



ISSN: 0067-2904

## A Comparative Study for Supervised Learning Algorithms to Analyze Sentiment Tweets

Fatema Hassan Fadel\*, Suhad Faisal Behadili

Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Received: 25/7/2021

Accepted: 25/9/2021

Published: 30/6/2022

### Abstract

Twitter popularity has increasingly grown in the last few years, influencing life's social, political, and business aspects. People would leave their tweets on social media about an event, and simultaneously inquire to see other people's experiences and whether they had a positive/negative opinion about that event. Sentiment Analysis can be used to obtain this categorization. Product reviews, events, and other topics from all users that comprise unstructured text comments are gathered and categorized as good, harmful, or neutral using sentiment analysis. Such issues are called polarity classifications. This study aims to use Twitter data about OK cuisine reviews obtained from the Amazon website and compare the effectiveness of three commonly used supervised learning classifiers, Naive Bayes, Logistic Regression, and Support Vector Machine. This is achieved by using two method of feature selection involving count Vectorizer and Term-Frequency-Inverse Data Frequency. The findings showed that the support vector machine classifier had achieved the highest accuracy of 91%, by feature selection: Count Vectorizer. But it is time consuming. For both accuracy and execution time concentrates, logistic regression is recommended.

**Keywords:** Social Networks, Data Mining, Sentiment Analysis, Opinion Mining, NLP, Confusion Matrix.

### دراسة مقارنة لخوارزميات التعلم الخاضعة للإشراف لتحليل تغريدات المشاعر

فاطمة حسن فاضل\* , سهاد فيصل شيخان

قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

### الخلاصة

زادت شعبية تويتر على نحو متزايد في السنوات القليلة الماضية ، مما يؤثر على جوانب الحياة الاجتماعية والسياسية والتجارية. ترك الناس تغريداتهم على مواقع التواصل الاجتماعي حول حدث معين ، وفي نفس الوقت يستفسرون لرؤية تجارب الآخرين إذا كان لديهم رأي إيجابي/سلبي حول هذا الحدث . ويمكن استخدام تحليل المشاعر للحصول على هذا التصنيف. ويتم جمع استعراضات المنتجات والأحداث وغيرها من المواضيع من جميع المستخدمين التي تشمل تعليقات نصية غير منظمة وتصنيفها على أنها جيدة أو ضارة أو محايدة باستخدام تحليل المشاعر. وهذه المسائل هي ما يسمى بالتصنيفات القطبية. تهدف هذه الدراسة الى استخدام بيانات تويتر عن استعراضات الاطعمه التي تم الحصول عليها من موقع امازون ومقارنه ثلاثه من

\*Email: [fatema.fadel1201@sc.uobaghdad.edu.iq](mailto:fatema.fadel1201@sc.uobaghdad.edu.iq)

مصنفات التعلم المشرفه المستخدمه عاده والتي باستخدام طريقتين لاستخراج الخصائص (naive Bayes) و (support vector machine), (logistic regression), تتضمن وقد حققت خوارزميه (Term-Frequency-Inverse Data Frequency) و ( count vectorizer ) اعلى دقه بنسبه 91 بالمئه وكانت الكثر استهلاكه للوقت باستخدام support vector machine

## 1- Introduction

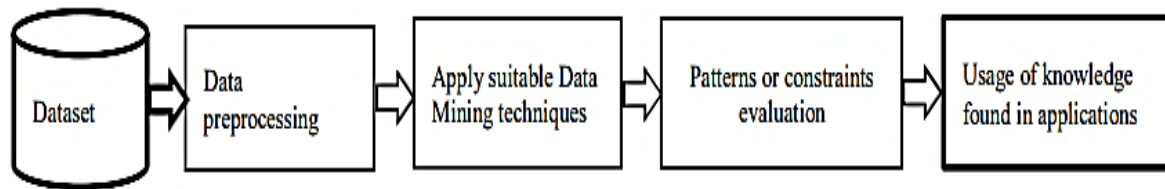
The recent decade showed increased progress in developing applications that employed on-screen sites of networking microblogging and assets like YouTube, Facebook, and Twitter. For several associations and organizations, such applications with their wealth of assets become valuable source of marketing information. Mostly, the organization conducted reviews and meetings to obtain the product quality and to enhance the response. However, these traditional methods were costly, intensive, and did not produce the results organizations were looking for due to ill-structured studies and eco-friendly issues.

Nowadays, a sizeable volume of client sentiment data, which is upsetting to get the meaning or has unstructured content for computers, is uploaded on the internet every day. Lately, feedback on services and products is collected using marking policies, mainly dependent on natural language processing and sentiment analysis. Note that a process of analyzing the text opinion is known as sentiment analysis [1]. For a particular event in social media, people are interested to know the experiences of the others (positive/negative) about that event before they post their comments. Analyzing sentiment is performed at three levels, document, aspect, and sentence level.

Twitter is a valuable information source for obtaining product quality reviews. It utilizes tweets that are in sentence forms to imply opinions. Companies' effective way to get people's opinions on their recently marketed products is sentiment analysis over Twitter [2]. The objective is to find the sentiment accuracy of the posted tweets.

Sentiment analysis classifies the tweets as negative or positive. In this analysis, the most significant part of the challenge is the opinion words. The opinion word could be negative or positive based upon the case. However, traditional text processing systems cannot alter the content, meaning if a slight variation in words is there. In contrast, if changes in two words are there, sentiment analysis can alter the content meaning. For instance, the sentence "The phone is not ringing" is not the same as "The phone is ringing". The user can understand the informal sentence, while the System cannot [3]. Furthermore, machines are restricted, and the type of text to be processed must be carefully considered, as this has a significant impact on the size of vocabulary that the system must learn and the quantity of the text to be analysed.

Natural Language Processing (NLP) is a branch of computer science and linguistics that aims to equip robots with the ability to understand natural languages like English as well as Arabic. Sentiment Analysis is a subdomain of NLP that examines the method of utilizing machines to process texts and assign each one a categorization that we can use and comprehend. This domain employs language processing techniques to extract features including word frequency, as well as supervised learning algorithms to learn from a set of data that's been previously classified by a person [4]. Classifier supervised learning models, clusters (unsupervised learning models), prediction, association rules and neural networks are aspects for Data Mining (DM) [5]. The current real-world problems are examined using the Data Mining (DM) approach, which is a new domain with various techniques. Numerous DM techniques were used in projects for acquiring knowledge from databases. Figure 1 illustrates steps of DM [6].



**Figure 1-** Architecture of Data Mining

This work used Twitter data about fine cuisine reviews obtained from the Amazon website in performing sentiment analysis and the main goal was to build a system that will be able to classify each input review as positive or negative emotion by comparing the effectiveness of three commonly used supervised learning classifiers, involving Naive Bayes, Logistic Regression, and Support Vector Machine (SVM). It transforms raw data into helpful information for different research fields and finds their patterns to decide future tendencies for the research domain. This study is arranged as follows: section two reviews the related work, section three explains the presented study, section four discusses the findings of sentiment analysis, and section five presents the conclusion.

## 2 -Related Work

Machine learning techniques have been applied to analyze tweet sentiments in the last two decades. [7] built a classifier based on query term to classify Twitter-message sentiments into positive and negative categories. Before buying a new product, consumers want to know the review sentiments. Thus, this classification was highly significant for both consumers and companies that monitor their products. Note that no previous work has been performed to classify message sentiments through microblogging services such as Twitter. However, machine learning algorithms have achieved an accuracy of about 80% after they trained with emotion data. For SVM, MaxEnt, and Naïve Bayes algorithms with unigram, they have achieved 82.9%, 80.4%, and 81% accuracy, respectively.

[8] Implemented an opinion mining system to determine sentence negation and polarity, the system generates summarized results that allow the user to make decisions based on these results. In general, sentiment analysis is a significant issue for those who need to know the other opinions about the product to arrange its pros and cons. The researcher in [9] dissected the tweets to obtain the Twitter-data sentiments. In contrast, in [10] they found that sentiment analysis is beneficial in many fields such as security, governance, etc., where it can be utilized for various tasks at various levels. For sentiment analysis, one good way is to find the accuracy of the used machine learning techniques. The efficiency of each technique, as compared to others, does not usually clarify the available evaluation metrics because the user dataset changes based on the specified work. In another work the same authors used random forest, SVM, and Naïve Bayes techniques to classify the movie reviews into positive and negative. The accuracies of random forest and SVM techniques are increased using the hyper-parameter. In contrast, the human-labeled document needs only simple effort using lexicon-based methods [11].

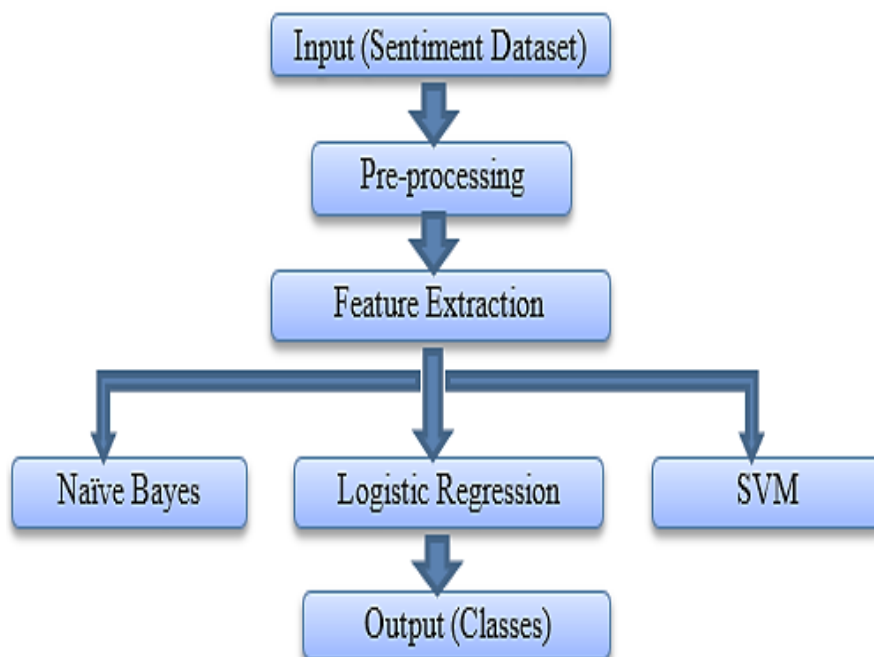
On the other hand, as found by [12]. Comparing to other models, they have achieved better accuracy with the bigram model in SVM. They also stated that cleaner data can obtain more accurate results. Recently, in [13] trained an SVM classifier to achieve the pre-labeled Twitter data. The polarity of tweets was decided using Twitter hashtags. The classifier conducted a test study for analyzing the accuracy of the introduced method. The result showed an accuracy of 85%. In contrast, [14] considered the hashtag polarity in the function of a tweet classification feature within the political domain. Based on positive/negative hashtags, they introduced the regulations for automated dataset labeling. Finally, they presented a method for enriching the tweet throughout extracting the hashtag term. The authors have achieved more

than 95% accuracy for sentiment classification and dataset labeling using positive/negative hashtags.

Furthermore, when Naïve Bayes, Logistic Regression, or SVM algorithms is combined with this hashtag feature, it will outperform the unigram feature. Conversely, when the Random Forest algorithm is combined, the accuracy will decrease depending on the time required for building the model. [15] studied sentiment analysis using two approaches: machine learning or sentiment lexicons. The first approach classifies the text using SVM and Naïve Bayes techniques, while phrases or words for a sentence are computed in the second approach. The machine learning approach obtained that the SVM technique has better accuracy of 85% than the Naïve Bayes technique. Better accuracy is obtained with the lexicon-based approach. Lastly, the techniques of Twitter sentiment analysis are classified into two approaches, term frequency-based, and machine learning. Their survey offers a classification of several current articles based on the technique used in sentiment analysis, which helps the user select the suitable technique for his/her work.

### 3 -Materials and methods

The proposed work uses three widely used machine learning classifiers, Naïve Bayes, logistic regression, and SVM, to classify the sentence in tweets. The procedure initially performed pre-processing tasks on the input dataset. Then, the feature extraction stage and applying the classification algorithm to obtain the results. The procedure steps of the presented work are shown in Figure. 2.



**Figure 2-**The procedure steps of the presented work.

#### 3.1. The dataset

Initially, the dataset was taken from the kaggle website [16], from October 1997 to October 2012, with over 5500 reviews. In addition, it comprises reviews from the whole other categories of Amazon as illustrated in Table 1.

**Table 1-** Attribute Information of dataset

Markers	Description
<b>Id</b>	Sequence
<b>Product ID</b>	unique identifier for the product
<b>user ID</b>	unique identifier for the user
<b>Profile Name</b>	Name of user profile

<b>Helpfulness Numerator</b>	number of users who found the review helpful
<b>Helpfulness Denominator</b>	number of users who found the review helpful or not
<b>Score</b>	rating between 1 and 5
<b>Time</b>	timestamp for the review
<b>Summary</b>	a brief summary of the review
<b>Text</b>	text of the review

The training dataset is in a CSV format file that Includes tweet date, time, sentiment, id, and when the tweets were posted. The product/event text opinion is referred to as a tweet. The dataset contains a collection of personal references, URLs, symbols, emoticons, and words. There are also misspelled words, excessive punctuation, and words including several repetitive letters within the words. As a result, pre-processing tasks are required to convert the data into an orderly format.

### 3.2. The Pre-processing steps

Twitter tweets collected in their raw form produce a noisy dataset. Such an issue is due to the informal nature of social media usage by people. For building a dataset that can train a variety of classifiers, the raw Twitter data must be standardized. Data is pre-processed to reduce its size and to standardize the dataset.

Pre-processed datasets are used for training and testing. Retweets, emoticons, user mentions, and other unique aspects of tweets must be retrieved appropriately. Lemmatization, stop-word removal, and remove undesirable tags, web-links, stemming (the use of word stemming to return derived words to their origins using the iterated Lovins stemmer), etc., may lead to erroneous results, which comprises Natural Language Processing [17]. This work includes various pre-processing like substituting single space instead of two or more spaces, strip quotes (" and ') and spaces from the tweet ends, substituting space instead of two or more dots (.), changing the tweet to lower case, etc. For instance, the sentence "product extremely flexible use" is the outcome of pre-processing the sentence "this product is extremely flexible to use!", and removing the noisy data or special characters like (@ # \* "/ : > , |?). The Porter stemming algorithm, which begins by examining the phrase and follows a set of criteria, is used to achieve stemming. Next, is removing any endings that turn the keyword into plurals, such as '-s' as '-es,' past tenses like '-ed,' or continuous tenses like '-ing.' The stemmer then validates and changes double suffixes to a single suffix. Suffixes with other endings '-ic', '-full', '-ness', '-ant', and '-ence' are among the words that have been eliminated. Such as the words "organized," "organizing," "organization," and "organizations" should all be expressed by the term "organize" [4].

### 3.3. The Feature extraction

Because many DM algorithms do not function with textual input, several strategies were applied to extract features in this study. Textual data should be transformed into a binary or numerical format to allow the algorithm to interpret and process the results. Two notable feature -extraction approaches used in the proposed system were [18]:

**1. Count Vectorizer:** It is called one-hot encoding, which generates a vector of similar size to the vocabulary. It is checked in the lexicon when a word appears in a phrase and gives a rating of 1. The word count is increased when a word appears more than once in a document. The word will be added if they do not already exist in the dictionary. The count vectorizer operation is explained in depth in that example showed in Figure.3 [19].

**2. TF-IDF:** For a given word ( $x$ ) in a document ( $N$ ), the term frequency (TF) score is defined as the ratio of the number of that word incidences to the total number of words in that document, while the inverse data frequency (IDF) is the ratio of the total number of documents to the number of documents where that word appeared. The TF score and IDF score are represented mathematically in equations 1,2 and 3 [20].

$$TF \text{ score of word}(x) = \frac{\text{number of times } x \text{ appeared}(n)}{\text{total words in the document}(N)} \dots \dots \dots (1)$$

$$IDF \text{ score of word}(x) = \log \frac{N}{D} \dots \dots \dots (2)$$

$$TF - IDF \text{ score} = TF \text{ score} * IDF \text{ score} \dots \dots \dots (3)$$

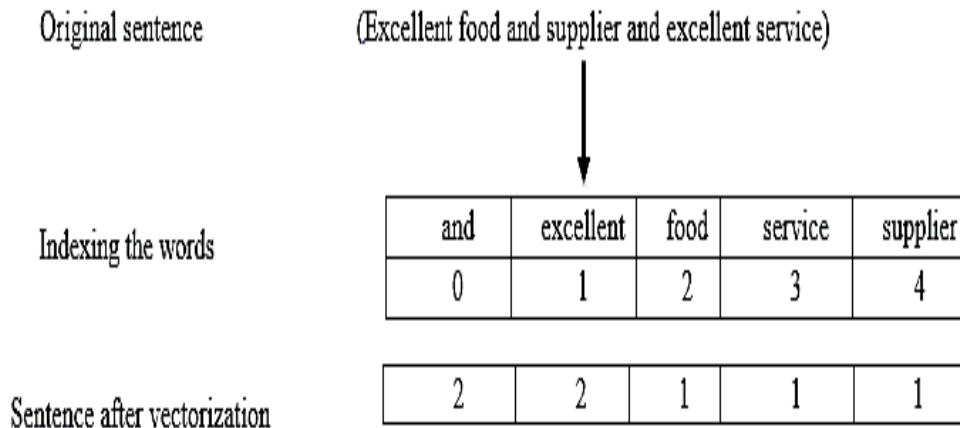


Figure 3- Example of the count vectorizer.

**3.4. The Classification steps**

It is an assigning process that assigns a class label to samples from the problem under consideration using machine learning algorithms. An easy example to understand this task is classifying emails as “read” or “not read”. Various types of classification tasks are available that may encounter within both specialized and machine learning approaches for modeling, which can be used for each. The input dataset must be divided into three sub-datasets for machine learning use: training, testing, and validating datasets [21].

**Training dataset:** It is used to define the optimal model parameters and to train the model.

**Testing dataset:** It is used to evaluate the capability of the model for generalization and its performance. After training the model, its capability for identifying patterns in a new, unused dataset. To avoid model overfitting, the same dataset must not be used for both model training and testing.

**Validating dataset:** It is used to tweak the hyper-parameters of the model, which cannot be learned directly from the dataset, as they are settings of high-level structures. Note that these settings are beneficial to obtain how quickly the model can find patterns in a dataset and how complex a model is. Figure 4 illustrates the input dataset partitioning process. In practice, data is divided either by K-fold cross validation, in which data is divided into k-number of blocks called fold (suppose k=5) and Figure 5 illustrates the structure of

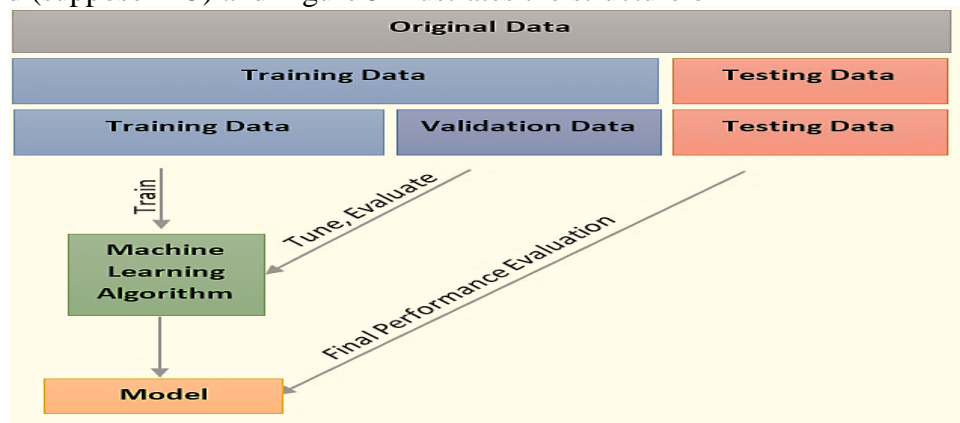


Figure 4-Input dataset partitioning process.

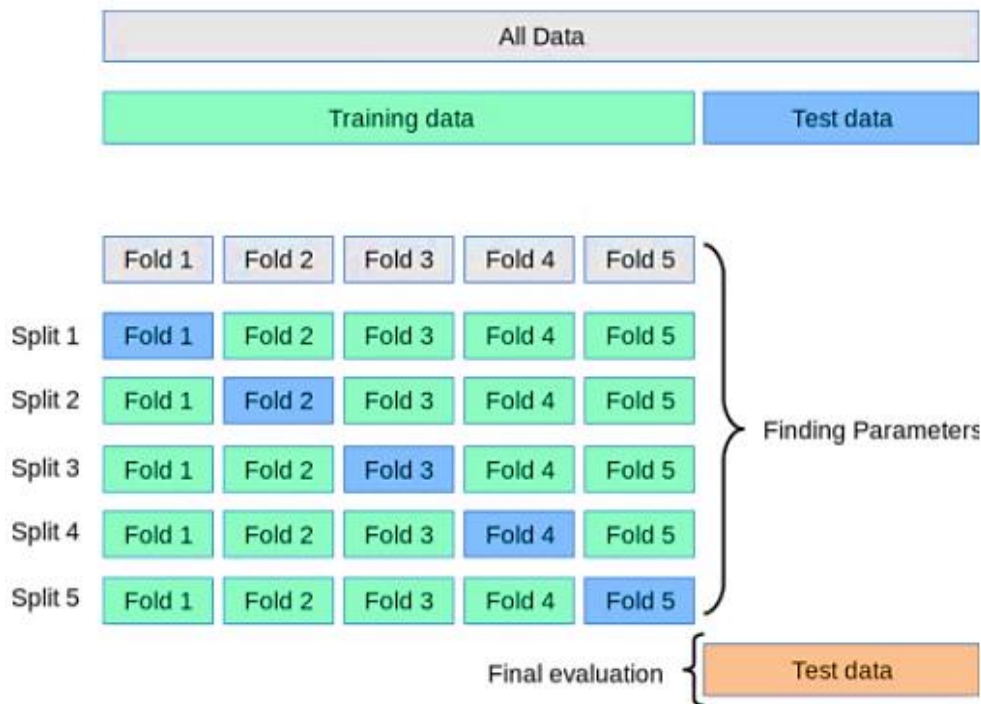
K-fold cross validation, the training dataset represents 80% of the input dataset, while the remaining data is 20% for the testing dataset. In our work, the internal data was classified by review to two classes positive and negative as shown in Figure 6. This work used three popular classifiers, which are explained in the following [22] [23]:

**3.4.1 Support Vector Machine (SVM) Approach**

It is the most widely used supervised machine learning classifier. Instead of the original input space, SVM uses a non-linear kernel function to turn the input data into a high-dimensional function space, making it easier to separate the data. This algorithm aims to make the best decision boundary or line that can segregate *n-dimensional* space into classes. Therefore, it is straightforward to put a new data point in the suitable category in the future. Note that a hyperplane is a name for the optimal choice boundary. SVM iterative learning processes will be influenced by the data provided. Finally, in *high – dimensional* feature space, create ideal hyperplanes with the greatest possible margin between classes. As a result, the greatest margin of hyperplanes will be used as separation boundaries for different data classes. Therefore, the increased distance between hyperplanes and group data will improve classification efficiency. This algorithm concept is depicted in Figure 7. The SVM classifier, on the other hand, is built as follows as in equation 5:

$$f(t) = \sum_i \alpha_i \gamma_i (X_i \cdot T) + b \quad \dots \dots \dots (5)$$

Where,  $\gamma_i$  are the class labels that are assumed to map to 1 or -1 of  $X_i$ , which are the support vectors. Vector  $T$  represents a test sample. Thus,  $(X_i \cdot T)$  is the dot product of one of the support vectors  $X_i$  with the test sample  $T$ ,  $\alpha_i$  and  $b$  are the learning method that will determine numerical parameters such as weights [24]. Furthermore, the algorithm steps for SVM firstly define the best hyperplane. Secondly, step I should be extended for nonlinearly separable issues. Then finally, transform data into a high-dimensional space where linear decision surfaces can easily classify it.



**Figure 5-** K-fold cross validation structure

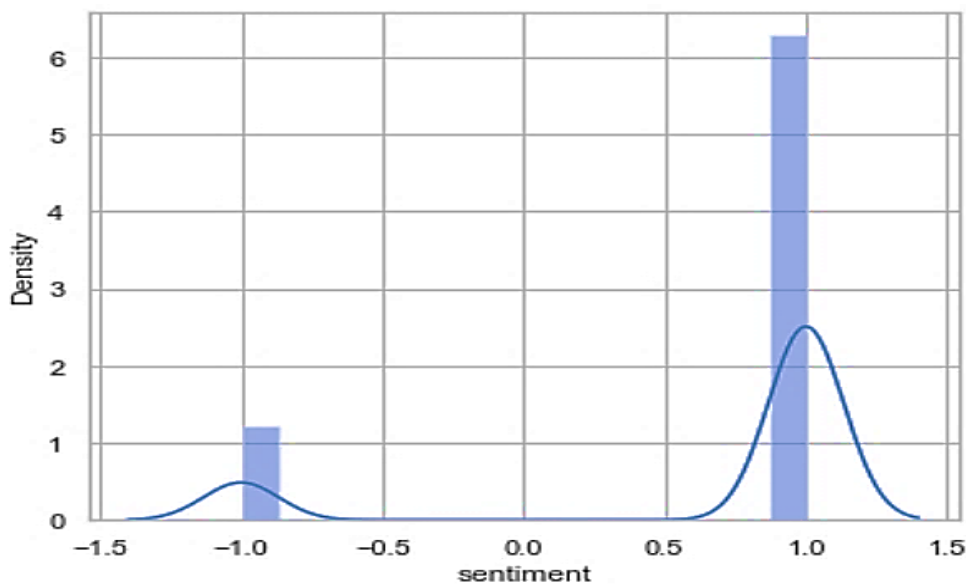


Figure 6-Classified sentiment classes as negative and positive ones.

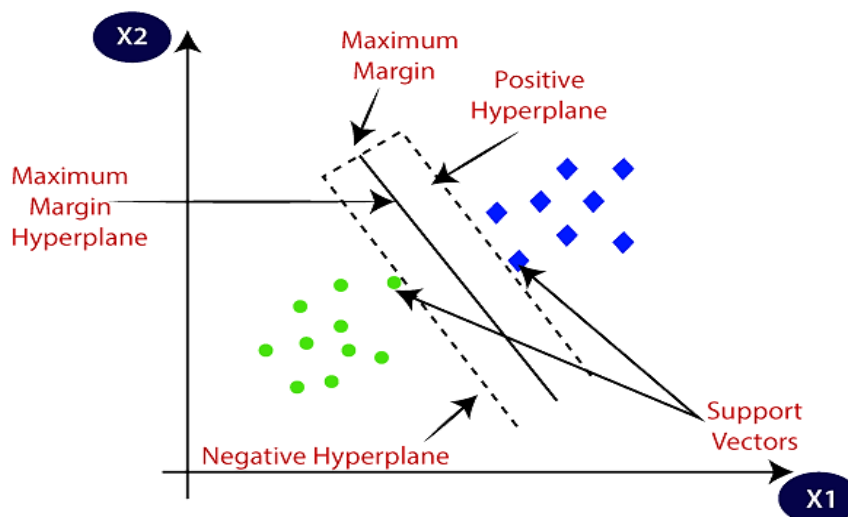


Figure 7- SVM algorithm concept.

### 3.4.2 Naïve Bayes Approach

It is an essential text classification model. This classifier determines the likelihood of an object having certain characteristics that belong to a specific class or group. Because it assumed that the incidence of one characteristic is independent of the incidence of other features, this method is also known as "Nave." As a result, it gives a method for computing conditional probability, or the probability of an incidence is dependent on prior knowledge about the events [25].

The following is the Bayes theorem as in equation 6:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \dots \dots \dots (6)$$

Where,  $P(A/B)$  - Conditional probability of incidence of event  $A$  given that event  $B$  is true. And  $P(A)$  and  $P(B)$  are the probability occurrence of the event  $A$  and  $B$ , respectively. Also,  $P(B/A)$  – the probability of incidence of event  $B$  given that event  $A$  is true.



### 3.4.3 Logistic Regression Approach

To simulate a linear relationship between one response variable and one explanatory variable, simple linear regression [26] can be utilized. As demonstrated in Figure 8, linear regression has been used to solve various significant scientific and societal challenges. It is assumed that the response variable and the explanatory variable have a linear connection. It simulates this relationship using a hyperplane, a subspace with one fewer dimension than the ambient space in which it exists. In basic linear regression, the response variable has one dimension, while the explanatory variable has another for a total of two dimensions. As a result, the regression hyperplane has only one dimension; a line [26] is a hyperplane with only one dimension. Typically, all estimators implement the *fit()* and *predict()* methods.

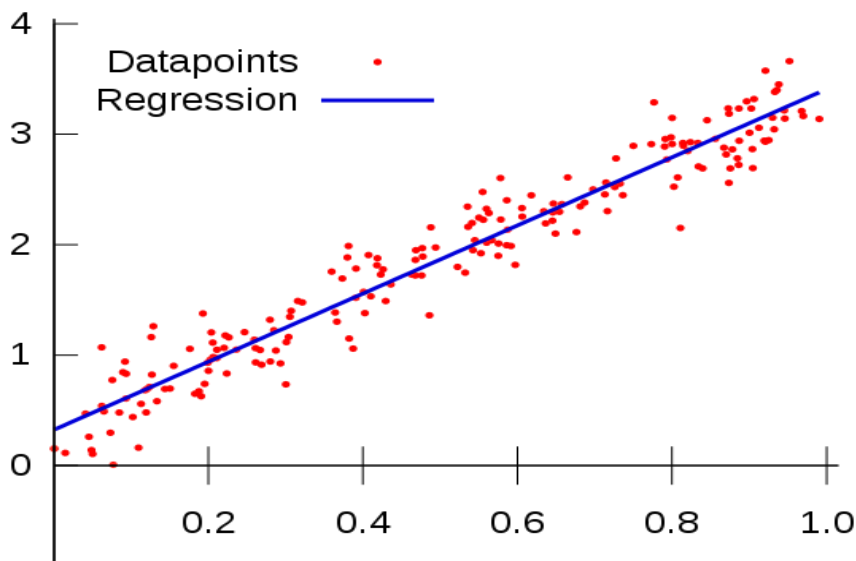


Figure 8- Represent of LR.

The former method is used to learn a model's parameters, whereas the latter method uses the learned parameters to forecast the value of a response variable given an explanatory variable. Because all estimators implement the fit and predict methods, it is simple to try different models using Scikit-learn. However, for simple linear regression, the fit technique of Linear Regression learns the following model parameters, as shown in equation 7.

$$y = \alpha + \beta x \quad \dots \dots \dots (7)$$

Where,  $y$  is the response variable expected value,  $x$  is the explanatory variable. The intercept term, on the other hand,  $\alpha$  and coefficient  $\beta$  are the model parameters that were learned by the Learning algorithm [26].

### 4 -Results and Discussion

This work aims to determine the accuracy of each classifier and make a comparison between them. The confusion matrix, or error matrix, is a table structure that visualizes the performance of a classification model and is used to evaluate a model's performance. Usually, it is a supervised learning situation (in unsupervised learning, it is ordinarily termed a matching matrix). Each property in the matrix explains the cases in an actual class, while each record in the matrix describes the cases in a predicted class (or vice versa) since there are two types of classes. Table 2 lists the detailed description of the confusion matrix [27].

**Table 2-** Structure of confusion matrix

	Class 1 predicted	Class 2 predicted
Class 1 actual	True positive	False positive
Class 2 actual	False negative	True negative

Note that True positive (*TP*) is a classification type that predicts positive outcomes and is true. False positive (*FP*) is a classification type that predicts a positive outcome but is incorrect. False negative (*FN*) is a classification type that predicts a negative outcome but is incorrect. True negative (*TN*) is a classification type that predicts a negative outcome and is accurate [28]. The following equations are the most often used metrics in a confusion matrix, including accuracy, precision, recall, and *F1 – score*.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots \dots \dots (8)$$

$$Precision = \frac{TP}{TP + FP} \dots \dots \dots (9)$$

$$Recall = \frac{TP}{TP + FN} \dots \dots \dots (10)$$

$$F1score = \frac{2}{\left(\frac{1}{precision} + \frac{1}{recall}\right)} \dots \dots \dots (11)$$

Table 3 lists the findings of the three classifiers for Count Vectorizer feature extraction, while Table 4 lists the findings of the three classifiers for *TF – IDF* feature extraction. The most critical factors, accuracy and execution time, and two classes and their related feature scores, are compared. It was observed from the data given that it went towards positive comments, and the most interesting parameter was the time when the commentary and score were written for comment. The system efficiency is demonstrated by the experimental results. This task indicates that it consumes extra time to read and analyze the bins and produces better results. The idea behind that is more likely as an emulation of the human brain so that the modal consumes much time to get more accurate results.

From both tables, it can be seen that the SVM classifier consumes the longest time to learn but has the highest accuracy. To decide which model is the best one is dependent on the purpose of the application. For example, the Logistic Regression accuracy is less than of SVM, but it is the fastest in terms of the execution time, making it very applicable for real-time purposes. In addition, the most relevant parameter was positive feelings that individuals displayed at proximity times, which are the proper periods for meals in the total of data mentioned earlier, based on the results obtained from the actual user data. Other parameters (*Precision, Recall, F1 – score, and Support*) and classes (*Negative, Positive, Macro Avg., and Weighted Avg.*) are affected directly by the accuracy that approves our results. The feature extraction that helped give him more accuracy is TF-IDF. As well as table 5 included the carried-out figures from results confusion matrix for used models. Meanwhile, the figures at first row within table 5 represent the counter vectorizer according to SVM show that makes 1103 correct predictions. Logistic regression makes 957 correct predictions, and Naïve Bayes makes 772 correct predictions respectively. However, the second row within table 5 represents the TF-IDF and show that the lowest ratio of inaccurate predictions was found. The concatenated features are responsible for the improved results, giving the model more features to learn and enhance accuracy, according to SVM makes 1118, Logistic regression makes 998, and Naïve Bayes makes 976 of correct predictions ,respectively

Additionally, figure 9 showed the performance measures of the three models used, this figure illustrates that the accuracy of Receiver Operating Curve (ROC) was used to compare the outcomes of all classifiers using 10-fold cross-validation. ROC curves that are recommended for use in classification issues when evaluating them can provide an overly optimistic image of an algorithm effectiveness.

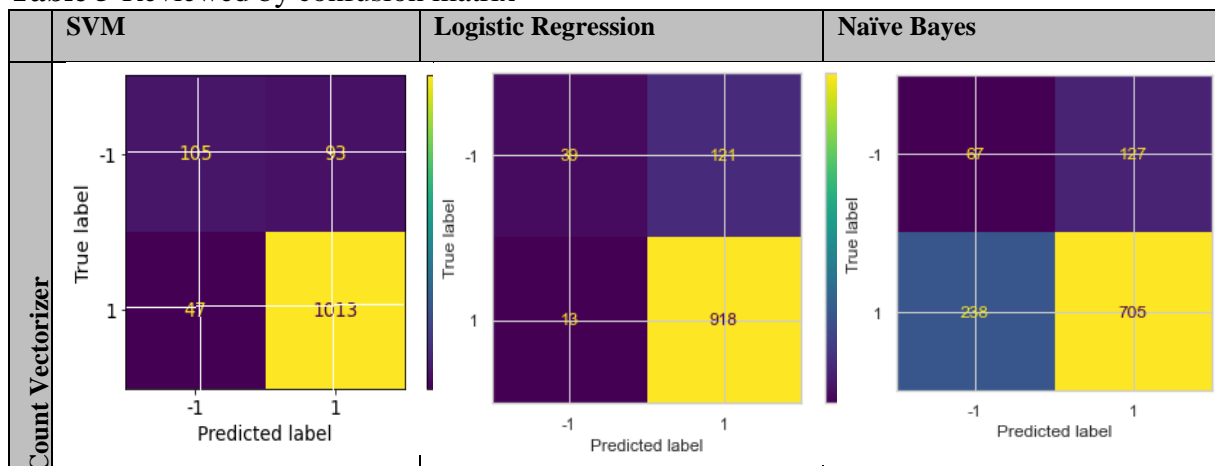
**Table 3-**Competitive for three machine learning algorithms for Count Vectorizer

Model	Accuracy	Classes	Precision	Recall	F1-score	Support	Testing time (sec.)
SVM	89%	Negative	0.53	0.69	0.60	152	3.4145
		Positive	0.96	0.92	0.94	1106	
		Macro Avg.	0.74	0.80	0.77	1258	
		Weighted Avg.	0.90	0.89	0.89	1258	
NB	68 %	Negative	0.35	0.22	0.27	305	1.3852
		Positive	0.75	0.85	0.79	832	
		Macro Avg.	0.55	0.53	0.53	1137	
		Weighted Avg.	0.64	0.68	0.65	1137	
LR	85%	Negative	0.21	0.75	0.32	51	1.0119
		Positive	0.99	0.86	0.92	1037	
		Macro Avg.	0.60	0.80	0.62	1088	
		Weighted Avg.	0.95	0.85	0.89	1088	

**Table 4-**Competitive for three machine learning algorithms for TF-IDF

Model	Accuracy	Classes	Precision	Recall	F1-score	Support	Testing time (sec.)
SVM	91 %	Negative	0.55	0.80	0.65	135	5.8271
		Positive	0.97	0.92	0.95	1082	
		Macro Avg.	0.76	0.86	0.80	1217	
		Weighted Avg.	0.93	0.91	0.91	1217	
NB	81 %	Negative	0.12	0.94	0.22	33	1.8554
		Positive	1.00	0.81	0.89	1168	
		Macro Avg.	0.56	0.87	0.55	1201	
		Weighted Avg.	0.97	0.81	0.87	1201	
LR	88 %	Negative	0.42	0.75	0.53	106	1.4192
		Positive	0.97	0.90	0.93	1084	
		Macro Avg.	0.69	0.82	0.73	1190	
		Weighted Avg.	0.92	0.88	0.90	1190	

**Table 5-**Reviewed by confusion matrix



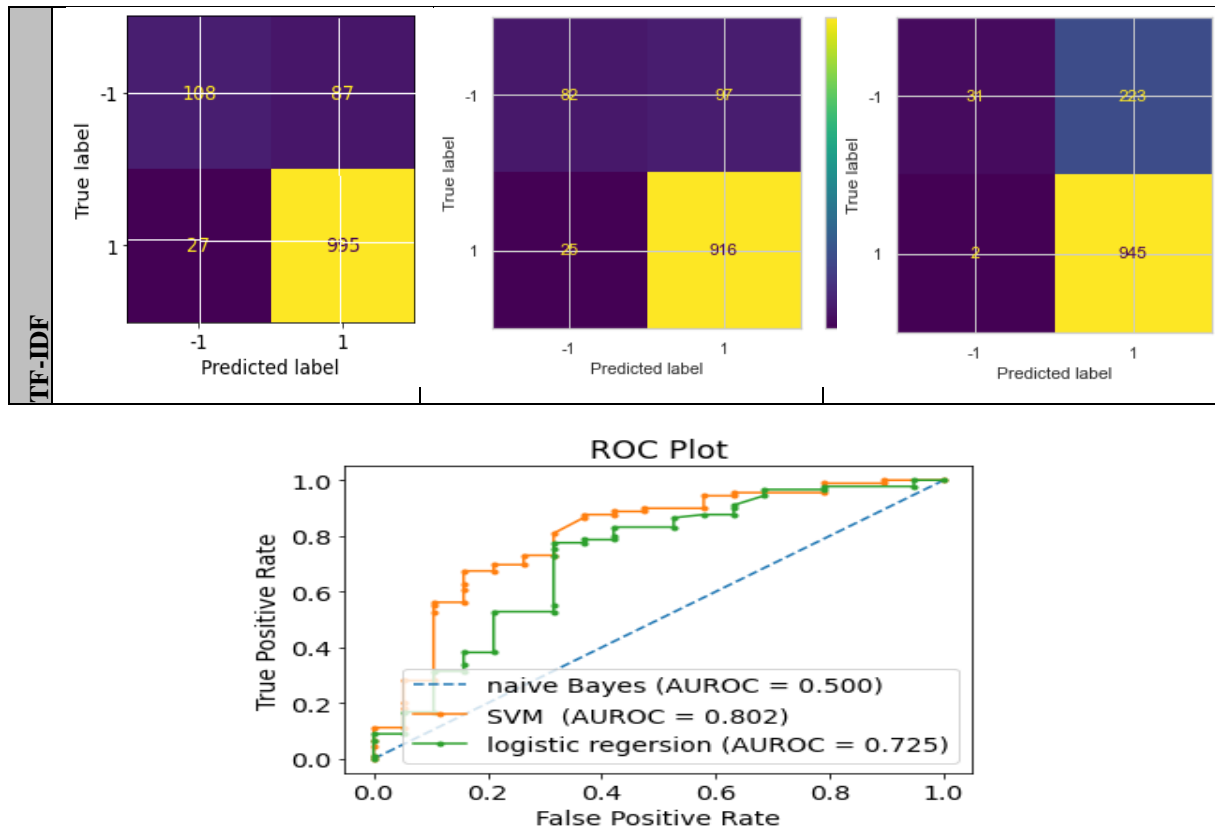


Figure 9-The Roc Plot of the three models.

5 -Conclusions

Sentiment analysis has become a popular technique in recent years, offering a unique service to individuals as well as organizations on a daily basis. When it comes to manufacturers, there is a huge demand for this technique. For instance, these businesses can acquire information on their products, services, policies, or anything. Individuals can use it to ask questions about a movie, a product they want to buy, a certain subject, or anything else. As a result, anyone who needs to learn more about a subject can use this strategy to save effort and time. However, the accuracy of the system is improved by good Feature Extraction effort, which is so important since it decreases the number of features available by selecting the most useful features collection. So, the number of features is reduced, which implies there are fewer of them. An opinion word of positive/negative is a difficult challenge in sentiment analysis. By adjusting the parameters of the machine learning classifiers, higher accuracy can be achieved. In this examination, a comparative study has been implemented using three machine learning classifiers that involved Naïve Bayes, Logistic regression, and SVM classifiers. The SVM classifier achieved the highest accuracy of 91% by the feature selection (TF-IDF). However, if both accuracy and execution time are taking into account as concentrates, then logistic regression is recommended. For future work, examining the effectiveness of the sentiment analysis algorithms with various features are considered. Other active learning techniques such as uncertainty sampling, pool-based sampling, and anticipated error reduction can be applied to identify Twitter sentiments and boost confident decision-making.

Reference

[1] S. G. Kanakaraddi, "Comparison Study of Sentiment Analysis of Tweets using Various Machine Learning Algorithms," 2020 *International Conference on Inventive Computation Technologies (ICICT)*, pp. 287-292, 2020.

[2] N. Nandal, "Machine learning based aspect level sentiment analysis for Amazon products," *Spatial Information Research*, pp. 601-607, 2020.

- [3] A. Poornima, "A comparative sentiment analysis of sentence embedding using machine learning techniques," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020.
- [4] H. A. Alatabi, "Sentiment analysis in social media using machine learning techniques," *Iraqi Journal of Science*, pp. 193-201, 2020.
- [5] T. K. Mustafa, "Proposed approach for analysing general hygiene information using various data mining algorithms," *Iraqi Journal of Science*, vol. 58, no. 1B, pp. 337-344, 2017.
- [6] M. S. Abd, "Recognizing job apathy patterns of Iraqi higher education employees using data mining techniques," *Journal of Southwest Jiaotong University*, vol. 54, no. 4, 2019.
- [7] W. Shi, "Sentiment analysis of Chinese microblogging based on sentiment ontology: a case study of '7.23 Wenzhou Train Collision'," *Connection Science*, vol. 25, no. 4, pp. 161-178, 2013.
- [8] S. N. a. R. J. R. Sharma, "Polarity detection at sentence level," *International journal of computer applications*, vol. 86, no. 11, 2014.
- [9] S. Khowal, "Review Of Natural Language Processing Inspired Sentiment Analysis Algorithms".
- [10] Z. Madhoushi, "Sentiment analysis techniques in recent works," in *2015 science and information conference (SAI)*, 2015.
- [11] S. V. Wawre, "Sentimental analysis of movie review using machine learning algorithm with tuned hypeparameter," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 6, pp. 12395-12402, 2016.
- [12] V. Kharde, "Sentiment analysis of twitter data: a survey of techniques," arXiv preprint arXiv:1601.06971, 2016.
- [13] O. H. A. ALQARYOUTI, "Aspect-Based Sentiment Analysis for Government Smart Applications Customers' Reviews," 2017.
- [14] I. Alfina, "Utilizing hashtags for sentiment analysis of tweets in the political domain," in *Proceedings of the 9th International Conference on Machine Learning and Computing*, 2017.
- [15] A. Hasan, "Machine learning-based sentiment analysis for twitter accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 11, 2018.
- [16] " <https://www.kaggle.com/snap/amazon-fine-food-reviews>".
- [17] A. Al-Bnd, "Sentiment analysis and opinion mining via microblogging in social media like: twitter," 2015.
- [18] G. M. Raza, "Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models," in *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, 2021.
- [19] S. R. Shah, "Opinion-mining on marglish and devanagari comments of youtube cookery channels using parametric and non-parametric learning models," *Big Data and Cognitive Computing*, vol. 4, no. 1, p. 3, 2020.
- [20] F. Rustam, "A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis," *Plos one*, vol. 16, no. 2, p. e0245909, 2021.
- [21] L. Alzubaidi, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, no. 1, pp. 1-74, 2021.
- [22] V.-H. Nhu, "Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms," *International journal of environmental research and public health*, vol. 17, no. 8, p. 2749, 2020.
- [23] L. Huiqing, "Effective use of data mining technologies on biological and clinical data," 2004.
- [24] L. K. Ramasamy, "Performance analysis of sentiments in Twitter dataset using SVM models.," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 11, no. 3, 2021.
- [25] P. N. Deepthi, "A review on bioinformatics using data mining techniques," *Journal of Physics: Conference Series*, vol. 1228, no. 1, p. 12023, 2019.
- [26] G. Hackeling, *Mastering Machine Learning with scikit-learn*, 2017.
- [27] J. R. Parker, "Rank and response combination from confusion matrix data," *Information fusion*, vol. 2, no. 2, pp. 113-120, 2001.
- [28] K. M. Ting, *Encyclopedia of Machine Learning and Data Mining, chap. Confusion Matrix, 260*, "Springer US, Boston, MA, 2017.