there is now a chance to examine this subject in more detail. This work focuses on categorizing real-time news based on its headlines. A system has been designed to classify each news headline to its pre-defined category. The model is trained such that the machine can predict the category of the news item accurately. The news headline will be fetched in real time and will be passed through this classifier. This whole process will not only lead to a better working model but also show a comparative study of naïve bayes and logistic regression models for classifying news headline. Logistic Regression produces the best outcomes compared to Naïve Bayes.

**Keywords:**

Text classification, Naïve bayes, Logistic regression, Tokenization, Natural Language Processing

1. **Introduction**

News article classification in machine learning is the process of using algorithms to automatically categorize news articles by their content. This process can be useful for a variety of applications, such as improving search engine results and identifying the main topics of a news article. Machine learning algorithms frequently combine supervised learning techniques with natural language processing (NLP) strategies to categorize news articles. NLP techniques are used to extract relevant features from the text of the article, such as the words used and their context. These features are fed into a model, which is trained on a news dataset. The features needed to forecast the category of a particular news article are extracted by the supervised learning model [1]. Based on the patterns and associations discovered from the training data, this prediction was made. As the model is exposed to more data over time, it may then be adjusted and enhanced. Overall, the use of machine learning in newspaper classification can help to improve the efficiency and accuracy of categorizing large volumes of articles. This can be useful for a wide range of applications, from improving search engine results to providing a better understanding of the topics and trends in the news.

1. **Literature survey**

In 2019, Azerbaijani news articles are processed with supervised models. Naïve Bayes and SVM are compared and found that Naïve Bayes yields better accuracy using count vectorization than SVM with count vectorization [1].

In 2020, the traditional method of TFIDF feature learning is compared with the proposed model of pre-trained embedding along with CNN. CNN provides better F1 score than traditional model [4]. Sree Devi et al, compared various models for newspaper article and analyzed SVM produced better results with moderate training time [3]. A comparison of news categorization was done in 2020. Dataset from BBC was used for this comparison and Naïve Bayes performed better than other algorithms [2].

In 2021, Bengali sports news article is considered for categorization. They used this dataset with six popular machine learning algorithms like Decision Tree (DT), Support Vector Classifier (SVC), Logistic Regression (LR), Random Forest (RF) etc. This dataset composed of 4 classes. SVC provides the F1 score of 97.60% with unigram, bigram and trigram features [5].

1. **Methodology**

**Dataset Description**

There are 45.5k rows and 5 columns in the dataset. Our model's goal column is called Category. Business, politics, food and drink, travel, parenting, style and beauty, wellness, world news, sports, and entertainment are among the categories into which the news stories are divided. Features like heading, description, and keywords make up this dataset [6].

**Handling missing data**

The missing data can be handled in two ways. Imputation and removal are the two most common methods to handle the missing data. Removal method can be used when the percentage of missing data is less than 30%. It does not produce required result**.** The imputation method can be performed o the basis of mean or median or mode based on its data type and dataset taken.

**Tokenization**

Tokenization is an excellent method for many types of data. Maintaining the data format is especially important for preserving referential integrity and database schemas. The raw text is divided into tokens, which are words and sentences. These tokens aid in context comprehension or model development for NLP. By examining the word order in the text, tokenization assists in analyzing the text's meaning.

**Count Vectorization**

By converting the provided text article into a sparse matrix, countvectorizer makes it possible to employ text data in machine learning models.

**Bag-of-words usage in the study**

The bag of words approach is a way of representing text data when working with natural language processing tasks. It is a simple representation of text data where each unique word in the text is represented by a single token or feature, and the presence (or absence) of each word is recorded in a vector. This approach allows the use of machine learning algorithms to analyze and classify the text data based on the presence or absence of certain words or combination of words [3].

**MACHINE LEARNING ALGORITHMS**

Figure 1 shows the methodology of the proposed work. The pre-processed training dataset is used for the learning the models.

**Naïve Bayes**

The first step is to train it on a labeled dataset of newspaper articles. The Bayes theorem is used to determine the likelihood of each category given the words in the article. Using the article's words as input and the trained model to determine the probability of each category, the algorithm may then be used to predict the category of new articles. The category with the highest probability is then selected as the predicted label for the article.

Overall, Naïve Bayes is a simple and effective algorithm for newspaper article classification, especially when the number of categories and the number of words in the vocabulary are large. It is also relatively fast to train and it is effective than other algorithms [1].

**Logistic Regression**

Another depiction of supervised learning is logistic regression. It is used to determine or forecast the likelihood that a binary (sure/no) event will occur. It’s miles a data analysis approach that uses arithmetic to find the relationships among information elements. Then, it uses this dating to predict the fee of one of those elements based totally on the other. The prediction usually has a finite wide variety of outcomes, like yes or no [5].

1. **Results and Discussion**

Once a machine learning model has been trained for newspaper article classification, it is important to evaluate its performance to determine how well it is able to predict the categories of new articles. This is typically done using a separate dataset of test examples, which the model has not seen during training. The model’s predictions on the test examples are then compared to the true label to access its accuracy. There are several specific metrics that may be used to evaluate the performance of a model on a classification task, such as newspaper article classification. Accuracy, precision, recall, F1 measure these are the common metrics. Table 1 provides the accuracy of two models implemented in this work.

***Accuracy =***

***Precision =***

***Recall =***

***F1 Score =***

***Table 1: Accuracy of various models***

|  |  |  |
| --- | --- | --- |
| **S.No** | **Evaluation Model** | **Accuracy** |
| 1 | Naïve Bayes | *76* |
| 2 | Logistic Regression | *79* |

Figure 2 provides the precision, recall and F1 score of individual news categories through logistic regression model. It is observed that Sports category has the highest metrics compared to all other news categories. Figure 3 provides the metrics for Naïve Bayes model with individual categories. This model also gives the highest accuracy for sports category.

**Fig 2:Precision, Recall and F1 Score of Logistic Regression**

**Fig 3:Precision, Recall and F1 Score of Naïve Bayes**

**4.Conclusion:**

Newspaper article classification using machine learning is a powerful and effective way to automatically organize and categorize large collections of text data. There are a variety of techniques and algorithms that can be used for this task, such as the bag-of-words model, n-grams, Naïve bayes, support vector machines, and decision tree. Each of these approaches has its own strengths and weaknesses, and choosing the right approach will depend on the specific goals and characteristics of the classification task. From the model, we conclude that Logistic regression method gave us a good result of 79% accuracy.

**Acknowledgement:**

We would like to express our gratitude to Ms. Bhuvaneshwari.C, Ms. Harini.M, Snehavarshini.S, Soundharya.M, Department of Artificial Intelligence and Data Science, Kongu Engineering College for their role in developing the proposed model.

**References:**

1. Suleymanov, Umid, and Samir Rustamov. "Automated news categorization using machine learning methods." IOP conference series: materials science and engineering. Vol. 459. No. 1. IOP Publishing, 2018.
2. Deb, Nabamita, et al. "A comparative analysis of news categorization using machine learning approaches." Int. J. Sci. Technol. Res 9.1 (2020): 2469-2472.
3. Sreedevi, J., M. Rama Bai, and C. Reddy. "Newspaper article classification using machine learning techniques." Int. J. Innov. Technol. Explor. Eng (2020): 2278-3075.
4. S. Ramraj, R. Arthi, S. Murugan and M. S. Julie, "Topic categorization of Tamil News Articles using PreTrained Word2Vec Embeddings with Convolutional Neural Network," 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE), Keonjhar, India, 2020, pp. 1-4, doi: 10.1109/CISPSSE49931.2020.9212248.
5. Barua, Adrita, Omar Sharif, and Mohammed Moshiul Hoque. "Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation." Procedia Computer Science 193 (2021): 112-121
6. Abinaya N, Anand R and Arunkumar T, ”An Exhaustive Survey on Automatic Text Summarization Using Machine Learning Approches”, Webology, (2021), pp.1184-1190.
7. https://www.kaggle.com/rmisra/news-category-dataset