第 48 卷第 9 期
2021 年 9 月

湖南大学学报（自然科学版）
Journal of Hunan University（Natural Sciences）

Vol. 48. No. 9.
September 2021

Open Access Article

BBC NEWS DATA CLASSIFICATION USING NAÏVE BAYES BASED ON BAG OF WORD

Hanan Abbas Salman
Department of Computer Systems, Al-Furat Al-Awsat Technical University, Najaf Technical Institute, Iraq

Tameem Hameed Obaida
Department of Computer Systems, Al-Furat Al-Awsat Technical University, Najaf Technical Institute, Iraq

Abstract

Sentiment analysis is used in practically every aspect of human life and has a substantial impact on our actions. There is a vast amount of data that presents people' opinions in numerous domains, such as business and politics, thanks to the expansion and use of online technology. Naïve Bayes (NB) is utilized in this paper to examine opinions by text classification and categorizing them into the relevant category (business, entertainment, politics, sprot, and tech). It examines the impact of combining NB classifiers (Multinomial, Gaussian, Bernoulli, and Complement) in conjunction with feature extraction methods in such of frequency-inverse document frequency (TF-IDF) on an accuracy of classifying BBC news data. Some techniques were used to measure the performance of classifiers such as recall, precision, and F1-score. The outcomes show that the Complement scored the highest in accuracy, with a score of 97.604 percent. The Complement was found to be the best fit.

Keywords: Multinomial, Gaussian, Bernoulli, Complement, Term Frequency, Sentiment Analysis, Opinion Mining

抽象的

情感分析几乎用于人类生活的各个方面，并对我们的行为产生重大影响。由于在线技术的扩展和使用，大量数据在商业和政治等众多领域呈现人们的观点。本文使用朴素贝叶斯 (NB) 通过文本分类来检查意见，并将其分类为相关类别（商业、娱乐、政治、sprot 和技术）。它检查了将 NB 分类器（多项式、高斯、伯努利和补）与频率逆文档频率 (TF-IDF) 等特征提取方法相结合对 BBC 新闻数据分类精度的影响。一些技术被用来衡量分类器的性能，如召回率、精度和 F1 分数。结果显示，Complement 的准确率最高，达到 97.604%。发现 Complement 是最合适的。

关键词：多项式，高斯，伯努利，补，词频，情感分析，意见挖掘

Introduction

In recent years, the usage of resources on the internet such as social networking, personal blogs, online review, and other similar sites has increased, allowing users to show or discuss their thoughts, ideas, opinions, and remarks on a

Received: July 02, 2021 / Revised: August 22, 2021 / Accepted: September 05, 2021 / Published: September 20, 2021
About the authors: Bushra Hameed
Corresponding author- Dr Yasir Arafat Elahi

variety of topics. It's critical to collect and analyse these remarks in real-life scenarios. For example, before purchasing a service or product, every customer would like to hear what other customers have to say. Similarly, a corporation wants to know what the consumer thinks so that they may improve and modify their product to meet his needs.

Because the number of subjective texts on forums, blogs, and social media has increased since the 2000s, several academics have employed a text classification technique as opinion classification [1]. opinions analysis, subjectivity analysis, review mining, opinion extraction, and opinion mining were all terminology used by a few studies for sentiment classification [2].

E-commerce, education, and opinion polls have all profited from sentiment analysis [3]. The organization analyses customer evaluations and watches social media to identify public perceptions of their services and products and take relevant action in a timely manner. According to a few academics, stock prices and social media sentiment analyses are linked, and future stock prices can be forecasted using opinion from Twitter [4].

A small number of words that reflect the opinions may be negative or positive are included in lexicon of the writer opinion. However, due to an opposite orientation of words in distinct domains, utilizing the sentiment lexicon for sentiment categorization is insufficient. Liu has provided additional information on the challenges surrounding the use of sentiment lexicon (2012). For sentiment categorization, a variety of machine-learning approaches are employed.

Machine learning was found to outperform lexicon-based approaches in previous investigations [5, 6]. Due to the parameter's versatility, ease of management, and high accuracy. In machine learning procedure, the text is first analysed as a collection of words before being turned into numerous characteristics. After that, classification is done using machine learning techniques such as the Naive Bayes (NB) algorithm [1, 7]. Sentence level, document level, and aspect level are the three stages of sentiment analysis [2]. This research focuses on level opinion analysis of document to find out classes and categorize it.

NB is a machine learning of type supervised learning technique which believes the whole of features are equally important with independent. NB is often utilized in the classification of documents. The machine learning structures such as multinomial and multi-variate Bernoulli are utilized in a text categorization without generalization. Even if the characteristics were not independent, the NB technique can be utilized to solve a variety of problems are related with machine learning. NB was believed to be a direct calculation dependent on probabilistic hypothesis [5] in comparison to several sophisticated machine learning techniques.

This paper's reminder is organized as follows. In part 2, NB classification will be discussed. The dataset is depicted in Section 3. Section 4 introduces the evaluation matrices. The model architecture is depicted in 5, followed by Section 6's explanation of the findings of our technique. Finally, in Section 7, we address our findings and plans for the future.

2 Naïve Bayes Classification Algorithm

The researchers used a utilized a conditional probability, as indicated in equation (1).

$$\mathrm{P.(c.|.d)}{-}= \frac{\mathrm{.P\ (d\ |\ c).P.(\ c\ )}}{-.\mathrm{P(.d.).}}. \qquad (1)$$

d is for the document to be classified, and c stands for the document class [8]. P(d) was deleted because it did not exhibit a meaningful effect in the real world. The Bays rule was devised by the researchers and is described in equation (2).

$$\mathrm{P\ (\ c\ |\ d\ ) =\ p\ (d\ |\ c)\ P(\ c\ )} \qquad (2)$$

They estimated the document's features or words [9] for each document using the following equation:

$$\begin{aligned} &\mathrm{P\ (\ c\ |\ d\ )} \qquad (3) \\ &= \mathrm{P\ (x_1,\ x_2, x_3, \ldots., x_n\ |\ c\ )\ P(\ c\ )} \end{aligned}$$

where d denotes the document's features from x_1 to x_n. The researchers utilized equation (3) to calculate each word in a vector in the document. The proposed model was integrated with the popular Maximum A Posteriori (MAP) decision algorithm in the NB classifier. The most likely hypothesis was chosen using this rule as equation (4).

$$\mathrm{y} = \mathrm{P(c_k)} \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(x_i|c_k)} \qquad (4)$$

where y is the final test class that it previously differentiates with the test label in the data. For finding the minimum or maximum y based on the naïve bays used, the researchers used equation (4). They also used the two formulae below to determine the values for $P(c_k)$ and $P(x_i|c_k)$. Where P (c_k) the identical each of NB types whereas $P(x_i|c_k)$ varies between algorithms depending on the classifier used.

$$\mathrm{P(c_k)} = \frac{\mathrm{documentP(c = c_k)}}{\mathrm{N_{doc}}} \qquad (5)$$

(6)

$$\mathrm{P(x_i|c_k)} = \frac{\mathrm{count\ (x_i|c_k)}}{\sum_{\mathrm{x \in v}} \mathrm{count(x|c_k)}}$$

The main downside of the NB was that the probability value based on frequency would be 0 if there isn't a single instance that related to class label and the trusted attribute value. For fixing this issue, the researchers employed the Laplace smoothing approach by adding 1 to equation (7).

$$\mathrm{P(x_i|c_k)} = \frac{\mathrm{count(x_i|c_k) + 1}}{\sum_{\mathrm{x \in v}} \mathrm{count(x|c_k) + 1}} \qquad (7)$$

This study used four different forms of NB, which are detailed below **[10]**.

2.1. Gaussian Naïve Bayes

The values of the numeric attributes in the Gaussian NB [11] were regularly distributed. The mean and standard deviation were used to depict this distribution. This aids in the computation of the likelihood of observed values based on guesses. The following formula was used to calculate the probability of the features:

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} exp(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}) \qquad (8)$$

where mean (μ) is the average and standard deviation (σ) is the standard deviation. The equation (4) was utilized. The binning technique is used to discretize the attribute values for continuous data.

2.2. Multinomial Naive Bayes

The information in the document's word frequency was acquired by the multinomial NB classifier. Because the multinomial distribution required integer features, it was well suited for discrete feature classification. [12].

$$P(x_i|c_k) = \frac{n!}{\prod_{i=1}^{|V|} x_i!} \prod_{i=1}^{|V|} P(x_i|c_k)^{x_i} \qquad (9)$$

$$P(x_i|c_k) = \prod_{i=1}^{|V|} P(x_i|c_k)^{x_i}$$

where V is the number of features. Normalization of $\frac{n!}{\prod_{i=1}^{|V|} x_i!}$ was unaffected by class k in the model [13]. After adding the V parameter, $P(x_i|c_k)$ was calculated using Eq. (7) [12, 14]. The Laplace equation was applied to avoid the zero-frequency problem.

$$-P\ (x_i|c_k) = \frac{count(\ x_i \mid c_k) + 1}{(\ \sum_{x \in v} count(x \mid c_k)\ ) + |R|} \qquad (10)\text{-}$$

where R is the vocabulary included within training dataset labels.

2.3. Complement of the Naïve Bayes (CNB)

Rather than estimating the chance of a word being in the class, the Complement Normal NB approach calculated the probability of it occurring in other classes [15]. Other classes' word-class dependencies P(x_i |c'_k) were so approximated. For the reverse class that was chosen, the researchers established the minimal value of y and c'_k.

$$y = P(c_k) \prod_{i=1}^{n} \frac{1}{P(x_i|c'_k)} \qquad (11)$$

We seek for a minimal y value and c'_k the reverse class we want to use.

2.4. Bernoulli of the Naive Bayes (BNB)

Assuming the binary system has two values as features so a classifier of Bernoulli naïve Bayes was built on that assumption features [16]. The distribution of Bernoulli equation is as follows:

$$P(x)\ = -P^x(.1 - P.)^{.1-x.} - \qquad =(.12.).$$

The Bernoulli distribution value ranging from zero to one was used as x. If it was 0, it was a failure, whereas if it was 1, it was a success as following equation [17]:

$$P(x = 1) = .\,P^1\ (.1 - P.)^{.1-1.} \qquad (13)$$
$$=.\,p -$$
$$P(x = 0) = .\,P^0\ (.1 - P\ .)^{1-0.}$$
$$=.\,(\ 1 - p\ ) -$$

The probability of the word not appearing in the class document was (1-p(x_i | c)), where x was a document word [18]. As a result, the following equation was created:

$$.\,\mathrm{P}\ (\ \mathrm{x_i} \mid \mathrm{c.}\ ).=.\,\mathrm{P}\ (\ \mathrm{x_i}|\mathrm{c}\ )\ .\,\mathrm{b_i}.+.\,(\ 1 - \mathrm{b_i}.\ ).\,(\ 1 - \mathrm{p}\ (\ \mathrm{x_i} \mid \mathrm{c.}\ )) \qquad (14)$$

Where: $x_{i=}$word, b_i =document

All of the terms can be used with this product. If x_i appeared in the b_i=1 and P(x_i|c) was the likelihood. b_i=0 if the word x_i is missing, and the probability is (1 − p(x_i|c)). As a result, Eq. (4) was utilized to calculate y.

3 Dataset Descriptions

We employed corpus in this study to test the provided strategy for finding the best answer. For this test, the BBC news dataset [19]

was utilized; for additional information, see the dataset on UCD[1].

This dataset is produced in English and is based on information gathered from BBC news websites between 2004 and 2005.

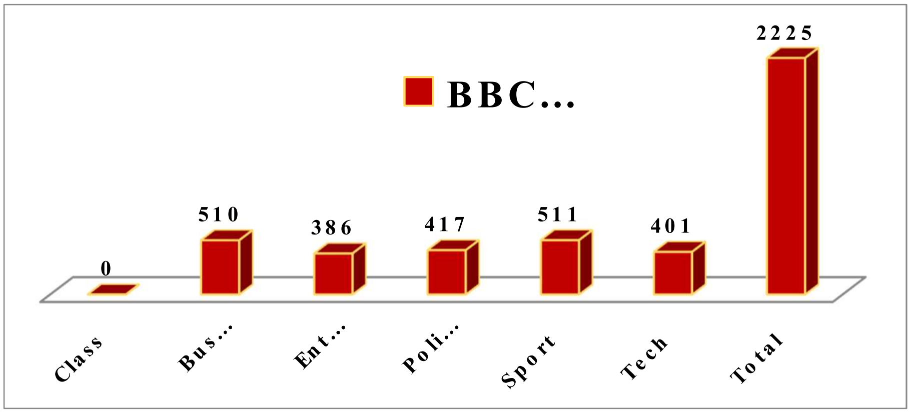
Table 1 shows the five classifications that were present in the dataset. There are 2225 items in this collection, which include a lot of articles.

TABLE 1: BBC DATASET

Class	**Total**

[1] Mlg.ucd.ie/datasets/bbc.html

. Business.	-510-_
. Entertainment.	-386-_
. Politics.	-417-_
. Sport. _	-511-_
. Tech._	-401-_
. Total. _	**-2225-_**



The above distribution of documents using PCA and T-SNE is shown in Figure 1, indicating that the BBC news dataset is non-linear. Every class had a well-separated data distribution. The BBC news dataset is non-linear in general, however it can be used with machine learning, as shown in the results section.

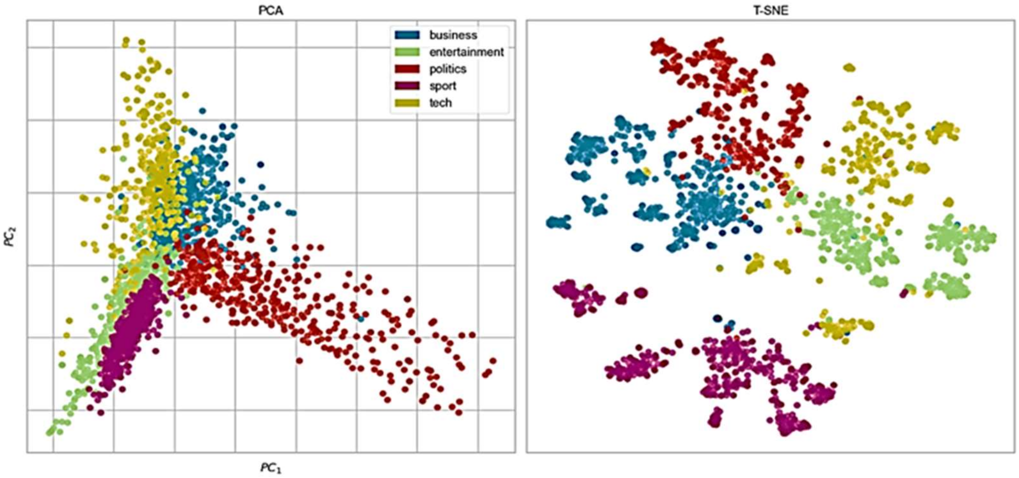


Figure 1: PCA and T-SNE were used to distribute the BBC news dataset.

4 Evaluation Metric

Precision, Recall, F1-score, and Accuracy are specified and calculated to assess the performance of our study's model. A variety of equations are used in this section to examine and quantify the degree of evolution that our approach has undergone, with the results shown in Table 2. These equations have parameters like

TP, TN, FN, and FP which must be intruded and their precise values computed. Positive value was appropriately assessed by TP suggest that its anticipated class value is yes, as well as its real class value. This means that when a class is predicted to be the same as the test class, it is expected to be true, however when the TN predicted class is the same as the test class, it is expected to be false. If the real FP class is positive, the predict class is FN negative. Furthermore, the error in our approach can be determined when the real class is positive but the projected class is negative. The formulations are as Table 2:

TABLE 2: EQUATIONS IN METRIC UNITS

Metric name	**Equation**
Accuracy-_	$\frac{-\mathrm{TP} + \mathrm{TN} -}{\mathrm{TP.} + \mathrm{TN.} + \mathrm{FP.} + \mathrm{FN}}^{-}$
Recall-_	$\frac{.\,\mathrm{TP.}}{.\,\mathrm{TP} + \mathrm{FN.}}^{-}$
Precision-_-	$.\frac{.\,\mathrm{TP.}}{\mathrm{TP} + \mathrm{FP.}}.^{-}$
F1-score-_	$.\,2 * \frac{\mathrm{Recall} * \mathrm{Precesion}}{\mathrm{Recall} + \mathrm{Precesion}}^{-}$

To calculate accuracy, use the Table 1. This equation is necessary for the model to work properly. The ratio of precisely calculated positive observations to overall projected positive observations is referred to as precision. Recall, on the other hand, is the proportion of appropriately calculated positive observations to total observations in real class. F1-Score is defined as the weighted average of Precision and Recall. As a result, this Score considers both false negatives and false positives. Accuracy is a challenging concept to grasp intuitively. F1-score is, on the other hand, often preferable to accuracy, especially if the class distribution is non-uniform. When the costs of false negatives and false positives are comparable, accuracy works best. However, if the costs of false negatives and false positives differ significantly, both Recall and Precision should be considered.

5 Proposed Approach

The goal of this research was to find the optimal representation of the BBC news dataset. The impact of the feature extraction approach on classification performance was also explored by the researchers (decreasing, increasing, or maintaining). The method for determining the orientation's classification of BBC news datasets was described here by the researcher such as classifiers used, text models..etc. Lexical based, machine learning, or hybrid approaches that were produced after merging the prior techniques were utilized to understand text classification. The researchers used machine learning techniques in this study, which required a labelled dataset to train and evaluate their classifier.

The methodology for categorizing the orientation of the BBC news dataset was described by the researcher in this paper (datasets, text models, and classifiers, for example). Machine learning procedures, lexical or hybrid approaches generated from merging the prior methodologies were utilized to better understand text classification. The researchers in this study employed machine learning techniques to train and evaluate their classifier, which required a labelled dataset.

The NB classifier was a widely used method for text classification. This was a decision rule of probabilistic model. The value that represent of the single feature in NB was thought to be independent of the values of other features, hence the name nave. This algorithm calculated the document's posterior likelihood of belonging to a different class, and it was assigned to the one with the highest posterior probability. The model for this investigation was shown in Figure 2.

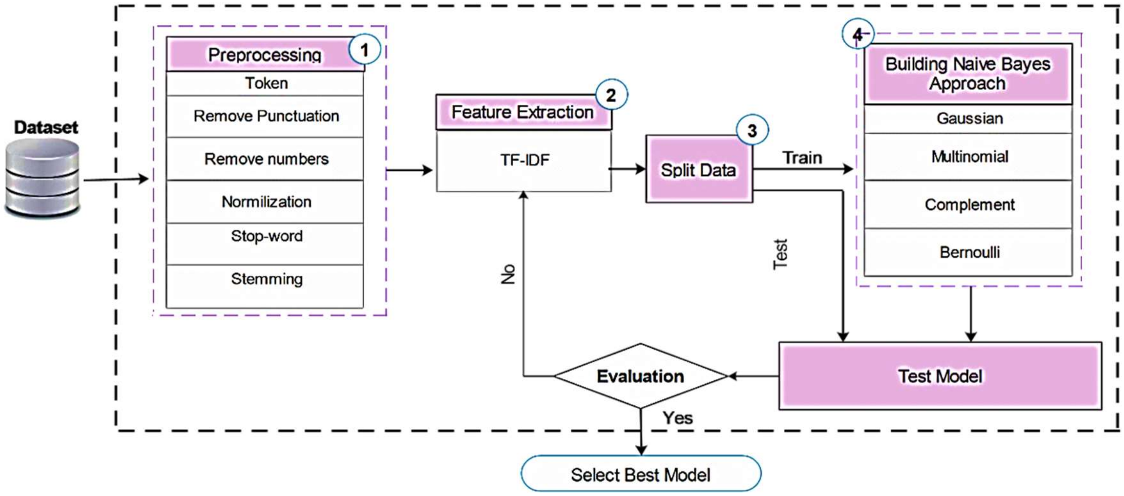


FIGURE 1: APPROACH USED IN THE STUDY

5.1. Pre-processing

To prepare and clean the text in preparation for further classification, the data must be pre-processed. The extraction of words is hampered by the presence of noise and uninformative data such as HTML tags, scripts, and adverts in online texts. Furthermore, the presence of English characters, improper spellings, or punctuation marks cannot be determined with accuracy. Furthermore, numerous words in the English content have no impact on the content's arrangement. A classification becomes more complex by preserving these terms because each word was handled as a separate dimension. There are several steps in the classifying process.

Tokenization: The text of a document is split into a sequence of tokens in this step. Typically, data acquired from online evaluations is coupled with noises such as HTML tags, URLs, advertising, scripts, and symbols such as hashes, asterisks, and other symbols that are neither relevant nor useful in classification. To increase the performance of the classifier, these noise and symbols must be removed, leaving only the words.

Remove unwanted characters: This step removes unwanted characters or strings from the text, such as non-English characters, English numerals, hashtags, punctuation marks, and so on. This task necessitates the use of a large number of regular expressions, as seen in Table 3.

TABLE 3: LIST OF REGULAR EXPRESSION

Regular expression-_	-Outcomes-_
-[0-9]+_	Remove English numbers-
[#-_?,.';]+_	Remove all punctuation marks-_

Normalization: converts all of the characters in a document to uppercase or lowercase. The majority of the evaluations employ a mix of capital and lowercase characters. The entire document collection is converted to lowercase in this process. It also aids in the removal of unnecessary characters from a few words, such as "soooon" being replaced with "soon" and "gooooood" being replaced with "good."

Stop-words removal: This function allows you to filter out English stop words from a review by removing each token and comparing it to the built-in stop-words list. Stop words are words that aren't particularly crucial to the opinion or sentence. e.

Stemming, entails removing affixes from a word in order to make it more brief by utilizing the fewest possible words while maintaining the meaning. Porter stemmer is used in the put forward system.

5.2. Feature Extraction

The features were extracted from the papers in the next stage. The weight of a word (feature) was determined in relation to the document that contained the word. Terms were used in this section Frequency weighting, Frequency-Inverse Document Frequency and the Inverse Document Frequency weighting (TF), (TF-IDF), and (IDF) respectively have all been utilized.

The frequency of a term in a document is determined by TF. Because the length of each document varied, it was likely that a term would appear more than once in the larger texts compared to the shorter records. As a result, as a normalization procedure depend on below an equation, the TF was divided by the document length in other words, the amount number of phrases in one document.

$$\mathrm{TF}(\mathrm{t}, \mathrm{d_i}) = \frac{\mathrm{N_{t,i}}}{\sum_{\mathrm{k}=1}^{|\mathrm{T}|} \mathrm{N_{k,i}}} \qquad (15)$$

Where $TF(t,d_i)$ denotes the TF of the word t in the document d_i; $N_{t,i}$ denotes the number of words t in the document d_i; and $N_{k,i}$ denotes the number of words in the document d_i.

Because it merged two prior methods of TD and IDF [20], TF-IDF is a prominent technique for identifying and retrieving information from texts. TF-IDF was calculated as the product of the TF and IDF values. It also assisted in determining the meaning of the word in the collection's text.

$$.\,\mathrm{IDF}. = .\log\left(\frac{.\,\mathrm{N}}{\mathrm{df}}\right) \qquad (16)$$

where N is the number of documents; df is the number of documents that include the phrase.

Because it detects the word in the same document or corpus, TF-IDF is quite advanced and provides the best results. As a result, the TF-IDF was determined using the Eq (17).

$$w(t, d_i) = TF(t, d_i) * IDF \quad (17)$$

where $w(t, d_i)$ denotes the weight of the term t in the document d_i. This method assisted in reducing the weight of all features found in a large number of documents that were chosen for feature selection in order to reduce the vector size and then exposed to machine learning methods [21].

5.4. Building Naïve Bayes Model

Created a model for each NB type. We used the TF-IDF feature extraction to create four models. Every model was put to the test, the results were calculated, and the best model with feature extraction was chosen. They compared the findings of each model to choose the best model that might be employed. NB provides training sets of documents as attributes $(x_1, x_2, x_{3,\dots,} x_n)$ and the document class $(y_{1,} y_{2,} y_{3,} \dots, y_m)$, where $x_i \in$ input features and feature become (x_n, y_m). The nave bayes technique was explained in Alg 1.

Alg 1: NB approach

```
Input
 T .= {.(x_i, y_i)| x_i ∈ n,. y_i ∈ m ,  i ∈ .{1,2,…,N}},
                  .N training samples.
 Z .= {.z_i | z_i ∈ m,i ∈ .{1,2,…,t}},..t test samples.
Initialization
    Y ← ∅
    .Read T input
    Calculate the parameter for predict class
Computation
    .for z_i ∈ Z. do.
    p_c ←calculate class document by equation 5.
    P_x ← calculate likelihood depend on (CNB, GNB, MNB, and BNB).
    y ← predicted label.
    Y ← Y. ∪ {y}}
Output
Y. = {y_i| y_i ∈m ,i ∈ {1,2,…,l}}, predict test in Z
```

6The Results

The researcher used the Python programming language, which provided pre-processing, visualization, and validation findings for text (data mining operations). Machine learning methods were used in this investigation. The researcher employed several NB parameters, as stated in Table 4, to determine accuracy, and then used the confusion matrix to do so. The researchers used the Multinomial, Bernoulli, Complement, and Gaussian tests, as well as TF-IDF for feature extraction.

TABLE 4: PARAMETERS USED IN THE NB APPROACH

Type of-Parameters-	**Value_**	**Details_**
Alpha-_	-1-_	By Laplace smoothing parameter using
Fit prior-_	-True-_	Whether or not to learn prior probability for a class. _

Class prior-_	-None-_	The classes' prior probabilities _
Binarize-_	-0-_-	Bernoulli's threshold for binarization of sample characteristics.
Norm-_	-False-_	Complement is used to perform the second normalization of the weights.
Var smoothing-_	-1e-9-_	In practice, the Gaussian model reveals that if the ratio of data variance between words is too tiny, numerical error will result. To solve this issue, we increase the variance artificially.

The dataset utilized in this paper is summarized in Table 5. The training and testing of instances from the dataset, as well as the class name, are displayed. We divided our dataset into two parts: 70% for training and 30% for testing.

TABLE 5: NUMBER OF TRAINING AND TESTING FOR EACH CLASS

-Class-_	-Training number-_	-Testing number-_
Business_	368_	142_
Entertainment_	274_	112_
Politics_	276_	141_
Sport_	359_	152_
Tech_	280_	121_
Total_	1557_	668_

Experiments on five groups or categories are done to assess the efficacy of the recommended technique. To choose main characteristic subsets, the NB algorithm is used. Table 6 shows how many of each class are included in the vector. There were 1557 documents for training and 21162 features from document training in the vector space.

The accuracy was determined using the confusion matrix. For several NB approaches, the TF-IDF findings were noted. The outcomes of the experiments are encouraging results are promising, demonstrating that the proposed technique outperforms others. Table 6 summarizes the findings of the evaluation.

TABLE 6: NB APPROACH MEASUREMENT FOR EACH CLASS

NB approaches	P_	R_	F1	-correct predict_	Acc	Label_
	90	98	94	139	97.887	Business
	100	88	93	98	87.5	Entertainment
	95	91	93	129	91.489	Politics

Multinomial	99	100	99	152	100	Sport
	94	98	96	118	97.520	Tech
Bernoulli	81	99	89	140	98.591	Business
	94	87	90	97	86.607	Entertainment
	95	89	92	126	89.361	Politics
	99	100	99	152	100	Sport
	97	84	90	102	84.297	Tech
Complement	97	96	96	136	95.774	Business
	100	95	97	106	94.642	Entertainment
	94	97	96	137	97.163	Politics
	99	100	99	152	100	Sport
	98	100	99	121	100	Tech
Gaussian	88	89	88	126	88.732	Business
	91	88	89	98	87.5	Entertainment
	90	89	90	126	89.361	Politics
	99	97	98	148	97.368	Sport
	85	89	87	108	89.256	Tech

In comparison to other models, complement was also more accurate. The number of correct prediction values for all NB models in TF-IDF is shown in Table 7. The training and testing curves are shown in Figure 3. The complement model was found to be the best because the training and testing curves revealed some space. When TF-IDF was applied to techniques of machine learning, the findings revealed a comparative change, with complement approach improving from 100 percent to 100 percent in the classes sprot and instruct, respectively as Table 7 demonstrates this. According to the results in Table 7, the Complement NB classifier had the best average accuracy, whereas the Bernoulli NB classifier produced the worst results.

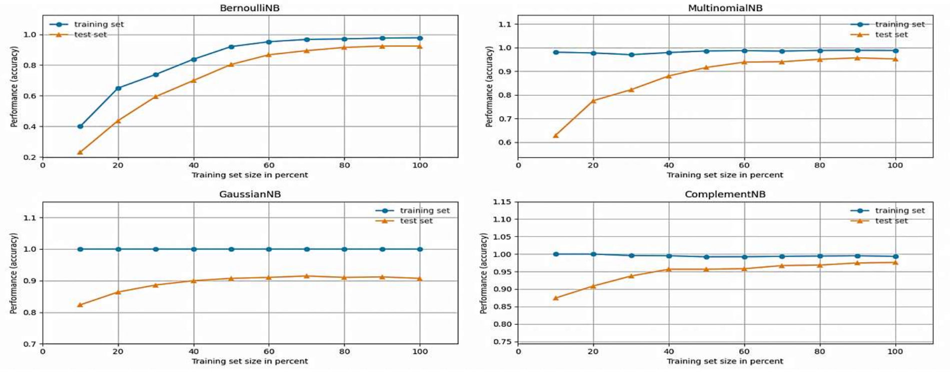


FIGURE 2: OUR APPROACH'S ACCURACY HAS BEEN TRAINED AND TESTED

It has been demonstrated in earlier trials and studies that when high-value training begins, the results are the best. Also, the testing starts with a high value and ends with the same value, which is the best-case scenario. Except for Bernoulli, which had a rough start and end for training and testing, we found the complement algorithms to be in great condition during training. Figure 4 depicted our approach's pression and recall.

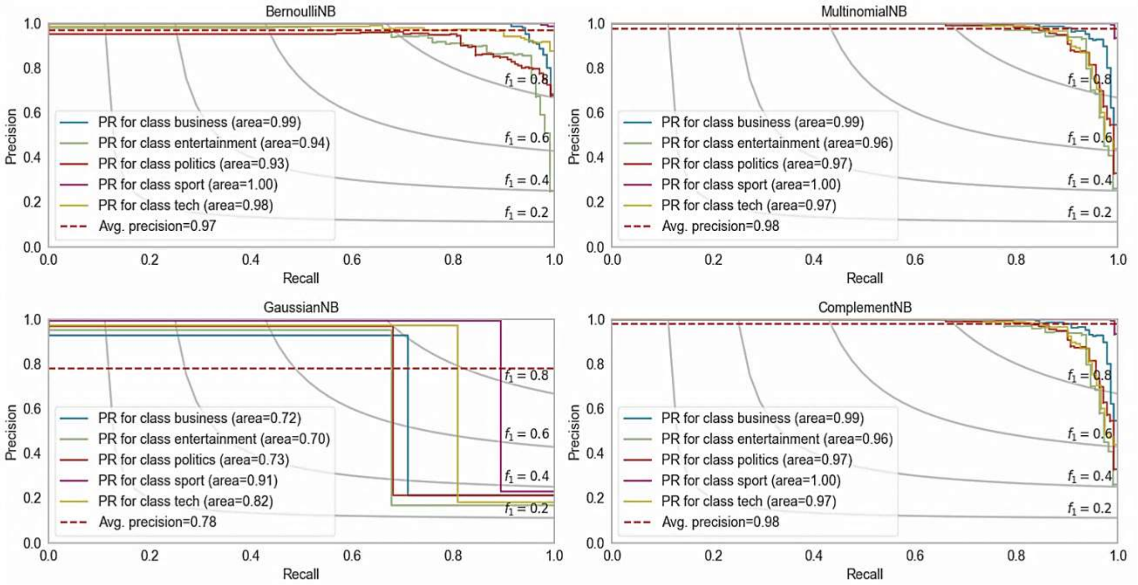


FIGURE 3: PRESSION-RECALL FOR NB APPROACHES

The F-scores for the five classes in the BBC news datasets are shown in Figure 4. Complement Naive Bayes is employed unigram in this experimental effort to attain the maximum accuracy compared to other NB techniques. As can be seen, the complement NB technique reported the best F-score line, with a precision of 1.00. The multinomial NB approach was also shown to be extremely close to 1.00. In the complement NB approach, all of the classes scored higher than f1 =0.8, indicating that they had a high F-sore of more than 0.95. The ROC carve of our NB methods is shown in Figure 5.

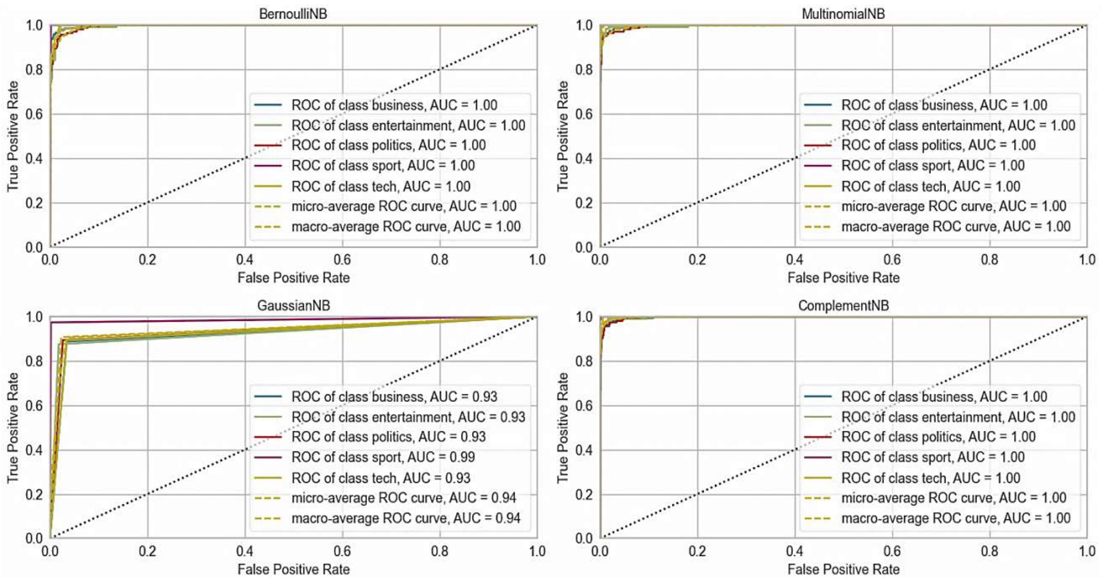


FIGURE 4: ROC CARVE FOR OUR NB APPROACHES

Figure 5 illustrates the best curve for any of the five classes in the BBC news datasets using ROC and AUC curves. We have two evaluation methods here (macro and micro). For both macro and micro, complement, Bernoulli, and multinomial attained AUC 1.00. For both macro and micro, Gaussian NB achieved an AUC of 0.94. The ROC obtained 1.00 for complement, Bernoulli, and multinomial techniques, according to the experimental data for the five classes. Table 7 shows that the supplement NB had a good accuracy of 97.604 percent.

TABLE 7: THE ACCURACY OF OUR PROPOSED APPROACH

Class	**Accuracy %**
Bernoulli	92.365
Multinomial	95.209
Gaussian	90.718
Complement	**97.604**

The supplement NB was shown to have the best effects. The TF-IDF approach was also employed to extract the data, with the complement NB displaying the highest accuracy numbers. It should be noted that the Gaussian NB technique, regardless of the method used, yielded a low value, however the multinomial NB method yielded a higher value than Bernoulli and gaussian. The supplement NB produced the best outcomes, according to the results of the experiments.

6.1 Discussion

The first experiment was carried out in this section to compare the proposed method to various NB methods. The precision, recall, and f-score of this experiment when employing different NB are shown in Tables 7 and 8. Figure 3 shows that our approach has a higher accuracy than the complement approach. In comparison to Gaussian, Multinomial, and Bernoulli distributions. The accuracy of the complement technique employing the unigram function was substantially greater than other approaches, as demonstrated in Tables 8 and 9. We evaluated the suggested model to existing nave bayes techniques using the same BBC news dataset to ensure its accuracy. In all approaches, the value for complement approach for sport and tech classes is higher than for other classes, as shown in Table 7. In general, class sprot in complement, Bernoulli, and multinomial obtained 100 percent accuracy, with the exception of gaussian, which achieved 97.368 percent. Through practical experience, it has been discovered that the proposed model achieves the best result (**97.604%**), which is superior to the results produced using alternative methods. The type or size of the feature's extraction approach has no effect on this.

7 Conclusions

The first purpose of this research was to discover the best model representation of BBC news for their category. The second goal was to see how well-known machine learning algorithms such as Multinomial, Bernoulli, Gaussian, and Complement affected the accuracy of four Nave Bayes approaches. The performance of the classifiers used for the classification problem is greatly improved in this paper. Four NB approaches were used to analyse the BBC news texts. They gathered 2225 documents for text analysis from the BBC news. They used the base paper approach with Nave Bayes classifiers to perform the text analysis. Text analysis was done in four steps: (i) pre-processing, (2) feature extraction, (3) split data, and (4) classification. For data classification, the researchers used four Nave Bayes algorithms. The maximum accuracy of the complement Nave Bayes classifier was 97.604%. We intend to continue this research by comparing the NB approach used in this research to some strategies of feature selection such as principal component analysis (PCA), Information Gain, and Relief in future studies. We also plan to look into additional issues linked to opinion mining, such as denial, sarcasm, and irony.

Reference

[1] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences, vol. 181, no. 6, pp. 1138-1152, 2011.

[2] B. Liu, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1-167, 2012.

[3] C. Quan and F. Ren, "Unsupervised product feature extraction for feature-oriented opinion determination," Information Sciences, vol. 272, pp. 16-28, 2014.

[4] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Stream-based active learning for sentiment analysis in the financial domain," Information sciences, vol. 285, pp. 181-203, 2014.

[5] C. Catal, U. Sevim, and B. Diri, "Practical development of an Eclipse-based software fault prediction tool using Naive Bayes algorithm," Expert Systems with Applications, vol. 38, no. 3, pp. 2347-2353, 2011.

[6] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Classifying political arabic articles using support vector machine with different feature extraction," in International Conference on Applied Computing to Support Industry: Innovation and Technology, 2019, pp. 79-94: Springer.

[7] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Political articles categorization based on different naïve bayes models," in International Conference on Applied Computing to Support Industry: Innovation and Technology, 2019, pp. 286-301: Springer.

[8] D. Zhang, "Bayesian Classification," in Fundamentals of Image Data Mining: Springer, 2019, pp. 161-178.

[9] S. Raschka, "Naive bayes and text classification i-introduction and theory," arXiv preprint arXiv:1410.5329, 2014.

[10] S. Geetha and R. Maniyosai, "An Improved Naive Bayes Classifier on Imbalanced Attributes," International Journal of Organizational and Collective Intelligence (IJOCI), vol. 9, no. 2, pp. 1-15, 2019.

[11] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," Journal of Information Science, vol. 44, no. 1, pp. 48-59, 2018.

[12] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in AAAI-98 workshop on learning for text categorization, 1998, vol. 752, no. 1, pp. 41-48: Citeseer.

[13] N. Sharma and M. Singh, "Modifying Naive Bayes classifier for multinomial text classification," in 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2016, pp. 1-7: IEEE.

[14] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 616-623.

[15] A. Anagaw and Y.-L. Chang, "A new complement naïve Bayesian approach for biomedical data classification," Journal of Ambient Intelligence and Humanized Computing, pp. 1-9, 2018.

[16] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," Expert Systems with Applications, vol. 36, no. 3, pp. 5432-5435, 2009.

[17] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive Bayes for text categorization," IEEE transactions on knowledge and data engineering, vol. 28, no. 9, pp. 2508-2521, 2016.

[18] H. Shimodaira, "Text classification using naive Bayes," Learning and Data Note, vol. 7, pp. 1-9, 2014.

[19] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 377-384.

[20] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in Mining text data: Springer, 2012, pp. 163-222.

[21] S. Alowaidi, M. Saleh, and O. Abulnaja, "Semantic sentiment analysis of arabic texts," International Journal of Advanced Computer Science and Applications, vol. 8, no. 2, pp. 256-262, 2017.