Machine Learning-based Opinion Extraction Approach from Movie Reviews for Sentiment Analysis

Mustafa Abdalrassual Jassim (mustafa944@mu.edu.iq)
MARS Research Laboratory, University of Sousse, Tunisia. Monastir Faculty of Science,University of Monastir, Tunisia

Dhafar Hamed Abd
Al-Maaref University College, Alanbar

Mohamed Nazih Omri
MARS Research Laboratory, University of Sousse, Tunisia. Monastir Faculty of Science,University of Monastir, Tunisia

Research Article

Keywords: Opinion mining, Sentiment analysis, Machine learning, movie review, word selection, Naive Bayes, Bag of words

Posted Date: July 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1780497/v1

Machine Learning-based Opinion Extraction Approach from Movie Reviews for Sentiment Analysis

Mustafa Abdalrassual Jassim · Dhafar Hamed Abd · Mohamed Nazih Omri

Received: date / Accepted: date

Abstract The field of sentiment mining, also known as sentiment mining, sentiment analysis, sentiment mining, sentiment extraction, etc., has seen a surge in the university environment. Various ways to automate the process of sentiment analysis have been tested by researchers in machine learning, data mining, natural language processing, and other fields. They examine the feelings embedded in people's opinions and beliefs that affect multiple areas, including companies' services and products. A movie review has to go through many processes to be able to detect and name feelings and achieve greater accuracy. Due to the structure of the language, the difficulties have been increased, its grammar and dictionary management. As part of this work, we propose a new approach for extracting words from a specific text and then classifying them. Thanks to a phase of pre-processing and extraction of words depending on the frequency, our approach makes it possible to select between 500 and 20,000 words with a vectorial construction by applying the techniques of "Term Frequency" (TF) and that of the "Term frequency-Inverse Document Frequency" ($TF - IDF$). Four different Naive Bayes models were thus considered and used (Complement, Multinomial, Bernoulli and Gaussian). We

*Corresponding author
Mustafa Abdalrassual Jassim*
MARS Research Laboratory, University of Sousse, Tunisia.
Monastir Faculty of Science,University of Monastir, Tunisia
E-mail: mustafa944@mu.edu.iq
ORCID iD: https://orcid.org/0000-0003-3360-6722

Dhafar Hamed Abd
Department of Computer Science Al-Maaref University College, Alanbar, Iraq.
E-mail: Dhafar.hamed@uoa.edu.iq
ORCID iD: https://orcid.org/0000-0003-0548-0616

Mohamed Nazih Omri
MARS Research Laboratory, University of Sousse, Tunisia.
E-mail: mohamednazih.omri@eniso.u-sousse.tn
ORCID iD: https://orcid.org/0000-0001-7803-0179

evaluated our proposed approach against different standard measures namely *precision*, *accuracy*, *recall*, $F1 - score$ and *kappa*. The results revealed that the Naïve Bayes multinomial system obtained the best results with an accuracy of 86.46%.

Keywords : Opinion mining; Sentiment analysis; Machine learning; movie review; word selection; Naive Bayes; Bag of words.

1 Introduction

1.1 Context and issues

The method of text's mining information to identify and realize the attitude that the writer is intending to communicate via words is known as Sentiment Analysis (SA)[1]. Most of the time, sentiment extractors classify between versions with a positive sentiment and those with negative sentiment (or occasionally neutral)[2][3]. A test determines if the web-based survey (of a movie, product, or book) contains a (positive or negative) rating among the most common sentiment functions. More in-depth SA may discern between multiple sentiments included inside a single text. When reading a restaurant review, a well-trained SA suite would deduce that the individual critic had a positive opinion about the cuisine but had a service poor opinion. The amount of information that are subjective are available on the Internet has increased significantly during the previous two decades[4][5]. The numbers of websites have sprung up because the users can create and collaborate whatever they wish [6]. Particularly valuable as information exporters are social networks [7], which allow users to express their opinions, thoughts, and feelings by sharing them with others. It will take little time to complete organizing the data and predict any other information. A critical component of SA is slang, misspellings, short formulae, repetitive characters, and modern dialects and symbols[8], making it more challenging to perform. It is not easy to define emotional words and phrases; one can use the exact words in different contexts[2]. Use SA to enhance the strategies of sales, business, services, and marketing[9][10].

The Naive Bayes (NB) workbook is one of the most widely used workbooks. On the base of the word distribution in the document, the model of NB classification produces the posterior probability[11] . The goal is to elicit appropriate and best opinions via keywords restricting or traits and reducing false/misleading ideas [12]. For this reason, some authors have relied on a proposed approach based on two algorithms, the Bernoulli and Multinomial NB Classifiers, to predict whether feelings are positive or negative. Each of these algorithms is used to classify documents but differs widely in their classification approaches. NB polynomial classifier depends on the term frequency concept, this means the times number of word appearing in the document. From another side, Bernoulli NB Classifier depends on the concept of binary which means if the term appears in a document or not. Unlike Multinomial Naïve Bayes, it doesn't mention the frequency word. The results for Multinomial NB were

73.4%, while for Bernoulli NB Classifier it was 69.15%[13]. Others determined the opinion in the text by extracting the feature set based on IDF (Inverse Document Frequency) with the NB classifier, and the classification accuracy was 85.25% [11]. He found that the polynomial model works best among four probability models, including the multivariate Bernoulli model. Both studies point out that the polynomial model takes the frequency of word appearance into account as significant differentiating factor for the two models.

In this work, the algorithm of supervised machine learning, a NB classifier implemented in the dataset, will be used. These data included preprocessing (tokenization, lower case of alphabetic, removing unwanted words, removing less than three alphabetic and stop word removals) and frequency-dependent word extraction in this study. TF was used in the first test for with 4 NBs and TF-IDF for the second. The researchers applied TF-IDF and TF and derived from 500 to 20,000 features. Next, the vector space contained 50,000 documents.

1.2 Motivation and contributions

In our study, among the most important was determining the better method for selecting texts features in order to be analyzed in the applications of opinion mining, for the largest movie review dataset. A model was built that extracts and categorizes words from a given text by reprocessing and frequency-dependent word extraction in this study, which uses a NB classifier for making decision. We were focused on the performance basing on time efficiency and accuracy, both of which are critical in real-world applications. From an opinion-mining perspective, we first ran a number of tests on the input texts features that we thought were important in text classification; Four NB models were used complement, multinomial, Bernoulli, and Gaussian). We deal with many data posted by people worldwide expressing their opinions on various topics. Furthermore, the development of the methods of opinion-mining is as fast and efficient as possible. For the current issue, NB classifier was used. After putting it through a series of tests to see how well it performed on different feature sets taken from the analyzed texts, we found that extracting an appropriate amount of texts words used for classifier training which is the best option for picking the real-time applications feature set. It performs admirably and quickly for a technique of this type. Finally, for testing purposes, a database of positive and negative movie reviews was employed to train and test the classifiers.

1.3 Paper organisation

The rest of this manuscript is organized as follow. Section 2 presents an overview of the related work of previous works on the domain of Social Analysis; Section 3 presents an overview on the Naive Bayes classification algorithm.

In this section, we present the classification algorithm we have used and its principle of operation. Section 4 presents the proposed approach in which we give the general architecture of the proposed model, the Pre-treatment phase, the select frequents words process, and the proposed algorithm we developped. The section 5, meanwhile, is reserved to includes the experimental study and the results analysis. Finally, section 6 concludes this work and gives the main prospects.

2 Literature review

SA is separated into four types: entity, phrase, document, and sentence levels. For classification use [8][14] the NB method, for feature extraction use TF-IDF, and for feature selection use Information Gain in conjunction with the NB approach. However, the many ways Naïve Bayes has been used for classifying the word conditionally independent of one another. To prevent this problem, uses Information Gain and TF-IDF; the result was accuracy is 88.27. To [15] investigate the relationship between online movie reviews, it was discussed how machine learning methods and SA; the paper illustrated how the model of sentiment-aware auto regression simplified version could predict movie ratings on online review data with high accuracy. To evaluate positive and negative attitudes in documents, document-level SA is used, including the use of values of the Term Frequency (TF) and Inverse Document Frequency (IDF) as features and Fuzzy Clustering to determine negative and positive attitudes. Processing and extracting textual data from documents is critical. As a result, one may gauge the quality of a movie based on other people's thoughts or reviews. Positive and negative reviews are grouped in this category. Where [16] tested two datasets for movie review (Cornell and Stanford datasets) on several algorithms, namely SVM, SVM+IG, NB, and KNN, NB accuracy has reached Cornell's 2,000 data 80.75 and the Stanford data for 25,000 data is 81.28. Others used[17] machine learning to rank polarity on movie review data for five machine learning classifiers types to analyze this data. Hence, the classifiers studied are Multinomial NB (MNB), Maximum Entropy (ME), Support Vector Machine (SVM), Decision Tree (DE), as well as Bernoulli NB (BNB). The highest results were among Multinomial NB and Bernoulli NB, reaching 88.50 and 87.50, respectively. Some authors[18] have captured the dataset from IMDb movie reviews. There are 50,000 instances contained in this dataset, with two columns with their sentiments being reviews. To extract valuable data, the dataset undergoes a cleaning process. Two features' types being investigated, namely, TF-IDF modelling with NB classifiers and a bag of words (BoW). The percentage of 89% is the accuracy score. Table 1 shows a comparison between the classifications of NB with our proposal.

Table 1: Comparison between the classifications of Naive Bayes with our proposal.

Authors	Year	Problem OR Objective	Methods and Description	Data Set	Accuracy %	No of items/Samples
Noor Adam et al [8].	2021	Designing and implementing a web-application system using the SA results from the reviews of IMDB movie.	NB (Two features' types were investigated: the TF-IDF modeling with classifiers of Naïve Bayes and the Bag of Words (BoW)).	IMDB Dataset of 50K Movie Reviews Kaggle	89	100 words
Meta Mahyarani et al [15].	2021	The ways in which Naïve Bayes is used for classifying the word are conditionally independent of one another. This study employs Information Gain and TF-IDF to address this issue.	NB (For classification Uses the NB approach, for extracting feature uses TF-IDF, and for selecting the feature uses Information Gain).	IMDB https://datasets. imdbws.com/	88.27	7842, 8665, and 9779
Reza Maulana et al[17].	2020	It is possible to solve issues faster and more reliably using the method of Information Gain as a feature selection technique.	(Dataset Cornell) NB (Dataset Stanford) NB	Two datasets of movie review: -Stanford and Cornell Datasets (cornell.edu) -The Experiment of Stanford Prison (2015) - IMDb	80.75 81.28	2,000 data of Cornell dataset And 25,000 data Stanford dataset
Atiqur Rahman et al[18].	2019	Use an ML method for classifying movie review data into negative and positive polarity.	Multinomial NB Bernoulli NB	movie reviews Sentiment-Analysis-on-Movie-Review-Data/Data set.xlsx at master · riyadatik/Sentiment-Analysis-on-Movie-Review-Data · GitHub	88.50 87.50	1000 positive and1000 negative
Yanuar Nurdiansyah et al[19].	2018	The construction of a system that can classify the feelings expressed in review documents into two classes: negative and positive sentiments.	NB Classifier	Was selected Movienthusiast, a movie review on the Bahasa Indonesia website. movienthusiast's films Letterboxd	88.37	collected 1201 movie reviews: negative reviews were 418 and positive reviews was 783

Table 1: Continued from previous page

Authors	Year	Problem OR Objective	Methods and Description	Data Set	Accuracy %	No of items/Samples
Mais Yasen et al [20].	2019	Tokenization is used to convert the inputted string into a word vector, stemming is used to get the words' root, feature selection is used for the most important terms extraction, and classification is used to categorize reviews as negative or positive.	NB (AUC, recall, Accuracy, f-measure and precision, utilized for results evaluation).	IMDB movies reviews Retrieved on: November 1, 2018, from: Kaggle: Your Data Science Community and Machine Learning	81.83	42926 review (positive or negative)
Nimesh V Patel et al [1].	2019	Reviews are classified into binary categories as either positive or negative opinions on reviews in different parts of the data set, subjected to sentiment analysis and rating.	Naïve Bayes (N-gram features were used).	movie reviews Bo Pang's Home Page (cornell.edu)	0.8953	1000 positive and1000 negative reviews
Kamoltep Moolthaisong et al [21]	2020	Data mining was utilized for movie reviews classification, also use a method for word clouds building from the word frequency in movie reviews.	NB (The review text then word-stemming got processed, stop words removing. Using TF-IDF for features selection).	From Metacritic website extracted Movie review data. Metacritic - TV, Movie, Game and music Reviews	80.25	multiple movies produce 700 reviews

The literature mentioned above comprises several approaches towards the opinion mining of large movie reviews; these works are limited in several aspects that mainly include:

- (i) Did not use frequency word selection as feature extraction with a different number of features.
- (ii) Not all work on large movie review datasets.
- (iii) Using NB with different models to compare them and select the best one.

Instead, the proposed study presents an intelligent opinion mining of a large move review dataset into positive and negative while utilising NB with varying configurations and a larger and diverse dataset.

3 An overview on the Naïve Bayes Classification Algorithm

The NB method is simple and popular in opinion mining, it also has a text mining's useful application [21]. The following mathematical equation describes how works NB and explains its mechanism:

$$y = P(l_j) \prod_{i=1}^{n} (P(x_i \mid l_j) \tag{1}$$

Where features in movie review refered to by $x1$ to x_n, and l denotes the document class. The final noted class for the test is y and compared to the data's test label. For determining the minimum or maximum y based on using the NB Equation (1). For calculating $P(l_j)$ and $P(x_i \mid l_j)$ values the following two equations are used, note that $P(l_j)$ was same for all types of NB, where in the value basing on the used classifier, $P(x_i \mid l_j)$ was differed among the used methods.

$$P(l_j) = \frac{documentP(l = l_j)}{N_{doc}} \tag{2}$$

$$P(x_i|l_j) = \frac{count(x_i|l_j)}{\sum_{x \epsilon v} count(x|l_j)} \tag{3}$$

Next sections describe the four used types of NB in the current study.

3.1 Gaussian Naive Bayes

The numeric attributes' values were regularly distributed in the Gaussian NB classifier [22]. This distribution was represented in terms of standard deviation and mean, which will aid in calculating the probabilities of observed values using estimations. Equation (4) was used to calculate the probability of the features:

$$P(x_i \mid l) = \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left(-\frac{(x_i - \mu_l)^2}{2\sigma_l^2}\right) \tag{4}$$

Where σ = standard deviation and μ = mean. y was calculated using equation (1). The attribute values are discretized using the technique of binning for continuous data.

3.2 Multinomial Naive Bayes

Equation (5) was used to capture the frequency of the word for the documents' presented information by the multinomial NB classifier [23]. Because the multinomial distribution requires integer features, it was well suitable for discrete feature classification.

$$\begin{aligned} P(x_i \mid l_j) &= \frac{n!}{\prod_{i=1}^{|V|} x_i!} \prod_{i=1}^{|V|} P(x_i \mid l_j)^{x_i} \\ P(x_i \mid l_j) &= \prod_{i=1}^{|V|} P(x_i \mid l_j)^{x_i} \end{aligned} \tag{5}$$

Where number of features denoted by V. In this model, the $\frac{n!}{\prod_{i=1}^{|V|} x_i!}$ normalization was independent class k. Hence, Laplace equation was used to calculate $P(x_i \mid l_j)$ after the addition of V parameter in order to prevent the *zero* frequency as demonstrated below:

$$P(x_i|l_j) = \frac{count(x_i|l_j) + 1}{\sum_{x \epsilon v} count(x|l_j) + |V|} \tag{6}$$

Where V = all classes vocabulary in the training dataset.

3.3 Complement Naïve Bayes

The technique of Complement Normal NB, have probability of its occurrence in other classes instead of the calculation of the probability of the word occurrence in the class [24]. Thus, other classes' word-class dependencies estimation $P(x_i|l'_j)$. y 's minimum value and l'_j for the selected reverse class was determined in equation (7).

$$y = P(l_j) \prod_{i=1}^{n} \frac{1}{P(x_i|(l')_j)} \tag{7}$$

3.4 Bernoulli Naive Bayes

The classifier of Bernoulli NB suggested binary features where it required just two values [25]. The following equation describes Bernoulli distribution:

$$P(x) = P^x(1-P)^{1-x} \tag{8}$$

Where Bernoulli distribution referred to by x with a value range between [0-1]. If it was 1, success occurred, while it was failed if it was 0 basing on the equation (9).

$$\begin{aligned} P(x=1) &= P^1(1-P)^{1-1} = p \\ P(x=0) &= P^0(1-P)^{1-0} = (1-p) \end{aligned} \tag{9}$$

The probability of non-occurring word in the class document was $(1-p(x_i \mid l))$, where word in the document referred to it by x. Thus, the equation (10) was:

$$P(x_i \mid l) = P(x_i \mid l)b_i + (1-b_i)(1-p(x_i \mid l)) \tag{10}$$

All the words can use this product. If the document's word x_i presented, then $bi = 1$ and the probability was $P(x_i|l)$. If x_i absent, then $bi = 0$ and the likelihood was$(1-p(x_i|l))$. So, equation (1) utilized for y determination.

4 Proposed Opinion Extraction Approach

By using the proposed approach, words can be extracted from the given text then it classified. The proposed approach procedure is as follows:

- Preprocessing: This step consists of (tokenization, lower case of alphabetic, remove unwanted words, remove less than three alphabetic and stop word removal).
- Extract words depend on the frequency in this study select from 500 to 20000 words.
- Build the vector by applying TF and TF-IDF.
- Used four NB model which are complement, multinomial, Bernoulli and Gaussian.

Evaluation proposed approach using accuracy, recall, precision, f-score, and kappa metric. Figure 1 showed the Steps for model classification.

4.1 Dataset description

Raw data were gathered from Kaggle [1] website . Large movie review dataset (IMDB) comes with two labels which are positive and negative in total of 50K reviews where positive label 25K reviews and negative 25K reviews. This dataset includes 438729 words with their binary ratings, and it represents a collection of movie reviews (negative or positive).

[1] www.kaggle.com

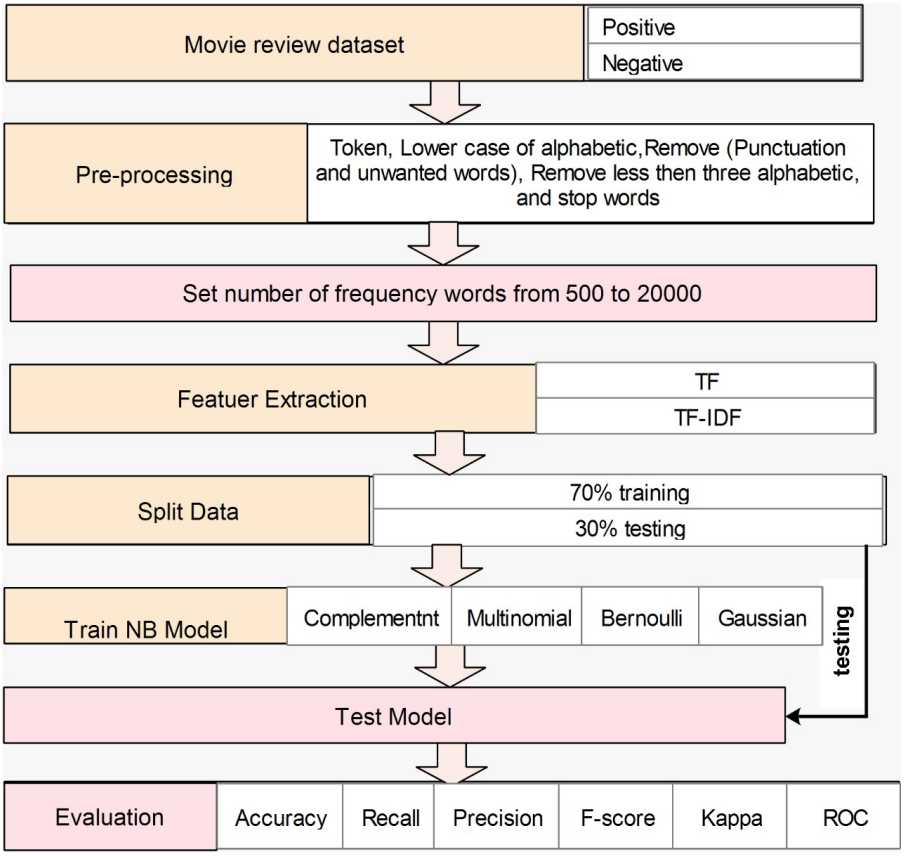

Fig. 1: General architecture of the proposed approach

4.2 Pre-treatment phase

This pre-processing phase is based on four sub-phases namely Tokenization, Eliminate unwanted characters, Normalization, Stop-words removal and the Remove less frequency word sub-phase described as follows:

- **Tokenization:** The document text is split into a sequence of tokens in this step. Typically, data acquired from online evaluations is coupled with noises such as scripts, advertising, URLs, HTML tags, and symbols like asterisks, hashes, and other symbols that are neither useful nor relevant in classification. To increase the effectiveness of the classifier, these symbols and noise must be removed, leaving only the useful words.
- **Eliminate unwanted characters:** This method removes unwanted characters or strings from the text, such as punctuation marks, hashtags, English numerals, non-English characters, and so on. To complete this work, a variety of regular expressions must be used, as shown in Table 2.

Table 2: Eliminate unwanted characters

Expression	Output
[# - _ ? , . ' ;]+	Removed all punctuation marks
[0-9]+	Removing the English numbers

- **Normalization:** All document's characters are transformed to lowercase or uppercase in this step. The majority of the evaluations employ a mix of lowercase and capital characters. The entire document set is converted to lowercase throughout this process.
- **Stop-words removal:** it enables removing English stop-words from a review by comparing the word to the built-in stop-words list and deleting each token. Stop words are words that aren't important for the opinion or sentence.
- **Remove less frequency word:** After each pre-processing stage, the features number decreases. According to the results of the experiments, we cannot avoid the preprocessing and cleaning data tasks related to the English language in order to reduce classifier complexity, save time, and reduce storage requirements.

4.3 Select Frequents words

In our approach, we consider two feature extraction models namely TF recurrence rate in a movie review and $TF-IDF$ given that reviews vary in length. It is normal for a term to be repeated more frequently in longer reviews than in shorter reviews. In order to obtain a normalized value, this frequency of the term is generally divided by the total number of revisions. It is given by the following equation:

$$TF = \frac{n}{\sum T_n} \tag{11}$$

In where n denotes the number of times a term appears in the review, and Tn is the total number of terms in the review. The products of TF and IDF are computed using the TF-IDF technique (IDF). The IDF is referred to in Equation (12).

$$IDF = \log\left(\frac{N}{rf}\right) \tag{12}$$

In which N denotes the total number of reviews and rf is the reviews containing a specific term. TF-IDF is a complex system that can distinguish terms in both the review and the dataset. The formula for calculating TF-IDE is as follows:

$$TF - IDF = TF * IDF \tag{13}$$

To lower the vector size, this technique reduces the weight of features in reviews prior to feature selection. Following that, machine learning techniques

are applied to the features. Select high frequency phrases, on the other hand, are of significant significance despite their frequency. As a result, frequently used phrases must be weighed. The frequency word threshold is set between 500 and 2000, and the highest accuracy is determined.

4.4 Proposed Algorithm

The Bayes rule of probability theory is used to create a supervised learning method called Naïve Bayes classification. The classification is based on labelled training data and the strong assumption that all of the training data's attributes are independent of each other, resulting in labels. To determine orientation from the vast movie review dataset, a nave Bayes classifier is built. The usefulness of Naïve Bayes classifiers has been proved utilizing training cases with numerous attributes, and they have good performance and classification speed. The independence assumption is largely responsible for this great performance.

Algorithm 1 provides training sets of features reviews as $(x_1, x_2, x_3, \ldots, x_n)$ and the reviews class $(l_{positiveand}l_{negative})$ where $x_i\epsilon$ input features. The form of the feature will take the form (x_n, l_2). In this case, for all methods (Bernoulli, Complement, Multinomial, Gaussian) that use the $P(l_2)$ for each class as in equation (2), where $P(x_i \mid l_2)$ calculate each class probability depending on the using method. Algorithm 1 make methods comparison and chooses the best one.

Algorithm 1: Naïve Bayes-based proposed Algorithm

Input: Set of training and testing samples
$Tr = \{(x_i, l_2)|i = \{1, 2, 3, \ldots, n\}, j = \{positiveandnegative\}\}$,training set
$Z = \{z_i \mid i = \{1, 2, 3, \ldots, e\}\}$, test set
Output: Set of predicted class
$Y = \{y_i \mid i\epsilon\{1, 2, 3, \ldots, e\}\}$- the test samples in Z with the set of predicted class lables.
Begin
/*Initialization:*/
$Y \leftarrow \emptyset$;
Read the training Tr
Calculate the parameter for predict class
/*Computation:*/
for $z_{i\epsilon}$ *in* Z **do**
    Computation:
        (a) $P(l_j) \leftarrow$ calculate articles class by equation (2);
        (b) $P(x_i \mid l_j) \leftarrow$ calculate the likelihood for each class depend on model (Gaussian, Multinomial, Complement, Bernoulli);
        (c) $y \leftarrow$ the label predicted by applying equation (1) on z_i, according to (a) and (b);
        (d) $Y \leftarrow Y \cup \{y\}$;
end

5 Experimental study and results analysis

5.1 Simulation setup

The proposed solution was implemented by the Python programming language, which provided preprocessing, visualization, and validation results for text (data) mining procedures. Machine learning algorithms were used in this investigation. The simulations are carried out on Intel® CoreTM i3-8100 CPU @ 3.60 GHz, 8 GB RAM, Windows 10 (Professional).

5.2 Evaluation metrics

Let TP, FP, TN, and FN be true negative rates, false negative rates, true positive rates, and false positive rates, respectively. The confusion matrix (see Table 3) summarizes these measures.

Table 3: Confusion matrix

		Predictive result	
		Positive	Negative
Actual result	Positive	TP	FN
	Negative	FP	TN

The following standard measures are then designed to evaluate the effectiveness of our approach: $F1-score$, $Recall$, $Precision$, and $Accuracy$ are all terms that describe how well something works. The percentage of positive instances is called accuracy, and the $F1-score$ is the harmonic mean of $Recall$ scores, $Precision$, and the $Kappa$. It is a good scale that can deal very well with the problems of multiple and unbalanced classes. Furthermore, $Accuracy$ is calculated using the following equation:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{15}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{16}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \tag{17}$$

$$Kappa = \frac{p_o - p_e}{1 - p_o} \tag{18}$$

5.3 Dataset description

We employed the confusion matrix to determine accuracy after adjusting different parameters for NB, as given in Table. Two tests have been used: the first was for TF with four NBs, and the second was for $TF - IDF$. We used TF and $TF - IDF$ to generate 500 to 20000 features. Following that, the vector space grew to accommodate 50000 documents. As stated in Table 4, the confusion matrix is utilized to determine accuracy.

Table 4: Naïve Bayes-based Algorithm parameters

Parameters	Value	Details
Fit prior	True	To learn previous probabilities of class or not learn it.
binarize	0	Bernoulli's threshold for binarizing sample features
Var smoothing	1e-9	Gaussian model in practice finds that if the data variation between words is too little, a numerical inaccuracy will result. We increased the variance in an attempt to combat this issue.
alpha	1	Use Laplace to smooth the parameters
norm	False	Complement is used to perform the second normalization of the weights.
prior class	None	Probabilities that a certain class will exist.

Table 5 shows numbers of words were collected from preprocess step.

As shown in table above the total number of words 96430 but still big. In this case need big memory and many processes, so need to select the number of words that will be achieve high accuracy and less time. Table 6 show the number of split datasets for training 70% and testing 30%.

Table 7 shows the number of words selected with model accuracy using TF feature extraction, as well as a Summary chart 2 for TF feature extraction.

Table 5: words collected from the pre-processing step

Step	class	Number of words	Total number of words (unique)
Raw data	Positive	7009136	101895
	Negative	6965330	
Lower alphabet	Positive	7045362	101881
	Negative	7002579	
Punctuation	Positive	5966173	137520
	Negative	5897408	
Remove unwanted and space	Positive	5728998	96998
	Negative	5643280	
Remove length word less than three alphabetic	Positive	4441116	96485
	Negative	4356226	
Remove stop words	Positive	3079961	96430
	Negative	3010186	

Table 6: split dataset for training and testing

Class	No. training 70%	No. testing 30%
Positive	17636	7364
Negative	17364	7636
Total	35000	15000

Figure 2 shows the summarization of Table 7 where used TF as feature extraction with four models of NB. As can see multinomial the best model where win 11 times than other models.

Table 7: TF feature extraction

Sets of words	**Model accuracy (%)**				**Best Model**
	Multinomial	**Bernoulli**	**Complement**	**Gaussian**	
500	81.93	80.52	**81.97**	80.11	Complement
1000	**83.3**	82.85	83.29	81.05	Multinomial
1500	**84.36**	**84.35**	84.34	79.9	Multinomial
2000	84.4	**84.53**	84.39	78.44	Bernoulli
3000	84.57	**85.06**	84.55	76.51	Bernoulli
4000	84.62	**85.07**	84.64	76.21	Bernoulli
5000	84.63	**85.03**	84.65	74.77	Bernoulli
6000	84.75	**85.03**	84.74	74.31	Bernoulli
7000	84.98	**85.19**	84.97	73.69	Bernoulli
8000	84.99	**85.12**	84.99	73.04	Bernoulli
9000	85.09	**85.11**	85.09	72.71	Bernoulli
10000	**85.22**	85.13	85.21	72.76	Multinomial
11000	**85.36**	85.34	85.33	72.3	Multinomial
12000	**85.39**	85.17	85.37	71.77	Multinomial
13000	**85.43**	85.14	**85.43**	71.7	Multinomial & complement
14000	85.49	85.12	**85.51**	71.76	Complement
15000	**85.52**	85.09	85.51	71.5	Multinomial
16000	**85.57**	85.21	**85.57**	71.14	Multinomial & complement
17000	**85.65**	85.25	**85.65**	70.4	Multinomial & complement
18000	**85.63**	85.24	85.61	69.91	Multinomial
19000	**85.61**	85.19	85.59	69.39	Multinomial
20000	85.59	85.16	**85.62**	69.39	Complement
Total	11	8	6	0	**Multinomial**

While Figures 3, 4, 5, 6 show the highest accuracy among four models of NB and compares it with word count. The accuracy of TF Multinomial was 85.65 at 17 000 words, TF Bernoulli 85.34 at 11 000 words, TF Complement 85.65 at 17 000 words, and finally got a TF Gaussian accuracy of 81.05 at 1000 words. We note that the two models TF(Multinomial and Complement) have equal accuracy at 17 000 words.Where the blue color represents the number of words, while the red color represents the percentage of accuracy.

Table 8 shows the number of words where select with model accuracy using TF-IDF feature extraction, as well as a Summary chart 2 for TF feature extraction.

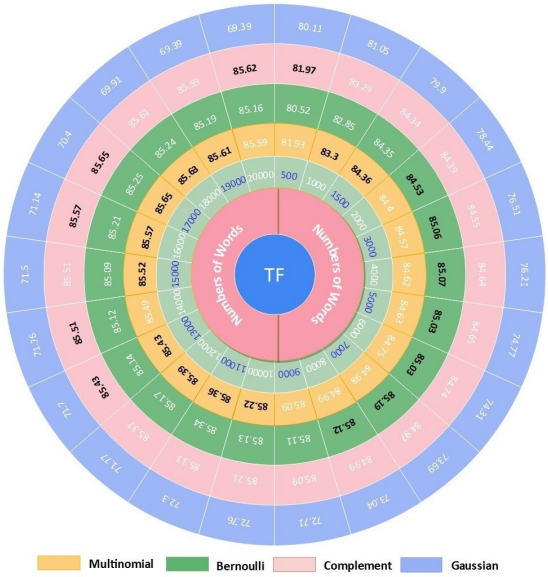

Fig. 2: Summary chart for TF feature extraction

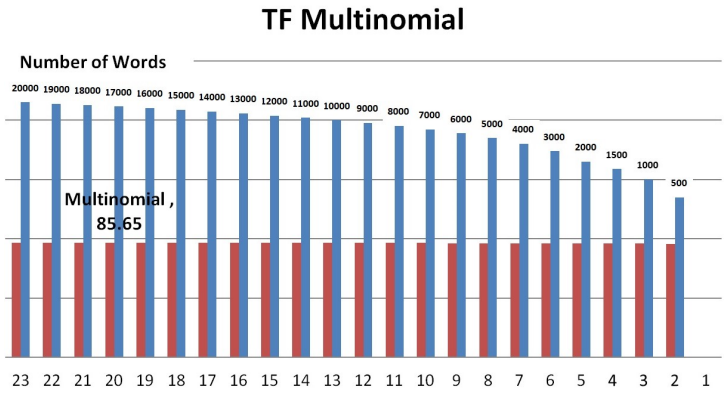

Fig. 3: Summary chart for TF Multinomial Accuracy Ratio

Figure 7 shows the best model used with TF-IDF feature extraction. As we can see multinomial model achieved high accuracy than others models. Multinomial model wins 12 times and complement 11 times, but Bernoulli and gaussian didn't achieved as shown in Table 8.

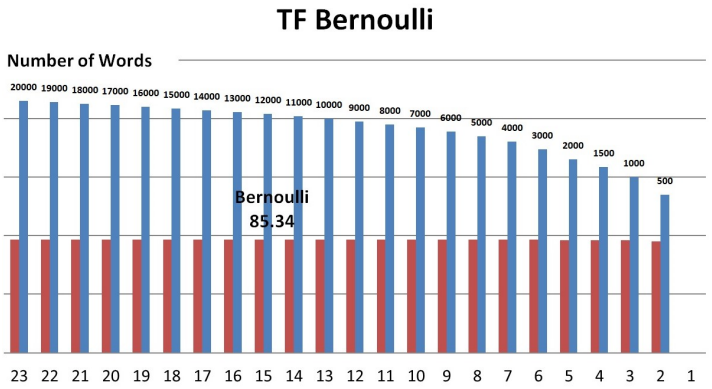

Fig. 4: Summary chart for TF Bernoulli Accuracy Ratio

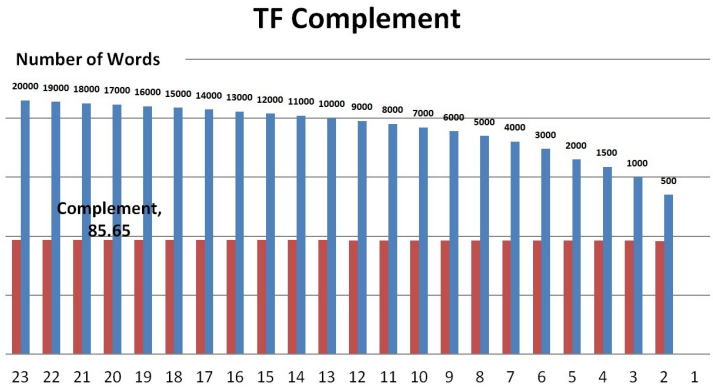

Fig. 5: Summary chart for TF Complement Accuracy Ratio

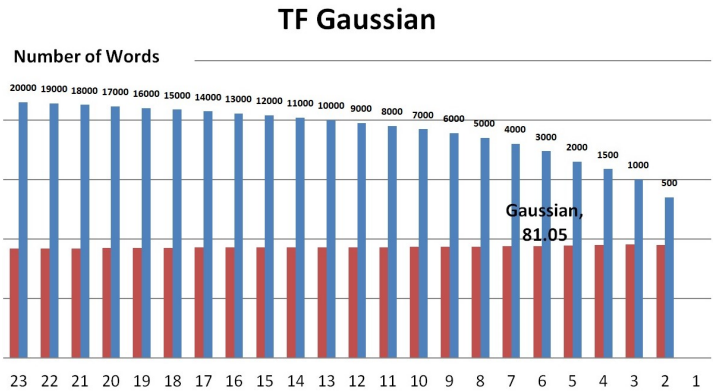

Fig. 6: Summary chart for TF Gaussian Accuracy Ratio

Table 8: TF-IDF feature extraction

Sets of words	**Model accuracy (%)**				**Best Model**
	Multinomial	**Bernoulli**	**Complement**	**Gaussian**	
500	82.39	80.52	**82.4**	80.49	Complement
1000	83.79	82.85	**83.87**	81.71	Complement
1500	84.91	84.35	**84.96**	81.74	Complement
2000	85.04	84.53	**85.09**	81.52	Complement
3000	**85.17**	85.06	85.06	80.79	Multinomial
4000	85.19	85.07	**85.22**	80.84	Complement
5000	**85.29**	85.03	85.24	80.33	Multinomial
6000	85.47	85.03	**85.49**	80.09	Complement
7000	85.47	85.19	**85.59**	79.35	Complement
8000	**85.59**	85.12	**85.59**	79.19	Multinomial & Complement
9000	**85.88**	85.11	85.85	78.79	Multinomial
10000	**85.9**	85.13	85.89	78.69	Multinomial
11000	**86.13**	85.34	86.1	78.24	Multinomial
12000	**86.12**	85.17	86.06	77.73	Multinomial
13000	86.24	85.14	**86.28**	77.54	Complement
14000	86.25	85.12	**86.27**	77.5	Complement
15000	**86.32**	85.09	86.24	76.79	Multinomial
16000	**86.29**	85.21	86.23	76.21	Multinomial
17000	86.26	85.25	**86.27**	75.47	Complement
18000	**86.46**	85.24	86.34	74.99	Multinomial
19000	**86.42**	85.19	86.37	74.39	Multinomial
20000	**86.45**	85.16	86.43	73.69	Multinomial
Total	**12**	0	11	0	**Multinomial**

While Figures 8, 9, 10, 11 shows the highest accuracy among four models of NB and compares it with word count. The accuracy of TF-IDF Multinomial was 86.46 at 18,000 words, $TF-IDF$ Bernoulli 85.34 at 11,000 words, $TF-IDF$ Complement 86.34 at 18,000 words, and finally got a$TF-IDF$Gaussian accuracy of 81.74 at 15,000 words. We note the superiority of the Multinomial model over all models to extract the advantage at 18000.

Table 9 shows three methods were taken for select the best feature extraction these methods are (multinomial, Bernoulli, complement). Words are a word that has been extracted and selected for use as a term in the TF-IDF. Initially, the word weighting assessment will be performed on the base of IDF and TF; This evaluation results will be incorporated into the classification stage's prediction model as a feature.

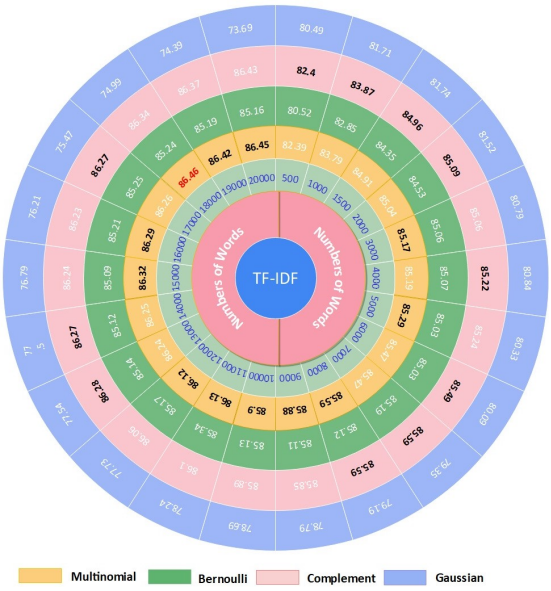

Fig. 7: Summary chart for TF-IDF feature extraction

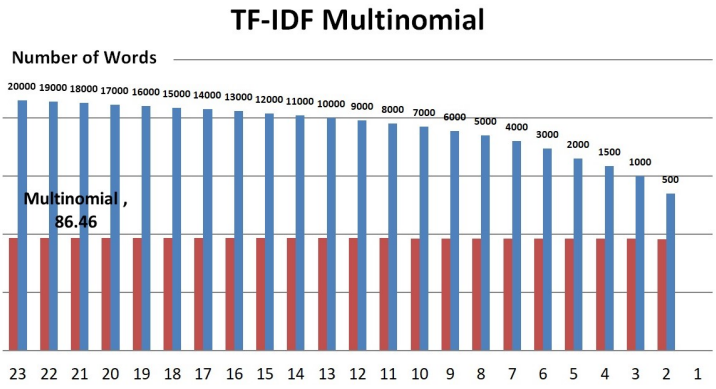

Fig. 8: Summary chart for TF Multinomial Accuracy Ratio

As we can see the best feature extraction with IMDB dataset is TF-IDF and for method both multinomial and complement. Belong to Table of TF-IDF we can achieved multinomial the best because achieved 12 best accuracies than complement. Table 10 shows the other metric and takes number of words 18000 because achieved higher accuracy.

Figure 12 shows the learning carve and figure 13 shows AUC and ROC curves.

The accuracy bar graph and the ROC curve show that Multinomial Naive Bayes performed well on the provided dataset. Because it has a 94 percent

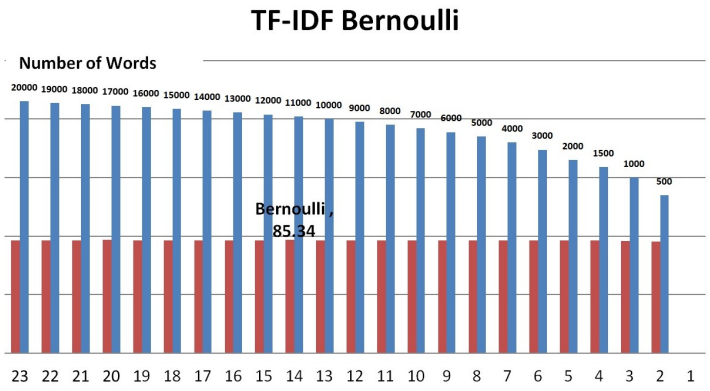

Fig. 9: Summary chart for TF Bernoulli Accuracy Ratio

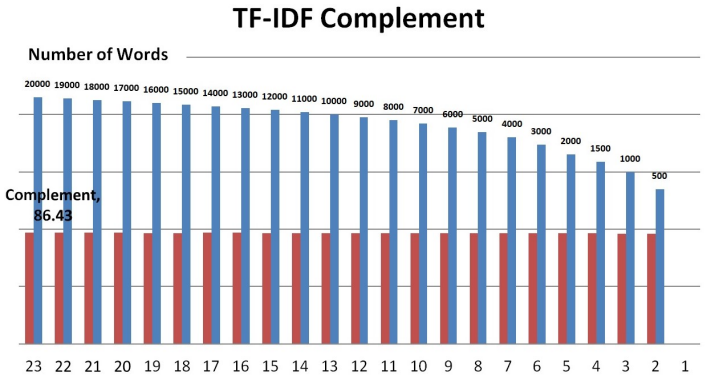

Fig. 10: Summary chart for TF Complement Accuracy Ratio

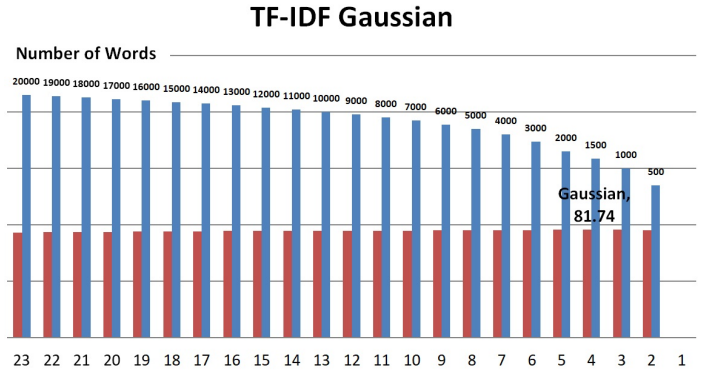

Fig. 11: Summary chart for TF Gaussian Accuracy Ratio

Table 9: Selected the best feature extraction

Sets of words	Multinomial		Bernoulli		Complement		Best feature
	TF	TF-IDF	TF	TF-IDF	TF	TF-IDF	
500	×	✓	-	-	×	✓	TF-IDF
1000	×	✓	-	-	×	✓	TF-IDF
1500	×	✓	-	-	×	✓	TF-IDF
2000	×	✓	-	-	×	✓	TF-IDF
3000	×	✓	-	-	×	✓	TF-IDF
4000	×	✓	-	-	×	✓	TF-IDF
5000	×	✓	-	-	×	✓	TF-IDF
6000	×	✓	-	-	×	✓	TF-IDF
7000	×	✓	-	-	×	✓	TF-IDF
8000	×	✓	-	-	×	✓	TF-IDF
9000	×	✓	-	-	×	✓	TF-IDF
10000	×	✓	-	-	×	✓	TF-IDF
11000	×	✓	-	-	×	✓	TF-IDF
12000	×	✓	-	-	×	✓	TF-IDF
13000	×	✓	-	-	×	✓	TF-IDF
14000	×	✓	-	-	×	✓	TF-IDF
15000	×	✓	-	-	×	✓	TF-IDF
16000	×	✓	-	-	×	✓	TF-IDF
17000	×	✓	-	-	×	✓	TF-IDF
18000	×	✓	-	-	×	✓	TF-IDF
19000	×	✓	-	-	×	✓	TF-IDF
20000	×	✓	-	-	×	✓	TF-IDF
Total	0	22	22	22	0	22	TF-IDF

Table 10: Other metric and take number of words

Set of words	Class	Precession	Recall	F-score	Correct	Incorrect	Kappa
18000	Positive	0.87	0.86	0.86	6651	985	0.729
	Negative	0.86	0.87	0.87	1046	6318	

accuracy rate, it outperforms the competition. Furthermore, the ROC curve clearly shows that the real positive rate for Multinomial Naive Bayes is higher, indicating that Multinomial Naive Bayes is superior in terms of performance. Below figure show the relation between precession and recall.

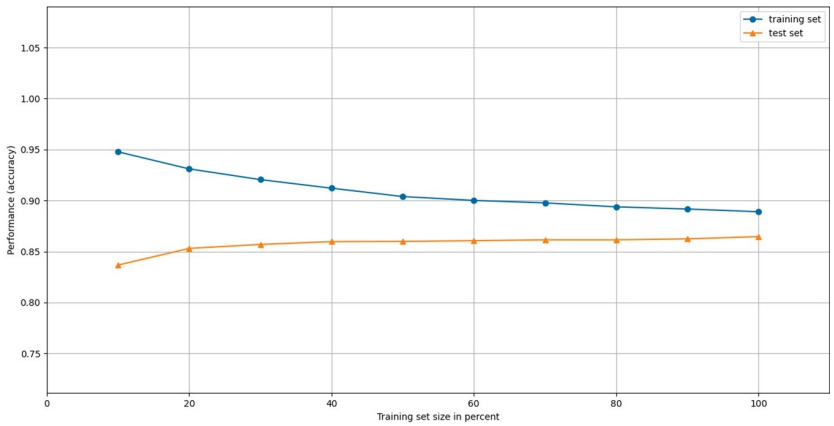

Fig. 12: Learning carve

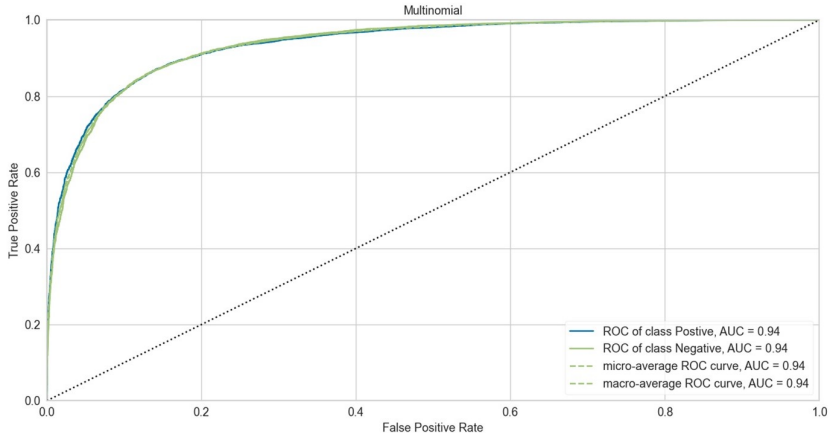

Fig. 13: AUC and ROC

Table 10 indicates that the positive class achieved 0.86 and the negative class achieved 0.87. The multinomial outperforms another Naïve Bayes model was completed at 86.46. Table 11 illustrates a comparison study between the proposed approach and other studies.

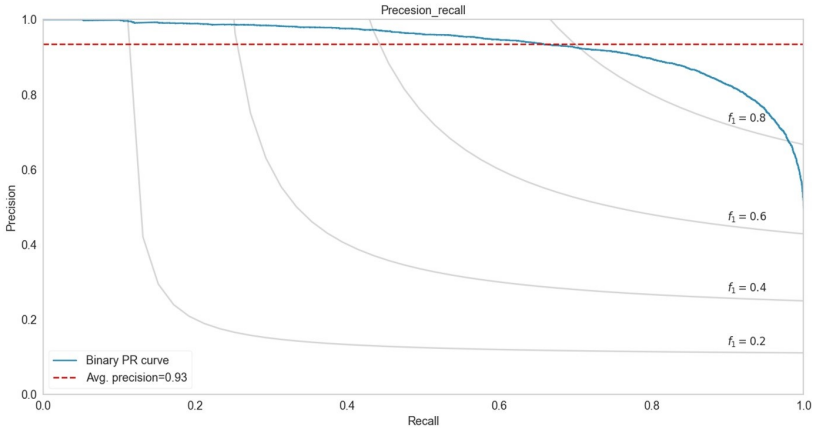

Fig. 14: Relation between precision and recall

Table 11: Comparison to other study

Ref	Objective	Data/Method	Confusion Matrix			Limitation
			Positive	Negative	Accuracy	
Moolthaisong et al[21]	Using Data Mining for movie reviews Classification.	Naïve Bayes, Random Forest J48 (using word frequency in movie reviews to create a word cloud).	462	238	80.25 79.83 68.06	Remove multiple lines and symbols, Remove stop words and Word Stemming
Boyko et al[26]	The influence of people's decisions on the opinions of others for movie review	SVM KNN NB (Knowledge of the frequency of words in documents).	12000	12000	76.35 62 73.87	Tokenization and delete stop words
Dupakuntla et al[27]	reviews or opinions expressed for movie review	Classify the polarization of reviews or opinions expressed into positive or negative feelings using the Naïve Bayes Classifier.	1693	1693	76.7	N/A
Tripathy et al[28]	The words with POS tags like adjectives and adverbs are being considered for SA.	Gaussian NB Multinomial NB Bernoulli NB (The term frequency is corpus's offset by the word's frequency).	8259 11107 11049	3744 1393 1451	75.7 83.1 82.7	Stop words, Numeric and special character removal and Vectorization
Nama et al[29]	Movie reviews SA and results in comparison with a rule-based approach using the AFINN-111 Sentiment Dictionary.	NB (Frequency distribution of the 10 most common words in the dataset).	700	700	80.10	Tokenize each word
Dhafar et al[30]	Computing the polarity score for each word occurring in the text for large movie review	System proposal that detects polarity through SA using dictionary methods.	25000	25000	76.585	Tokenisation, Unwanted characters elimination, Normalisation, Stop-words removal, Stemming and Remove less frequency word
Our approach	Opinion mining for large movie review	Using words frequency with four naïve Bayes models	25000	25000	86.46	Using only preprocessing not use light stemming or root stemming

6 Conclusion

6.1 Summary

Based on the results of the tests performed, it can be inferred that the number of words from the feature extraction results, the data pre-processing procedure and the cross-validation parameters significantly affect the classification process to achieve optimal results. . Four steps in the sentiment analysis process were considered: (i) pre-processing, (ii) feature extraction, (iii) data segmentation and (iv) classification. As part of our work, we used four Naïve Bayes data classification techniques and various parameters. Then, we implemented two types of tests: the first test was used for the TF coefficient and the second for the $TF - IDF$ coefficient. We applied these two coefficients which allowed us to derive between 500 and 20,000 characteristics. This results in better feature extraction with the IMDB and TF-IDF dataset and for both the multinomial and complementary method. Our approach gave the best results and that its performance can be improved by obtaining the best multinomial coefficient. This is possible given that our approach achieved 12 better accuracies than the complement when the word count reaches 18,000 words. The results obtained by our approach are 0.87% for precision, 0.86% for recall and 0.86% for F1-score. Our approach also has limitations and the accuracy of the classification of film reviews, given by this approach, needed to be improved.

6.2 Prospects

Our future work includes three directions. The first direction is to conduct a more in-depth comparative study, on other data collections, between our approach and the main approaches studied in the literature in order to give academics and practitioners more amplification on how to handle the problem. of opinion extraction from a film. Notice for SA. Through this study, we hope to definitively confirm the performance and robustness of our proposed approach. In a second step, we plan to study the possibility of hybridizing different algorithms, corresponding to the best solutions proposed in the literature, with deep learning models. This hybridization could improve the performances which could exceed those of the sub-approaches considered separately. This hybridization gave this proof for the case of SA and we believe it will also work for the problem of extracting opinion from movie reviews. In the third direction, we plan to include the use of ontologies in the deep learning model, which should further improve the results since current deep learning methods do not use ontology.

References

1. Nimesh V Patel and Hitesh Chhinkaniwala. Investigating machine learning techniques for user sentiment analysis. *International Journal of Decision Support System Technology (IJDSST)*, 11(3):1–12, 2019.

2. G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.
3. Fethi Fkih and Mohamed Nazih Omri. Hidden data states-based complex terminology extraction from textual web data model. *Applied Intelligence*, pages 1–19, 2020.
4. Sarra Ouni, Fethi Fkih, and Mohamed Nazih Omri. Toward a new approach to author profiling based on the extraction of statistical features. *Social Network Analysis and Mining*, 11(1):1–16, 2021.
5. Olfa Mabrouk, Lobna Hlaoua, and Mohamed Nazih Omri. Fuzzy twin svm based-profile categorization approach. In *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 547–553. IEEE, 2018.
6. Mondher Sendi, Mohamed Nazih Omri, and Mourad Abed. Possibilistic interest discovery from uncertain information in social networks. *Intelligent Data Analysis*, 21(6):1425–1442, 2017.
7. Kabil Boukhari, Mohamed Nazih Omri, et al. Approximate matching-based unsupervised document indexing approach: application to biomedical domain. *Scientometrics*, pages 1–22.
8. Noor Latiffah Adam, Nor Hanani Rosli, and Shaharuddin Cik Soh. Sentiment analysis on movie review using naïve bayes. In *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pages 1–6. IEEE, 2021.
9. Bee Wah Yap, Nazira Abdullah, Shuzlina Abdul-Rahman, and Michael Loong Peng Tan. Text mining and sentiment analysis on reviews of proton cars in malaysia. *Malaysian Journal of Science*, 37(2):137–153, 2018.
10. Shuhaida Mohamed Shuhidan, Saidatul Rahah Hamidi, Soheil Kazemian, Shamila Mohamed Shuhidan, and Maizatul Akmar Ismail. Sentiment analysis for financial news headlines using machine learning algorithm. In *International Conference on Kansei Engineering & Emotion Research*, pages 64–72. Springer, 2018.
11. T Sumathi, S Karthik, and M Marikannan. Performance analysis of classification methods for opinion mining. *International Journal of Innovations in Engineering and Technology (IJIET) Vol*, 2:171–177, 2013.
12. Vidisha M Pradhan, Jay Vala, and Prem Balani. A survey on sentiment analysis algorithms for opinion mining. *International Journal of Computer Applications*, 133(9):7–11, 2016.
13. DR Jeevanandam Jotheeswaran and YS Kumaraswamy. Opinion mining using decision tree based feature selection through manhattan hierarchical cluster measure. *Journal of Theoretical and Applied Information Technology*, 58(1):72–80, 2013.
14. Kabil Boukhari and Mohamed Nazih Omri. Dl-vsm based document indexing approach for information retrieval. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2020.
15. Meta Mahyarani, Adiwijaya Adiwijaya, Said Al Faraby, and Mahendra Dwifebri. Implementation of sentiment analysis movie review based on imdb with naive bayes using information gain on feature selection. In *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)*, pages 99–103. IEEE, 2021.
16. P Nagamma, HR Pruthvi, KK Nisha, and NH Shwetha. An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In *International conference on computing, communication & automation*, pages 933–937. IEEE, 2015.
17. Reza Maulana, Panny Agustia Rahayuningsih, Windi Irmayani, Dedi Saputra, and Wanty Eka Jayanti. Improved accuracy of sentiment analysis movie review using support vector machine based information gain. In *Journal of Physics: Conference Series*, volume 1641, page 012060. IOP Publishing, 2020.
18. Atiqur Rahman and Md Sharif Hossen. Sentiment analysis on movie review data using machine learning approach. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE, 2019.
19. Yanuar Nurdiansyah, Saiful Bukhori, and Rahmad Hidayat. Sentiment analysis system for movie review in bahasa indonesia using naive bayes classifier method. In *Journal of Physics: Conference Series*, volume 1008, page 012011. IOP Publishing, 2018.
20. Mais Yasen and Sara Tedmori. Movies reviews sentiment analysis and classification. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 860–865. IEEE, 2019.

21. Kamoltep Moolthaisong and Wararat Songpan. Emotion analysis and classification of movie reviews using data mining. In *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, pages 89–92. IEEE, 2020.
22. Anjali Agarwal, Roshni Rupali Das, and Ajanta Das. Machine learning techniques for automated movie genre classification tool. In *2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE)*, pages 189–194. IEEE, 2021.
23. Rini Wongso, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli, et al. News article text classification in indonesian language. *Procedia Computer Science*, 116:137–143, 2017.
24. Berna Seref and Erkan Bostanci. Sentiment analysis using naive bayes and complement naive bayes classifier algorithms on hadoop framework. In *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–7. IEEE, 2018.
25. Hiroshi Shimodaira. Text classification using naive bayes. *Learning and Data Note*, 7:1–9, 2014.
26. Nataliya Boyko and Karina Boksho. Application of the naive bayesian classifier in work on sentimental analysis of medical data. In *IDDM*, pages 230–239, 2020.
27. Vb Parthiv Dupakuntla, Hemish Veeraboina, M Vamsi Krishna Reddy, M Mohana Satyanarayana, and Ysai Sameer. Learning based approach for hindi text sentiment analysis using naive bayes classifier. *LEARNING*, 7(8), 2020.
28. Abinash Tripathy and Santanu Kumar Rath. Classification of sentiment of reviews using supervised machine learning techniques. *International Journal of Rough Sets and Data Analysis (IJRSDA)*, 4(1):56–74, 2017.
29. Vihaan Nama, Vinay Hegde, and B Satish Babu. Sentiment analysis of movie reviews: A comparative study between the naive-bayes classifier and a rule-based approach. In *2021 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE, 2021.
30. Dhafar Hamed Abd, Ayad R Abbas, and Ahmed T Sadiq. Analyzing sentiment system to specify polarity by lexicon-based. *Bulletin of Electrical Engineering and Informatics*, 10(1):283–289, 2021.