Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 3 (2024)

Machine Learning Methods for Forecasting News Trends

Dr. Rajesh Saturi[1], Panuganti Hanumantha Rao[2], Bandamedi Naveen[3], Bairagoni Sanjana[4], Ayiti Dhanunjay[5]

[1]Associate Professor,[2]Assistant Professor,Department of computer science & engineering
[3,4,5]B-Tech student, Department of computer science & engineering
Vignana bharathi institute of technology,
Aushapur, ghatkesar, hyderabad, telangana, india

Abstract:- News Classification project includes activities like data gathering and preprocessing, selection of appropriate algorithms like Random Forest, Logistic Regression, K Nearest Neighbor, Gaussian Naïve Bayes, Multinomial Naïve Bayes, Decision Tree Classifier training and evaluation of the model's performance. We currently have a news-related dataset with a variety of data kinds, including political, sports, entertainment, and educational information. An expanding amount of information has to be categorized and arranged using automated techniques as the volume of electronic data grows dramatically. This research paper explores the application of machine learning approaches for news classification, aiming to categorize news articles into predefined topics or themes.The paper delves into the challenges posed by linguistic nuances, evolving language, and the need for adaptability in news classification systems. Evaluation metrics, including recall, F1-score, accuracy, and precision are investigated to evaluate the effectiveness of these systems. Real-world applications in information retrieval, recommendation systems, and personalized content delivery are highlighted, emphasizing the practical significance of news classification. The study concludes with a discussion on ongoing developments, emphasizing the impact of deep learning and pre-trained language models on the accuracy and efficiency of news classification systems.

All things considered, this study offers profound knowledge of modern techniques and challenges in the field of machine learning for news classification, offering a foundation for further advancements and applications in the evolving landscape of digital information. To achieve the best results, we have applied a classification algorithm along with various word vectorizing approaches to this data. We compared our findings based on several measures, such as precision, F1 Score, accuracy, and recall with the goal to boost performance. We employ all the methods of classification and select the one that has the best precision, recall, accuracy, and F1 Score.

Keywords*:* Categorization, Machine Learning, News article, Random Forest, K Nearest Neighbor, Support vector classifier, Gaussian Naïve Bayes, Multinomial Naïve Bayes, Logistic Regression and Decision Tree Classifier..

1. Introduction

Through news stories on the internet, we are constantly confronted with an enormous amount of information [1]. Data indexing and retrieval involving text requires the use of text mining technologies [1]. Text mining is an extremely powerful technique used for text categorization that is used to extract hidden information from massive datasets [1]. This text is in the form of unstructured text. A variety of classifiers, including Random Forest, K Nearest Neighbor, Support Vector Classifier, Gaussian Naïve Bayes, Multinomial Naïve Bayes, Logistic Regression and Decision Tree Classifier are available for classifying news [1].

People nowadays inhabit a civilization in which the internet connects all individuals [2]. Humans may easily receive, process, and exchange information through news portals that are widely distributed online [2]. The textual data is created from several sources and may be found in publications, journals, web pages, emails, conference materials, editorials, digital/electronic papers, and publications. Instead of being restricted to printed materials like newspapers, magazines, and books, a lot of individuals use these internet sources for daily

information access. Although it is now easier to get this kind of information, managing it is still tough due to the complexity of knowledge organization. It is thought that organizing this type of digital data is essential to properly categorizing it [3]. Here, we categorize the news stories according to their genre or category. Consider a person who likes to read political news, for instance. Given the abundance of news stories on the internet, it would be difficult for a certain individual to read solely political pieces [3]. Text classification involves the use of a number of different methodologies [4]. We compare the outcomes for different classifiers in this study. Issues with news classification [4]. This study's main objective is to use machine learning models to solve the challenge of news text classification [5].

As a result, there is a lot of study focused on the algorithms and methods that successfully categorize news [6]. A range of digital media formats, including like websites, social networking feeds, news on the internet platforms, and others, replaced publications and magazines. Customers now have a simpler time getting their hands on the newest news [7]. Fortunately, a variety of computational methods are available to classify specific articles according to their textual content [7]. Sorting texts into specific groups or tags is the main goal of text categorization. Furthermore, the TC heavily relies on news classification. It describes the process of creating a model that determines a text's topic [8].

Accurately predicting the target class for every case in the data is the aim of grouping [9]. Text categorization techniques have drawn more attention than they have in the past due to the Internet's text information resources expanding so quickly [10].

2. Literature Survey

In the journal IJISRT, an author MAHAJAN S.Dsuggested a News Classifications [1] utilising Machine Learning with certain word vectorization approaches in the year 2021, demonstrating that Naive Bayes can successfully categorise news with 80-85% accuracy. This article uses word vectorization in conjunction with the Naive Bayes approach.

The authors named Muhammad Hatta, Rahmatul kholiq published a journal in IJEST [2]. In the year 2022, a light gradient-boosted machine method to identifying news and demonstrated that their approach got great results. The framework is analysed using different extracting features procedures.

In journal Scopus, the authors Jeelani Ahmed, Muqeem Ahmed [3] suggested a topic on categorization of news using the methods of machine learningin the year 2021, demonstrating that the higher performance utilising multiple news data sets provides sufficient classification accuracy by 93%.

In the journal HINDAVI, the authors suggested a topic on categorization of news in the year 2022. This study found that their approach [4] performed well in the classification task on the data set with a success rate of 97.86%.

In the year 2022, the authors suggested a topic on categorization of news in the journal IEEE [5]. They transformed TTE and LFC parts into textual vectors of features using BERT and CNN as their frameworks with accuracy of 99.0% and 96.2%, respectively. According to the experimental findings, their approach performs better in terms of accuracy and categorization impact than the other approaches.

Within the journal IEEE [6], an author named. Roher has been published a paper on categorization of fake news. Several Machine Learning approaches have been trained to identify bogus news items on the self-aggregated dataset. After training the model on several items of news, their best approach demonstrated the best result in identifying false news, while NB demonstrated the highest recall.

In the journal HINDAVI, the four authors [7] suggested a approach on detecting false news in the year 2020.They employed feature set and ensemble approaches in this study. With their model they achieved a maximal result of 99% [7]. The problem of spotting bogus news reports using ensemble approaches and machine learning models has been covered in this research.

Three authors named Shuo Lv, Peimin Cong, and Xinying Chen published a journal in IEEE in the year 2022[8]. They suggest an approach to address the challenge of identifying lengthy text in Chinese news documents. The new technique considerably increases the results of text categorization, according to experimental data.

In the year 2020, the authors namedBlessy, Lubna Juveria, M. Sundarababu, Ch. Chandra Mohan, Mahendra Suthar, and CH. Devi Harsha have proposed a Research Journal in JETIR. The Multinomial Naive Bayesian classifier will be implemented in the suggested framework in order to increase the accuracy of online news categorization predictions. When compared to other examined classifiers, multinomial naive bayes is shown to provide a greater exactness [9].

In the journal HINDAVI, the authors Ningfeng Sun and Chengye Du suggested an article to build an approach on deep learning. In order to classify text in network news, this study primarily investigates two topics: text feature selection and representation [10]. The findings of this experiment demonstrate that the categorization approach outperforms the old method in terms of comprehensive performance.

3. Methodology

Our methodology gives the output based on the description. Fig 1 shows the whole process of the methodology.

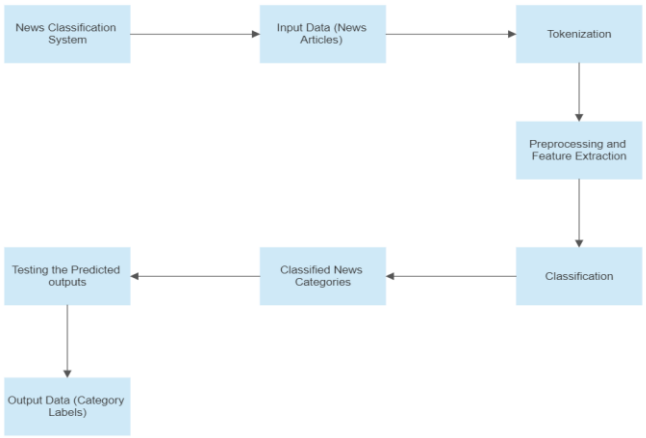

Fig:1 General view of Proposed System

A. Modules :

1.News Classification System: It is a framework where the user provides an input from the news dataset and the system displays the category for the user requirement.

2. Input Data: In this module we are giving the BBC news dataset for the different models.

3.Preprocessing and Feature Extraction: Preprocessing and feature extraction are crucial steps in preparing text data for news classification. Convert the tokenized text into numerical features using the Bag-of-Words model. Using libraries like Count Vectorizer for this purpose.

4.Tokenization: Tokenization is a crucial step in news classification, as it entails segmenting the text into discrete components, often words or sub words. Word tokenization break downs each sentence into individual words using a word tokenizer. For example, in Python with NLTK.

5.Classification: Divide the dataset into sets for testing and training. Build down the Random Forest Model and Train It the RF Classifier must be created and trained using the training set of data. Evaluate Model Performance by Assess the model's performance using accuracy and other relevant metrics.

6.Classified News Categories: After the completion of classification of news data by applying various models it will be going to classify the different categories.

7.Testing the Predicted outputs: By using the trained model to make predictions on the test set.

8. Output Data: Display the output data to the user.

B. Dataset description

In order to do machine learning-based news categorization, I have gathered a dataset from the BBC that comprises news articles along with their article id, text, and categories.

The BBC provided us with 1490 data, which is divided into 5 categories overall. We are looking to identify the following five categories: sports, business, politics, technology, and entertainment.

C. Data preprocessing

Newspapers, periodicals, and other publications are among the various sources from which we gather news. The dataset may be found in a variety of forms, including.csv, json,.pdf,.doc, and html upon the completion of news collecting. We get the dataset from several sources; thus, it must be necessary for cleaning in order to remove irrelevant and noisy data [1]. The procedure of turning unprocessed data into an understandable structure is known as data preparation. Preprocessing your text essentially means putting it in a format that can be easily analysed and predicted for the task at hand.

1.Label Encoding: In order for machine learning models—which can only accept numerical data—to fit categorical columns, a method known as label encoding is employed to transform them into numerical ones. Label encoding transforms categorical data into numerical data while allocating a distinct number (beginning at 0) to every data type [3].

	Category	CategoryId
0	business	0
3	tech	1
5	politics	2
6	sport	3
7	entertainment	4

Table 1: Label encoding for different news category.

2.Lowercasing: One of the easiest and most efficient ways to preprocess text is to lowercase ALL of your text data, even though this is frequently forgotten. It helps a great deal with predicted output consistency and is relevant to most text mining and natural language processing challenges, even if your dataset is not particularly large.

3. Stop words: A group of frequently used terms in a language are called stop words. The English words "a," "the," "is," "are," and so on are examples of stop words. The idea behind stop words is that they help us concentrate on the crucial words in a text by eliminating words with little information.

4. Lemmatizing the Words: The procedure of classifying a word into a single group so they may be examined as a unit. Lemmatization and stemming are comparable, but lemmatization gives words context. Thus, it unites terms with related meanings into a single term. Lemmatization is favoured over stemming as it analyses the morphology of the words.

5. Exploratory Data Analysis (EDA): An strategy of analysing datasets to bring out their salient characteristics is known as Exploratory Data Analysis. We utilize this to look at what the data can tell us before we start analysis.

Determining significant features of the data from a spreadsheet or a column of numbers is a difficult task. Analysing simple statistics to get insights can be tiresome, monotonous, and/or intimidating.

To help in this circumstance, approaches for exploratory data analysis have been developed.

D. Train-test split

We need to separate a dataset into sets to train and test in order to evaluate how well our approach performs. The model is fitted using the known statistics of the train set. Second collection, referred to as the test data set, is only utilized for forecasting purposes.

The initial set of data was split into train (80%) and test (20%) sets after being separated into features (X) and target (y). As a result, one set of data would be used to train the algorithms, while another set of data would be used to test them.

E. Data Visualization:

To categorize the data, we are using a bar chart method of data visualization. We display the distribution categorization in a two-dimensional bar chart that has both counts and category of data, along with the number of counts in each category.

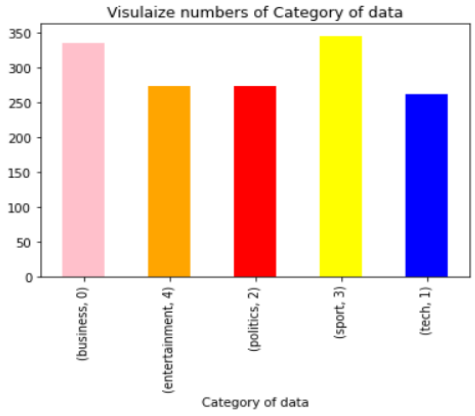

Fig:2Bar Chart for Category Distribution

F. Comparative analysis of algorithms

We employed seven algorithms using ensemble models. Test accuracy, precision, recall, and F1 scores were all compared. Out of all the machine learning models, we discovered that Random Forest Classification provides the highest accuracy for this dataset.

	Model	Test Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
0	LR	97	97	97	97
1	RF	98	98	98	98
2	MNB	97	97	97	97
3	SVM	96	97	97	97
4	DT	83	83	83	83
5	KNN	74	74	74	74
6	GNB	77	76	76	76

Table 2: Performance metrics of various models.

4. System Design

Random forest is useful for both categorizing and predicting the label of text data especially while working with text data. The random forest system uses techniques like TF-IDF vectorization and count vectorizer.

These techniques are also useful for increase in both accuracy and precision.

Count Vectorization: Count vectorization is just counting the number of words in the feature set. It is necessary to extract the feature set for machine learning from text documents. Every unique word is treated as a single feature in the feature set, which has many dimensions due to the large number of unique numbers in the whole dataset. Each collection of features is shown as a document. The term "count vectorization method" refers to the number of words assigned in a text that correspond to that characteristic [1].

TF-IDF Vectorization: Based on rescaling the frequency, these terms are prevalent. The number of times terms like "that," "is," and "then" are used in the current report. Team frequency is the quantity of times a document appears. Downscaling terms that exist throughout the document is known as inverse document frequency. This frequent term is permitted to lighten the load. For instance, a phrase like "of," "is," or "that" may be used frequently, but its significance is modest, thus we must give the more common terms less weight [1].

We have applied classification technique that is random forest. The random forest classifier checks accuracy with both approaches of TF-IDF and count vectorization methods.

5. Implementation And Results

For implementation, the random forest approach is taken into consideration. The demonstration is done using python programming on Visual Studio code.

The Random Forest approach has given the greatest results among all the models, with greater Precision, Recall, and Test Accuracy, according to the findings we have. It suggests that when compared to the other models, the Random forest model is significantly more accurate and reliable.

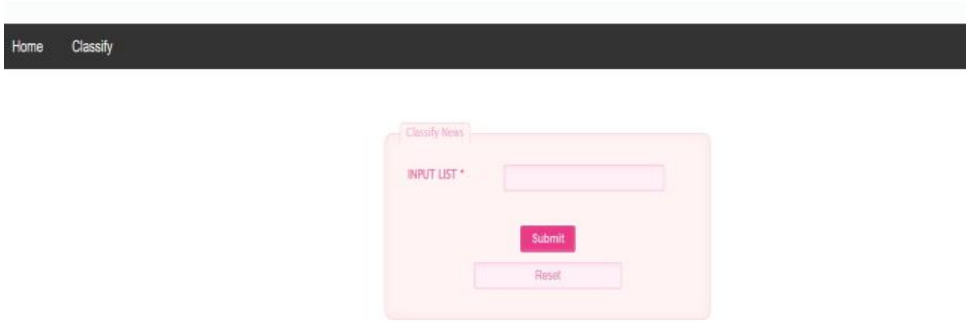

Fig: 3 Home page of Implementation

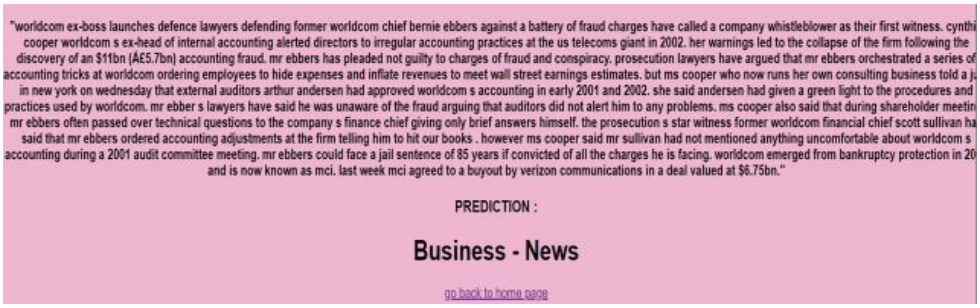

Fig: 4Final Result of Classification

6. Performance Metrics

The performance measures we employ include support, recall, precision, accuracy and F1-score. The ratio of accurate outcomes to total results is known as precision. The ratio of accurate outcomes to correct results that ought to have been returned is known as recall.

The F1 score depends on both recall and accuracy. Support is the number of times an item appears in a dataset [1].

	Precision (%)	Recall (%)	F1-score	Support (%)
0	96	99	97	75
1	100	91	95	58
2	96	98	97	56
3	98	100	99	63
4	98	100	99	46
accuracy			98	298
macro avg	98	98	98	298
Weighted avg	98	98	98	298

Table 3: Performance Metrics of Count-Vectorizer

	Precision (%)	Recall (%)	F1-score	Support (%)
0	95	96	95	75
1	100	93	96	58
2	93	96	95	56
3	100	100	100	63
4	98	100	99	46
accuracy			97	298
macro avg	97	97	97	298
Weighted avg	97	97	97	298

Table4: Performance Metrics of TF-IDF Vectorizer

7. Conclusion And Future Scope

TheRandomForest classifier has been effectively employed in order to create a new categorization model.When compared to the other classifiers, Random Forest performed the best on the dataset we used.

Lastly, we may conclude that Random Forest will provide the most accurate results based on the Test Accuracy scores.

The future scope of a news classifier developed using machine learning is promising.

1. Better Categorization Accuracy:We'll work to improve the model's accuracy moving ahead.

2.In real time and Dynamic Classification: Real-time classification on huge websites concurrently with the inclusion of news articles.

3.BiasedNews Identification and Prevention: In the future, efforts might utilize use of methods to identify and reduce biased news content. This helps to the fair and balanced classification of the news.

4. Integrated Methods: Combining various AI approaches could result in improved news classification systems capable of managing any kind of media.

References

[1]. MAHAJAN, S. “News Classification Using Machine Learning”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 9, no. 5, May 2021, pp. 23-27, DOI:10.17762/ijritcc.v9i5.5464.

[2]. Muhammad Hatta Rahmatul Kholiq, Wiranto Wiranto, Sari Widya Sihwi."News classification using light gradient boosted machine algorithm" DOI:http://doi.org/10.11591/ijeecs.v27.i1.pp206-213

[3]. Ahmed, J., & Ahmed, M. (2021). ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES. IIUM Engineering Journal, 22(2), 210–225. DOI:https://doi.org/10.31436/iiumej.v22i2.1662

[4]. Saigal, P., Khanna, V. Multi-category news classification using Support Vector Machine based classifiers. SN Appl. Sci. 2, 458 (2020). DOI: https://doi.org/10.1007/s42452-020-2266-6

[5]. Lujuan Deng, Qingxia Ge, Jiaxue Zhang, Zuhe Li, Zeqi Yu, Tiantian Yin, Hanxue Zhu, "[Retracted] News Text Classification Method Based on the GRU_CNN Model", International Transactions on Electrical Energy Systems, vol. 2022, Article ID 1197534, 11 pages, 2022. https://doi.org/10.1155/2022/1197534

[6]. D. Rohera et al., "A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects," in IEEE Access, vol. 10, pp. 30367-30394, 2022, doi: 10.1109/ACCESS.2022.3159651.

[7]. Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, Muhammad Ovais Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods", Complexity, vol. 2020, Article ID 8885861, 11 pages, 2020. https://doi.org/10.1155/2020/8885861

[8]. X. Chen, P. Cong and S. Lv, "A Long-Text Classification Method of Chinese News Based on BERT and CNN," in IEEE Access, vol. 10, pp. 34046-34057,2022.DOI:10.1109/ACCESS.2022.3162614.

[9]. "NEWS CLASSIFICATION USING MACHINE LEARNING", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.7, Issue 3, page no. pp657-659, March-2020.

[10]. Ningfeng Sun, Chengye Du, "News Text Classification Method and Simulation Based on the Hybrid Deep Learning Model", Complexity, vol. 2021, Article ID 8064579, 11 pages, 2021. https://doi.org/10.1155/2021/8064579.