Online News Article Classification Using Machine Learning Approaches

[1]Bellam Sruthi, [2]M Prasanna Kumar,
[1]Student, Department of CSE, Krishnaveni Engineering college for women, kesanupalli(Narasaraopeta).
[2]Professor , Department of CSE, Krishnaveni Engineering college for women, kesanupalli(Narasaraopeta).

Abstract –

Internet news portals receive a variety of information from a wide range of sources. It is very desirable to organize this information into the appropriate categories in most real-world situations, and having a trustworthy system in place for doing so is essential. Since there hasn't been much research on categorizing news headlines, there is now an opportunity to look more closely at this topic. As a result, the classification system is strengthened and enhanced using machine learning. The main goal of this project is to classify current events according to their headlines. To assign each news headline to its predetermined category, a mechanism has been developed. The machine can correctly estimate the news item's category thanks to the model's training. This classifier will process the news headline after it is retrieved in real time. In addition to producing a better working model, this entire procedure will compare the effectiveness of logistic regression and naïve bayes models for news headline classification. When compared to Naïve Bayes, Logistic Regression yields the best results.

Keywords: Natural language processing, text categorization, Naïve Bayes, logistic regression, and tokenization.

I. INTRODUCTION

In machine learning, news article classification is the process of automatically classifying news articles according to their content using algorithms. Numerous applications, including enhancing search engine rankings and pinpointing the primary subjects of news articles, can benefit from this procedure.

In order to classify news items, machine learning algorithms usually integrate supervised learning approaches with natural language processing (NLP) techniques. NLP approaches are used to extract relevant features, such as the words used and their context, from articles' textual content. A model is trained using a news dataset once these features are fed into it. The supervised learning model extracts the information required to predict the category of a certain news article [1]. This prediction was produced in light of the correlations and patterns found in the training data. The model may then be modified and improved as it is subjected to additional data over time. In conclusion, the effectiveness and accuracy of classifying large collections of articles can be improved by using machine learning into newspaper classification. This feature has potential for a number of uses, including as improving search engine results and encouraging a more thorough understanding of news subjects and patterns.

II. LITERATURE SURVEY

In 2019, supervised models are used to process news stories from Azerbaijan. When comparing Naïve Bayes and SVM, it is discovered that Naïve Bayes uses count vectorization to produce more accuracy than SVM [2].

In 2020, the suggested model of pre-trained embedding in conjunction with CNN is contrasted to the conventional approach of TFIDF feature learning. CNN outperforms the conventional model in terms of F1 score [3]. When comparing several models for newspaper articles, Sree Devi et al. found that SVM yielded superior results with a moderate training time [4]. In 2020, a comparison of news classifications was conducted. For this comparison, the BBC dataset was used, and Naïve Bayes outperformed the other algorithms [5]. Bengali sports news articles are taken into consideration for classification in 2021. Six well-known machine learning techniques, including Random Forest (RF), Logistic Regression (LR), Support Vector Classifier (SVC), and Decision Tree (DT), were applied to this dataset. There are four classes in this dataset. SVC uses bigram, trigram, and unigram features to provide an F1 score of 97.60% [6].

III. METHODOLOGY

A. Dataset Description

The dataset consists of 5 columns and 45.5k rows. Category is the name of the target column in our model. The news stories are broken down into many categories, including business, politics, food and drink, travel, parenthood, wellness, style and beauty, foreign news, sports, and entertainment. This dataset consists of features such as heading, description, and keywords [7].

B. Handling missing data

There are two methods to deal with the missing data. The two most popular techniques for dealing with missing data are imputation and elimination. When the percentage of missing data is less than 30%, the removal procedure may be applied. It doesn't yield the desired outcome. The imputation approach can be used to the mean, median, or mode, depending on the dataset and data type.

C. Tokenization

For a variety of data types, tokenization is a great technique. Database schemas and referential integrity are particularly dependent on maintaining the data format.The process of breaking up raw data into distinct units, like words or phrases, is known as tokenization. These tokens are crucial for understanding context and creating NLP models. Tokenization facilitates successful text analysis and meaning extraction by breaking the text up into tokens while maintaining word order.

D. Count Vectorization

Countvectorizer enables the use of text data in machine learning models by transforming the supplied text article into a sparse matrix.

E. Bag-of-words usage in the study

When working on projects involving natural language processing, the bag of words technique is one method of encoding text data. Every unique word in the text is represented by a single token or feature in this simple form of text data, and each word's occurrence—or lack thereof—is recorded in a vector. By taking into account the presence or lack of particular words or word combinations, this approach facilitates the analysis and classification of text input [4].

MACHINE LEARNING ALGORITHMS

A.Naïve Bayes

Training it on a labeled dataset of newspaper articles is the initial step. Given the terms in the article, the likelihood of each category is ascertained using the Bayes theorem. The technique may then be used to predict the category of new articles by using the words from the article as input and the trained model to calculate the probability of each category. The label with the highest likelihood is used to establish the article's anticipated category. All things considered, Naive Bayes is a simple and effective method for categorizing newspaper articles, particularly when there are a lot of classes and vocabulary. In addition to being more efficient than previous algorithms, it can be trained rather quickly [2]. Equation 1 gives the posterior probability, or the chance that an article belongs to a specific class based on its characteristics.

$$P(C|X) = \frac{[P(X|C)*P(C)]}{P(X)} \qquad (1)$$

Calculate P(C|X) for each class in order to classify a new article. Then, use Equation 2 to determine which class has the highest likelihood of being the anticipated class.

$$\text{Predicted Class} = \text{argmax}(P(C \mid X)) \qquad (2)$$

B. Logistic Regression

Logistic regression, a technique used to forecast or estimate the likelihood of a binary event occurring, such as a yes/no result, is another way to demonstrate supervised learning [6]. This approach is a type of supervised learning that works well for tasks involving binary classification and can be modified for multi-class classification. Any numerical input can be transformed into a bounded range between 0 and 1 using the sigmoid function. This method is used to calculate the likelihood that an item belongs to a particular class. Equation 3 provides the logistic function formula:

$$P(Y = 1|X) = 1 / (1 + e^{(-z)}) \qquad (3)$$

where the likelihood that the article pertains to class 1 is represented by P(Y=1|X).
X represents the features of the article.
Z represents linear combination of the features weighted by the coefficients learned during training.

The dot product of the feature vector X and the coefficient vector W, where W stands for the weights assigned to each feature, yields the linear combination z. Equation 4 provides the linear combination z.

$$z = W1 * X1 + W2 * X2 + \ldots + Wn * Xn + b \quad (4)$$

where

W1, W2, ..., Wn are the weights for each feature.

X1, X2, ..., Xn are the feature values.

b is the bias term.

IV. RESULTS AND DISCUSSION

It's critical to evaluate a machine learning model's performance once it has been trained for newspaper article categorization in order to determine how well it predicts the categories of new articles. Usually, to accomplish this, a different dataset of test cases that the model hasn't seen during training is used. The accuracy of the model is then assessed by comparing its predictions on these test examples with the actual labels. A model's success in a classification assignment, like classifying newspaper articles, can be evaluated using a variety of unique indicators. Common measures include accuracy, precision, recall, and F1 measure. The accuracy metrics for the two models used in this study are shown in Table 1.

The metrics for every news category as established by the logistic regression model are shown in Figure 1. When compared to all other news categories, the sports category is found to have the highest metrics. The metrics for the Naïve Bayes model with distinct categories are shown in Figure 2. Additionally, this model provides the best accuracy for the sports category.

TABLE I. ACCURACY OF VARIOUS MODELS

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

V. CONCLUSION

A reliable and effective method for automatically organizing and classifying large text data collections is to use machine learning for newspaper article categorization. Numerous methods and algorithms, including the bag-of-words model, n-grams, Naïve bayes, support vector machines, and decision trees, can be applied to this problem. Each strategy has advantages and disadvantages, and the particular objectives and features of the categorization task will determine which strategy is best. We deduce from the model that the 79% accuracy rate obtained using the logistic regression approach was good.

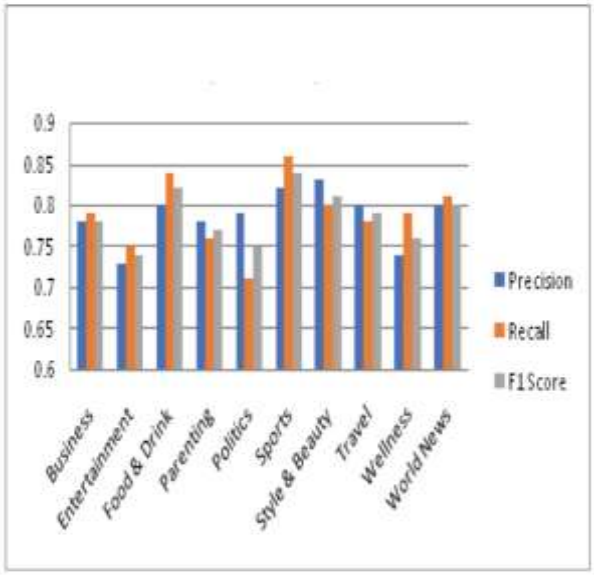

Fig.1. Performance metrics of Logistic Regression

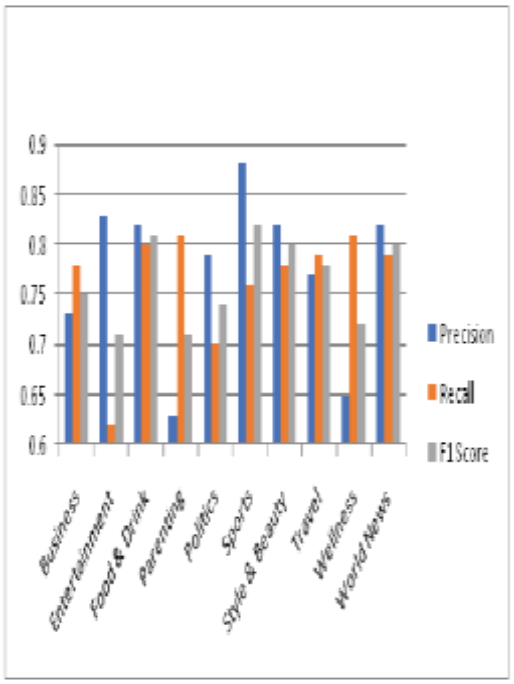

Fig.2. Performance metrics of Naive Bayes

REFERENCES

[1]. G. K. Kamalam, K. Lalitha, E. Priyadarshini, V. C. Janani and P. M.Sasidhar, "Forecasting the spread of COVID-19 using supervisedmachine learning models." AIP Conference Proceedings. Vol. 2387.No. 1. AIP Publishing, 2021.

[2]. 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.

[3]. 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

[4]. Sreedevi, J., M. Rama Bai, and C. Reddy. "Newspaper article classification using machine learning techniques." Int. J. Innov. Technol. Explor. Eng pp.2278-3075, 2020.

[5]. Deb, Nabamita, et al. "A comparative analysis of news categorization using machine learning approaches." Int. J. Sci. Technol. Res 9.1, pp.2469-2472, 2020.

[6]. Barua, Adrita, Omar Sharif, and Mohammed Moshiul Hoque. "Multi-class Sports News Categorization using Machine Learning Techniques:Resource Creation and Evaluation." Procedia Computer Science vol.193, pp.112-121, 2021. https://www.kaggle.com/rmisra/news-category-dataset