Performance Analysis of Naïve Bayes for News Text Classification

Dharmendra Thapa[1] | Madhav Dhakal[2,*]

[1]Central Department of Computer Science and Information Technology, Tribhuvan University, Kathmandu, Nepal, Email: dharmendrathapa3@gmail.com,

[2]Graduate School of Science and Technology, Mid-West University, Surkhet, Nepal,
Email: madhav.dhakal@mu.edu.np

* **Correspondence Email:** *madhav.dhakal@mu.edu.np*

Keywords:
News, Multinomial Naïve Bayes classifier, News Classification.

Received: 6 November 2024
Revised: 27 November 2024
Accepted:23 December 2024

ISSN: 3102-0763 (Print)
3102-0771 (Online)

Copyright: @Author(s) 2025

Abstract

This study proposes a multinomial Naïve Bayes Classifier technique for identifying news categories and analyzing their classification. To achieve the objective, 1490 data (BBC News) are employed, including 336 business categories, 261 technology categories, 274 politics categories, 346 sports categories, and 273 entertainment categories. The datasets' performance is evaluated using accuracy, recall, precision, and the F1-score. These data are utilized to train the model, resulting in 98.85% and 97.09% accuracy on train and test data, respectively, with an 80-20 split.

1. Introduction

Online news platforms organize their content into various categories like "Politics", "Sports", "Entertainment", and others. However, manually assigning these labels becomes increasingly challenging when a large volume of news arrives from multiple sources. In situations where applications aim to deliver real-time trending news to users, manual classification becomes practically unfeasible. Therefore, there is a strong need for an automated system that can accurately categorize news articles into appropriate classes. Moreover, such a classification technique can be effectively applied to other types of text documents as well (Shah et al., 2020).

Automatic classification of text into predefined categories is considered a key approach for handling and organizing large volumes of data. This textual information is generated from diverse sources, including conference proceedings, editorials, online articles, websites, emails, periodicals, and academic journals. Nowadays, many people prefer accessing information through these digital platforms rather than relying on traditional sources like newspapers, magazines, or books. Although access to information has become more convenient, the challenge of effectively organizing this knowledge persists, making data management difficult. Therefore, structuring such digital content is crucial for the efficient categorization of information(Singh et al., 2021).

This study evaluates most widely used machine learning techniques i.e. Naïve Bayes for automatic news classification problem. To experiment the system, BBC news articles with 5 different categories and total 1490 documents collected from BBC news portal is used. Stop words, special characters, symbols are removed which is present in the article. After that tokenization process occurs to convert text into individual tokens. The BBC news dataset is retrieved from kaggle. After retrieving data from kaggle, data are pre-processed. Naïve Bayes classifier is popular for text classification. Due

to multiple category classes, multinomial Naïve Bayes is preferred among other Naïve Bayes classifier (i.e. Bernoulli and Gaussian).

The Naive Bayes algorithm is based on Bayes' theorem. It gives a formula for computing the conditional probability P(A|B);.

According to Bayes, the following formula may be used to compute the conditional probability:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

2. Problem Statement

The vast majority of textual data is unstructured, making it difficult and time-consuming to organize, modify, and manage. Automatic news categorization has gained popularity among news organizations in recent years due to its speed and low cost. This study presents a machine learning architecture that uses the Naïve Bayes classifier to categorize BBC news stories into business, technology, entertainment, politics, and sports. A dataset of 1490 news stories was used.

3. Objective

The objectives of this study are mentioned as:

- To classify news into their respective category.
- To evaluate the performance of multinomial Naïve Bayes classifier for news classification on the chosen datasets.

4. Literature Review

In (Ahmed & Ahmed, 2021), the authors reviewed several existing methods for classifying online news content and introduced a framework for the automated categorization of news articles. They evaluated multiple classifiers to achieve high classification accuracy. Their experimental approach, using a Bayesian classifier, reached an accuracy rate of 93% and included the presentation of confusion matrices.

(Barua et al., 2021) explored several widely used machine learning algorithms—namely Logistic Regression, Support Vector Classifier (SVC), Decision Tree, Multinomial Naïve Bayes, and Random Forest—for the task of automatic sports news classification in Bengali, utilizing Term Frequency-Inverse Document Frequency (TF-IDF) features. In response to the absence of a standard benchmark dataset, the authors also developed a Bengali News Corpus (BNeC) consisting of 43,306 news articles with 202,830 unique terms across categories such as cricket, football, tennis, and athletics. Experimental results revealed that the Support Vector Classifier, using a unigram+bigram+trigram feature set, achieved the highest weighted F1-score of 97.60%, outperforming other classifiers and feature combinations.

In the study (Abd et al., 2020) used Naïve Bayes (NB) to analyze opinions in a text and classify them as Reform, Conservative, or Revolutionary. This study examines how employing two feature extraction approaches, Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF), together with Naïve Bayes, affects the accuracy of Arabic article classification. Precision, recall, F1-score, and number of accurate predictions were utilized to assess the performance of the applied classifiers.The findings show that utilizing TF with TF-IDF increased accuracy to 96.77%.

The authors in (Kaur & Khiva, n.d.) used a machine learning approach to analyze NN classifiers. Text categorization aims to group articles into predetermined categories. Archives can be classified as none, several, or only one. Four categories have been proposed, including politics, finance, and sports. A

neural network classifier was used to implement the classification procedure. Simulation findings indicate that NN achieves 99.93 accuracy and outperforms previous approaches.

5. Methodology

5.1. Dataset

For this study, news articles are collected from kaggle. It contains 1490 data which is classified into 5 categories in which 336 to business, 261 to tech, 274 to politics, 346 to sports and 273 to entertainment are represented in Table 1.

Table 1: Dataset of news articles

ArticleId	Text	Category
1833	worldcom ex-boss launches defence lawyers defending former worldcom chief bernie ebbers against a battery of fraud charges have called a	business
154	german business confidence slides german business confidence fell in february knocking hopes of a speedy recovery in europe s largest econ	business
1101	bbc poll indicates economic gloom citizens in a majority of nations surveyed in a bbc world service poll believe the world economy is worseni	business
1976	lifestyle governs mobile choice faster better or funkier hardware alone is not going to help phone firms sell more handsets research sugges	tech
917	enron bosses in $168m payout eighteen former enron directors have agreed a $168m (Â£89m) settlement deal in a shareholder lawsuit over th	business
1582	howard truanted to play snooker conservative leader michael howard has admitted he used to play truant to spend time with his school frier	politics
651	wales silent on grand slam talk rhys williams says wales are still not thinking of winning the grand slam despite a third six nations win. that s	sport
1797	french honour for director parker british film director sir alan parker has been made an officer in the order of arts and letters one of france s h	entertainment
2034	car giant hit by mercedes slump a slump in profitability at luxury car maker mercedes has prompted a big drop in profits at parent daimlerchry	business
1866	fockers fuel festive film chart comedy meet the fockers topped the festive box office in north america setting a new record for christmas day.	entertainment
1683	blair rejects iraq advice calls tony blair has rejected calls for the publication of advice on the legality of the iraq war amid growing calls for an i	politics
1153	housewives lift channel 4 ratings the debut of us television hit desperate housewives has helped lift channel 4 s january audience share by 12	entertainment
1028	uk coal plunges into deeper loss shares in uk coal have fallen after the mining group reported losses had deepened to Â£51.6m in 2004 from Â	business
812	bp surges ahead on high oil price oil giant bp has announced a 26% rise in annual profits to $16.2bn (Â£8.7bn) on the back of record oil prices.	business
707	ireland 21-19 argentina an injury-time dropped goal by ronan o gara stole victory for ireland from underneath the noses of argentina at lansdo	sport
1588	wenger signs new deal arsenal manager arsene wenger has signed a new contract to stay at the club until may 2008. wenger has ended specul	sport
342	u2 s desire to be number one u2 who have won three prestigious grammy awards for their hit vertigo are stubbornly clinging to their status a	entertainment
486	hantuchova in dubai last eight daniela hantuchova moved into the quarter-finals of the dubai open after beating elene likhotseva of russia 7-	sport
1344	melzer shocks agassi in san jose second seed andre agassi suffered a comprehensive defeat by jurgen melzer in the quarter-finals of the sap o	sport
1552	moving mobile improves golf swing a mobile phone that recognises and responds to movements has been launched in japan. the motion-sen	tech
1547	hewitt overcomes wobble in sydney lleyton hewitt gave himself the perfect preparation for next week s australian open with victory over ivo	sport
177	carry on star patsy rowlands dies actress patsy rowlands known to millions for her roles in the carry on films has died at the age of 71. rowlan	entertainment
1785	serena becomes world number two serena williams has moved up five places to second in the world rankings after her australian open win. v	sport

5.2 Pre-processing

Pre-processing phase deals with the tokenization process, word stemming, removal of special characters, tags, symbols, etc. Dataset is split into train data and test data where 80% is training data and 20% is test data. Each category is uniquely labelled (i.e. 0 for business, 1 for tech, 2 for politics, 3 for sports and 4 for entertainment) as shown in Table 2.

Table 2: Labeled category

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

5.3 Naïve Bayes Classifier

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combination of values. Naïve Bayes technique used Bayes theorem to determine that probabilities of occurring data in particular category.

$$P(B) = \frac{P(A).P(A)}{P(B)}$$

Where A and B are events and P(B) ≠ 0.

The Naïve Bayes classifier comprises three main variants: Multinomial, Bernoulli, and Gaussian. Among these, Multinomial Naïve Bayes is a widely used supervised learning algorithm, particularly effective for handling categorical text data, and is commonly adopted as a baseline in text classification tasks. It is well-suited for multiclass classification problems, as it computes the probability of each class for a given input and assigns the class with the highest likelihood. In contrast, the Bernoulli model is designed for binary features, determining whether a feature is present or absent, while the Gaussian model is used for continuous data distributions. Given that news classification typically involves multiple categories, the Multinomial Naïve Bayes classifier is generally favored over the other two variants

For example: 6 news articles of two categories are taken (i.e. Sport and Business)

A1: match is interesting	Sport
A2: football share is interesting	Business
A3: game is closed	Sport
A4: today football match is interesting	Sport
A5: today market is closed	Business
A6: today match closed	Sport

Table 3. Categories of News

P(Sport)=4/6 P(Business)=2/6		
Sport	Business	
3/4	0/2	match
2/4	2/2	is
2/4	1/2	interesting
2/4	1/2	today
0/4	1/2	share
0/4	1/2	market
1/4	0/2	game
2/4	1/2	football
1/4	0/2	wins
1/4	1/2	closed

New article news: "football is interesting"

P(football is interesting | Sport)= P(0,1,1,0,0,0,1,0,0)

$$= \left(1-\frac{3}{4}\right)\left(\frac{2}{4}\right)\left(\frac{2}{4}\right)\left(1-\frac{2}{4}\right)\left(1-\frac{0}{4}\right)\left(1-\frac{0}{4}\right)\left(1-\frac{1}{4}\right)\left(\frac{2}{4}\right)\left(1-\frac{0}{4}\right)\left(1-\frac{1}{4}\right)$$

$$=0.0088$$

P(football is interesting | Business)= P(0,1,1,0,0,0,1,0,0)

$$=\left(1-\frac{0}{2}\right)\left(\frac{2}{2}\right)\left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\left(1-\frac{0}{2}\right)\left(\frac{1}{2}\right)\left(1-\frac{0}{2}\right)\left(1-\frac{1}{2}\right)$$

$$=0.016$$

P(Sport | football is interesting)= $\frac{\frac{0.0088*4}{6}}{\frac{0.0088*4}{6}+\frac{0.016*2}{6}}$ = 0.52

- This is sport article news.

6. Performance Measures

Performance of model can be measured using different performance factor like Accuracy, Precision, Recall and F1 Score. These factors are explained below:

6.1. Accuracy Score

In classification, Accuracy Score is the ratio of correct predictions to the total number of input data points.

Accuracy Score = (TP + TN)/ (TP + FN + TN + FP)

In which TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false-negative instances, the number of false-positive instances and the number of true negative instances.

6.2. Precision

Precision is the ratio of number of True Positive to the total number of Predicted Positive. It measures out of the total predicted positive, how many are actually positive.

Precision Score = TP / (FP + TP)

In which TP and FP refer respectively to the number of true positive instances and the number of false-positive instances.

6.3. Recall

Recall is the ratio of number of True Positive to the total number of Actual Positive. It measures out of the total actual positive, how many are predicted as True Positive.

Recall Score = TP / (FN + TP)

In which TP and FN refer respectively to the number of true positive instances and the number of false-negative instances.

6.4. F1 Score

F1 Score is an important evaluation matric for binary classification that combines Precision & Recall. F1 Score is the harmonic mean of Precision & Recall.

F1 Score = 2* Precision Score * Recall Score/ (Precision Score + Recall Score)

7. Implementation

The dataset is retrieved from kaggle and the implementation is carried out using python and its library. They are:

Panda: In this report panda is used to import/load the dataset as well as evaluating the nature of dataset. It is also used to remove unnecessary data and null data present in dataset.

Sklearn: This library is used for many purposes like splitting data, train model and testing data. Sklearn libary include different module for different proposed among them following are used during implementation.

Model_selection: This module is used to split data into train and test data using train_test_split() method.

Metrics: This module is used to analyze the performance of this model. The performance analysis include accuracy, precision, recall and f1 score with its respective classes.

Multinomial NB classifier: MultinomialNB() method is used to train and classify news article category.

Nltk: nltk library is used to remove stop words, special characters, tags and to convert all the text into lowercase.

Matplotlib: matplotlib is used to visualize the data (i.e. in bar graph).

8. Result

predict() method is used to classify news category, which is provided by MultinomialNB(). When new article is needed to be predicted then the input article is preprocessed and passed to a predict() method which returns result of category.

8.1 Train Data Analysis

Out of 1490 datasets, there are 1043 training data in which business category contains 233 data, tech(184), politics(193), sports(249) and entertainment(184) data. Out of total train data of each category all are not correctly predicted which is shown below:

Table 4. Categories of Train data

Category	Correct Prediction	Other
Business(233)	226	3(tech), 4(politics)
Tech(184)	183	1(entertainment)
Politics(193)	192	1(tech)
Sports(249)	247	1(business), 1(entertainment)
Entertainment(184)	183	1(tech)

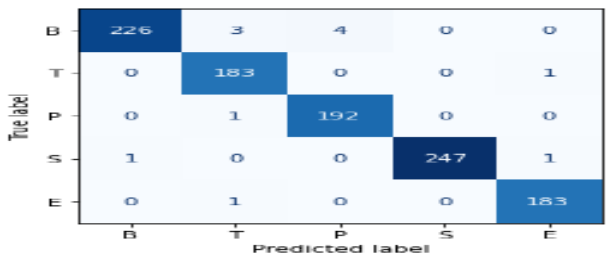

Fig. 1: Confusion Matrix of Training Data

The achieved accuracy, precision, recall and f1 score of the train model are 98.85%, 98.85%, 98.85% and 98.85% respectively with the help of confusion matrix.

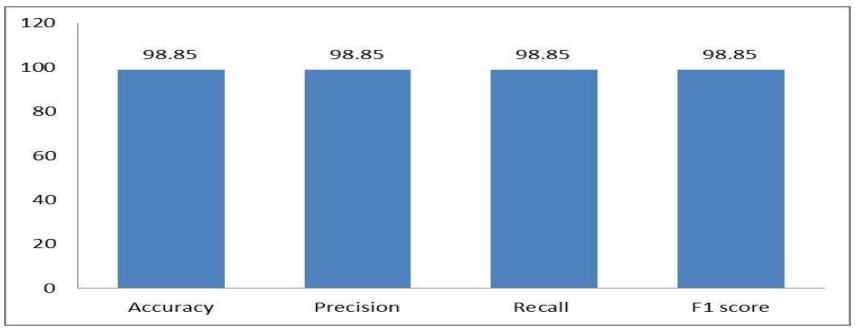

Fig.2: Performance Analysis on Training Data

8.2 Test Data Analysis

Out of 1490 datasets, there are 447 test data in which business category contains 103 data, tech(77), politics(81), sports(97) and entertainment(89) data. Out of total test data of each category all are not correctly predicted which is shown below:

Table 5. Categories of test data

Category	**Correct Prediction**	**Other**
Business(103)	96	5(tech), 2(politics)
Tech(77)	75	2(entertainment)
Politics(81)	78	1(politics),1(sports),1(entertainment)
Sports(97)	97	0
Entertainment(89)	88	1(politics)

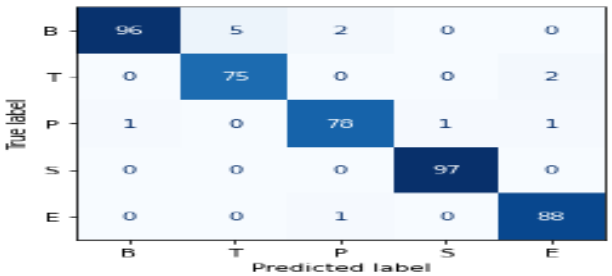

Fig.3: Confusion Matrix of Test Data

The achieved accuracy, precision, recall and f1 score of the train model are 97.09%, 97.09%, 97.09% and 97.09% with the help of confusion matrix.

Table 6: Performance Table of Test Data

Accuracy	Precision	Recall	F1 score
97.09%	97.09%	97.09%	97.09%

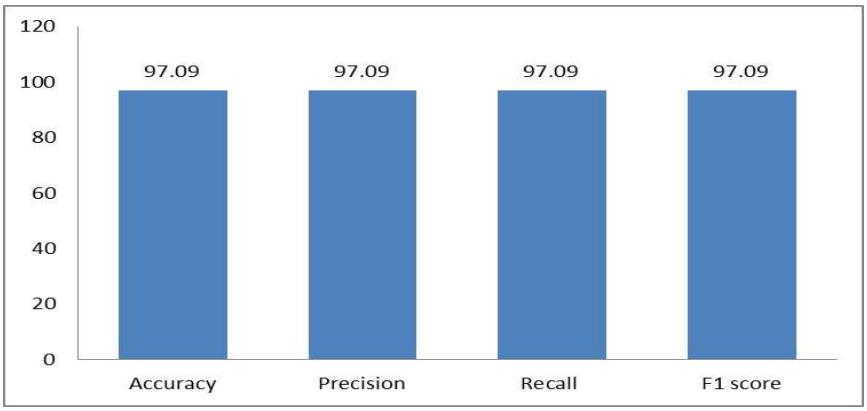

Fig. 4: Performance Analysis on Test Data

9. Conclusion

After data preprocessing (cleaning data, train_test_split model) and using Multinomial Naïve Bayes algorithm, it is concluded that the news are classified to their respective category with 97.09% accuracy, precision, recall and f1 score. The findings of this study demonstrate that Naïve Bayes can effectively categorize news articles into their respective categories.

References

Abd, D. H., Sadiq, A. T., & Abbas, A. R. (2020). Political Articles Categorization Based on Different Naïve Bayes Models. In M. I. Khalaf, D. Al-Jumeily, & A. Lisitsa (Eds.), *Applied Computing to Support Industry: Innovation and Technology* (Vol. 1174, pp. 286–301). Springer International Publishing. https://doi.org/10.1007/978-3-030-38752-5_23

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

Barua, A., Sharif, O., & Hoque, M. M. (2021). Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation. *Procedia Computer Science*, *193*, 112–121. https://doi.org/10.1016/j.procs.2021.11.002

Kaur, S., & Khiva, N. K. (n.d.). *Online news classification using Deep Learning Technique*. *03*(09).

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1), 12. https://doi.org/10.1007/s41133-020-00032-0

Singh, Y. V., Naithani, P., Ansari, P., & Agnihotri, P. (2021). News Classification System using Machine Learning Approach. *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 186–188. https://doi.org/10.1109/ICAC3N53548.2021.9725409