Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Predicting Consumer Sentiment from Social Media by Text Mining

Fatima Hassan Fadel [a] **,** ***Fatin Kadhim Nasser*** [b] **,** ***Suhad Faisal Behadili*** [c] *****

[a] *Departments of compute, college of science, university of Baghdad, Iraq. Email fatema.fadel1201@sc.Uobaghdad.edu.iq*

[b] *Departments of compute, college of science, university of Baghdad, Iraq. Email fatin.nasser1201@sc.Uobaghdad.edu.iq*

[c] *Departments of compute, college of science, university of Baghdad, Iraq. Email suhad.f@sc.Uobaghdad.edu.iq*

ARTICLE INFO

Article history:
Received: 27/07/2025
Rrevised form: 08/10/2025
Accepted : 12/10/2025
Available online: 30/12/2025

Keywords:

Feature Extraction,
Machine learning,
Sentiment Analysis,
Sentiment Classification,
Text Mining.

ABSTRACT

Currently, social media has expanded and disseminated, and a vast quantity of information is available to individuals of all ages and is being disseminated over the Internet. This information is not only vast, but also rapidly disseminated and diverse. Consequently, conventional tools and methodologies are inadequate for managing this information. Given the rapid advancement of this field, it is imperative to cultivate capabilities and investigate solutions that enable the extraction of specific values from data sets and their accurate analysis. Data analysis is one of these solutions. For example, the classification of emotions through data mining, which employs machine learning, involves the treatment of information that is transmitted through a variety of communication channels as emotions and the development of analytical models. Using three text mining and machine learning techniques: k-means, Decision Tree, and Classification and Regression Tree (CART), a significant amount of information was collected in this study and transmitted to individuals via Twitter from the Amazon website (comments). The results indicated that the utmost accuracy was achieved by employing two methods to extract properties from non-structural data and convert them into usable numerical representations. It is obtained at a rate of 95% through the use of Bag of Word feature extraction in conjunction with CART. So, it outperformed to Decision Tree, while K-means, which the desired outcome did not get.

MSC..

https://doi.org/ 10.29304/jqcsm.2025.17.42551

1. Introduction

In our current era, there has been a tremendous expansion in social media all over the world, which contains information available in a huge amount on the Internet, such as Facebook, Instagram and Twitter, users of these applications share their feelings about anything, whether it is opinions about a specific type of product or a service provided to them by a party Certain or watching movies. The computational study of sentiment analysis depends here on mining texts of people's attitudes, emotions, or opinions regarding a specific event. This event may represent individuals or topics, and these topics are covered through reviews (Martinovic, et al., 2018). In fact, computational analysis of these opinions and feelings is not a simple task because it is possible in a certain place to understand someone's opinion as positive, and it is possible in another place to understand this person's opinion as

∗Corresponding author: Fatima Hassan Fadel

Email addresses: fatema.fadel1201@sc.Uobaghdad.edu.iq

Communicated by 'sub etitor'

negative. In this case, there must be very specific circumstances for individuals because people's thinking differs from each other in addition to some. They immediately express their feelings while there are people who take time to express them (Alam, et al., 2025). Because reviews of people's thoughts include good and bad situations, it makes analyzing texts one by one difficult. Therefore, the most important problem is determining the validity of feelings in the comments posted by individuals (Medhat, et al., 2014) . On the other hand, For the application user, he understands phrases and comment sentences simply, but for the computer system, this is not an easy matter because traditional word processor algorithms cannot distinguish information if there is a slight difference in these texts. In addition, if there are changes in two sentences, sentiment analysis may affect the meaning; for example, **"The bell is not ringing"** is not the same as **"The bell is ringing"**. Additionally, Computers have certain limitations, making it essential to carefully consider the type of text being processed. The size of the vocabulary indicates the number of symbols or words that the model uses to understand and represent texts, as each symbol represents a word or part of it and is used in training and prediction processes (Younis, 2015) . Blogging platforms and means of communication such as Facebook and Twitter enable the user to use any language, whether formally or informally, so it was found necessary to find a natural language processor (Martinovic, et al., 2018). Natural language processing is a branch of artificial intelligence that helps computers interpret and understand human language, whether spoken or written. This processing is used in many applications such as Alexa and machine translation programs. Therefore, the subfield of Arabic language processing is sentiment analysis, which focuses on using computers to analyze texts and identify expressed feelings or opinions (Cai & Zhang , 2023). In the field of analyzing people's opinions, in addition to using machine learning algorithms, special language processing techniques must be used to obtain pure data, such as deleting repeated words or symbols that have no meaning, to analyze large data sets that have been recorded by people regarding a specific topic and play significant roles in the process of data and text mining. Present-day real-world problems are addressed through the application of data mining, which represents a relatively new field benefiting from a variety of innovative techniques. In this context, sentiment analysis focuses on identifying the sentiment conveyed within a given text and subsequently analyzing it. Sentiment analysis aims to reveal the opinions of individuals, determine their feelings that are reflected in them, and classify them whether they are positive or negative, through which it is possible to determine the best, as depicted in Figure 1 (Younis, 2015) , (Medhat, et al., 2014), (Fadel & Behadili, 2022).

The main goal of this work is to build a system that can classify every evaluation entered as expressing positive or negative feelings for a group of evaluations obtained from the Amazon website using three techniques in the field of machine learning, which are the K-means algorithm, which is used to classify data without supervision, and Decision Tree, it is classified based on data characteristics, in addition to CART which can be used to classify or predict numerical values. The construction of these models was based on two methods for extracting features from texts: Bag of Word and Term-Frequency-Inverse Data Frequency. Comparing these two methods' performance to determine which is more effective in improving the accuracy of the models. Organizational structure of the research. The second section is a presentation of previous work on the topic. The second section presents the research and methodological model. The fourth section the analysis and interpretation of model results is presented and the conclusions are in section five.

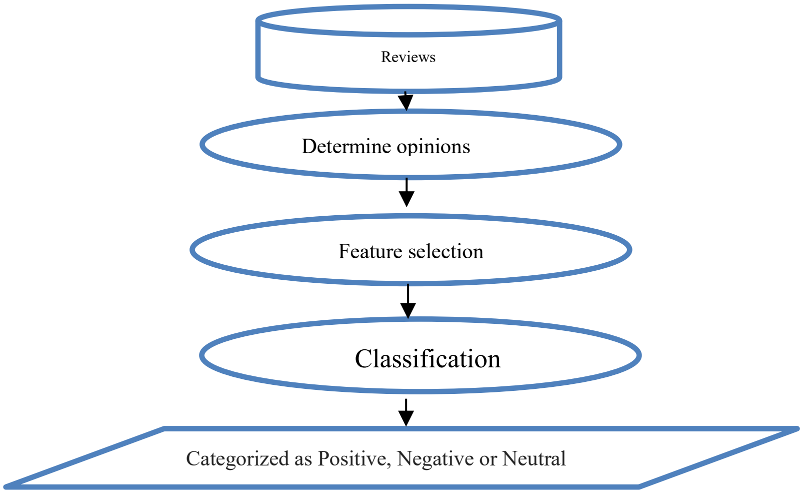


Fig.1 -Add flowchart to explain methdedology.

2. Literature Review

In the context of analyzing opinions, many studies have been conducted, starting from the use of grammar methods to machine learning techniques. The two main directions of studying sentiment analysis are the document and sentence levels, as these two systems focus on discovering words and phrases that include viewpoints or feelings. In (Riaz, et al., 2019)Sentiment analysis was conducted on real-world customer review data at the phrase level to identify customer preferences by examining subjective expressions. The researchers measured the strength of sentiment words to determine the intensity of each expression. They then applied k-means clustering to group the words into various clusters based on their intensity levels The results of the same data were compared with the star rating, and better and neutral feelings towards the products were found through analyzing the text of the references, where a group of references was found that carried high negative feelings while others carried positive feelings. This work helped new customers to use the products for the first time and also made it easier for those in charge of the work to know how they like them. Customers use the product to increase profits.

Researchers in (Araque, et al., 2019) proposed a method using dictionaries that are based on the evidential similarity between dictionary words and opinion terms. Opinion-based characteristics and implication were examined using semantic probability based. The method was performed on seven sets of general data and four dictionaries of emotions. Through several statistical methods, features were extracted, where the goal was to integrate Feature extraction with existing representation.

Authors of another article (Alatabi & Abbas, 2020) Two different datasets were used: the first related to comments on Facebook, where 4,000 comments were classified into 2,000 positive and 2,000 negatives. The second relates to movie reviews, where 2,000 reviews were classified into 1,000 positive and 1,000 negatives. In this study, the Bayesian Rough Decision Tree algorithm was used to develop an emotion analysis system aimed at classifying texts into positive or negative emotions. Experimental results demonstrated the success of the system, with classification accuracy of over 95% when applied to social media data.

In (Nalini, et al., 2021) the researchers found that data mining leads to better knowledge of the problems that can occur in sailors' medical documents. Centro International Radio Medico collected digital medical data, an Italian system for marine medical aid. Patient data was for the period from 2018 to 2020, and naive Bayes technology was adopted to conduct analysis and carry out experiments. On the R statistical tool, the correlation between medical problems and diagnosis was found through verbal withdrawal at 96% the accuracy of the sentiment analysis was more than 80%. While in (Rustam, et al., 2021) the data used in this study was taken from the IEEE Data Port platform on May 31, 2020. The collection includes Tweet IDs and Sentiment Scores for 7,528 tweets. Two language processing techniques for tweets, word bag and TF-IDF, were combined to improve text representation. Tweets were classified into positive, negative, and neutral categories to understand the different emotions expressed by tweets. To conduct advanced analysis, the "long- and short-term memory" architecture of neural networks was applied deep. This architecture is characterized by its ability to process time sequences and capture long dependencies in the data, which enhances the accuracy of emotion classification and has an accuracy of 93%.

In (Fadel & Behadili, 2022) comparative study was conducted utilizing Twitter data containing reviews related to Oklahoma cuisine, which were sourced from the Amazon website. The supervised learning classifiers: **Naive Bayes**, **Logistic Regression**, and **Support Vector Machine (SVM)**, used in this study, two methods were used to extract the characteristics: the count vectorizer and Term-Frequency-Inverse Data Frequency .The results obtained showed that the highest classification accuracy was 91% using the SVM method using the count vectorizer, but it was more expensive from a mathematical standpoint in terms of time consumption, while the Logistic Regression was more Balanced between accuracy and runtime performance.

In this work, a decision tree algorithm is developed using the CART algorithm. The post-processed text data, collected from Twitter, underwent efficient feature extraction, resulting in high accuracy.

3. materials and methods

In this study, decision tree-based classification algorithms, including the traditional decision tree algorithm and the iterative decision tree algorithm (CART), were applied to classify data. The procedures begin by performing the tasks of initializing and arranging the entered data set, followed by the feature extraction stage, where texts are represented using custom techniques and converted into numerical values. After that, the aforementioned classification algorithms are applied to classify sentences into specific categories. In addition, the K-means clustering method is used to group data based on similarities between texts.

3.1. Data Collection Summary

The dataset utilized was sourced from the Kaggle repository "Amazon Fine Food Reviews," provided by the Stanford Network Analysis Project "https://www.kaggle.com/snap/amazon-fine-food-reviews ". It included the details shown in Table 1 ,reviews were categorized based on their scores: Positive for reviews with scores of 4 or 5.Negative for reviews with scores of 1 or 2,while reviews with a score of 3 was normal and was excluded from the analysis due to its ambiguity, and it was divided into two subsets: a training set and a test set, facilitating the development and evaluation of predictive models. Table 2 illustrates the format of the data entries.

Table 1- Dataset details.

Data set statistics	Number of records
Appraisal	568,454
Participants	256,059
Items	74258
Active Reviewers	260
Median no. of words per review	56

Table 2 - Attribute Information of dataset.

Markers	Description
Identifier	Sequence
Product Code	Product Code
Account Number	unique identifier for the user
Display Name	Name of user profile
Helpful Indicator Count	Count of Helpful Votes
Total Helpfulness Responses	**Helpful Votes Total**
Assessment	1–5 Rating Scale
Period	Review Date and Time
Outline	Brief Digest
Document	**Review Content**

3.2. Data Preparation Phase

Processing data is necessary to achieve more effectiveness in machine learning projects. This is because different social media are used naturally for people's informal languages and terms in other words, in their natural expression. So, Twitter tweets, Facebook comments, and other sentences are incomplete and collected in raw form, producing a noisy total of data (Alam, et al., 2025). Therefore, the raw Twitter data need to be considered to build a set of data that can be trained by diverse group of manufacturers. In addition, preprocessing leads to the reduction and consolidation of data aggregation. All these aspects are under the principle of Natural Language Processing. In this case, explains that these processing methods in the field of analysis focus on deriving features such as repeated words from text data. It must be emphasized that machines are limited, restricted, and receive orders from their personnel, such as their applications, microblogging sites as Twitter, Facebook, which are used in a closed-door manner, depending on the nature of the individuals (Cichosz, 2024). For example, the following phrase **" food tastes very good**" is the result of an arrangement of this phrase **"this food is tastes too very good!**"

3.3. Feature Representation

The technique of converting text into numerical feature vectors for use by Machine Learning Algorithms is known as text feature extraction. Machine Learning Algorithms cannot be directly fed text documents or corpora since these algorithms require data input in the form of numerical feature vectors or matrices. There are several ways of converting text into a feature vector in this article uses Bag of Words and Term-Frequency-Inverse Data Frequency Weighting. Therefore, a good knowledge of feature extraction techniques is considered a necessary stage for representing texts for use in the advanced stages and employing them effectively (Nalini, et al., 2021).

3.3.1. Bag of Words

Characters and words are incomprehensible to machines. So, dealing with text data must be expressed numerically in order that the computer can understand it. Bag of Words is a Vectorization strategy that involves tokenization, frequency counting, and normalization. Alternatively, it converts text into digital data, where an equal vector is created to the volume of vocabulary. Every time a word occurs in a sentence, it is inspected in the vocabulary and

labeled as 1. If the word is repeated in the text, the count gradually increases. If it is not in the vocabulary, it is added to the vector. This example describes the work of these technologies in detail (Zhang, et al., 2020).
Original sentence: **(Excellent food and cost and excellent service)**
after indexing the words, they will be as in Table 3 and the sentence after vectorization is expressed as in Table 4.

Table 3 - Indexing the words.

And	Excellent	Food	Service	cost
0	1	2	3	4

Table 4 - Sentence after vactorization.

And	Excellent	Food	Service	cost
2	2	1	1	1

3.3.2. Term-Frequency-Inverse Data Frequency

Text analysis tasks employ the Term-Frequency-Inverse Data Frequency (TF-IDF) feature extraction Method. When determining the frequency of each token, TF-IDF weight in guarantees that the whole corpus is taken into account. TF-IDF weighting can be added to each token to deal with the problem of frequent, uninformative tokens against. infrequent, informative tokens. If a term appears too often in a text or corpus, it is unlikely to provide much value to the study. The TF-IDF weighting system reduces the importance of phrases that appear frequently and raises the importance of terms that infrequently occur (Alqaryouti, et al., 2020). In other words, it provides weighted characteristics to improve performance. TF-IDF uses the product of term frequency (TF) and inverse document frequency (IDF) to determine the weight of each feature in a document. The frequency of a feature in a document is determined by the document’s length. Equation 1 can be used to define it.

$$tf_{t,d=\frac{count_{t,d}}{total\ count_{d}}} \qquad (1)$$

Where $count_{(t,d)}$ is the number of terms t in the document $d$and total count$_{\boldsymbol{d}}$ is the total number of all terms in the document d. IDF measures the extent of a term $t$being informative in a document for model training. IDF can be computed as in Equation 2.

$$idf = \frac{N}{DF_t} \qquad (2)$$

Where N is the number of documents in the corpus and DF_t is the number of documents that contain the term t. IDF measures the weight of a term t low when term t frequently occurs in many documents. For instance, stop words have low IDF value. Finally, TF -IDF can be defined in Equation 3 (Alqaryouti, et al., 2020).

$$tf - idf = tf_{t,d} * log(idf) \qquad (3)$$

For more clarification the following example: defines the corpus, which is a list of comma-separated text documents.
Corpus=

”Do you like Chicken and Meat?”
”I do not like them, Ali-I-am.”
”I do not like Chicken and Meat!”
”would you like them here or there.”
”I would not like them here or there.”

Get alist of all the tokens/features found over the entire corpus: am’,’and’,’do’,’ Chicken’,’ Meat’,’here’,’like’,’not’,’or’,’ Ali’,’them’,’there’,’would’,’you’.
Obtain mapping of tokens to feature ID in the feature vector:
’am’=0, ’and’=1, ’do’=2, ’ Chicken’=3 , ’ Meat’=5, ’here’=6, ’like’=7, ’not’=8, ’or’=9, ’ Ali’=11, ’them’=11, ’there’=12, ’would’=13, ’you’=14.
All the techniques and tools that could help accomplish the research and lead to good results were mentioned in an area of the conducted work search.

3.4. Performance Evaluation Measurements

The N-Fold Cross-Validation technique is a popular statistical technique for evaluating the performance of classification models. In this technique, the original data set is randomly divided into N subsets (called "folds") of similar sizes. Training and testing are performed N times so that each piece of data is used as test data once, while the rest of the pieces are used as training data. The accuracy of the model is calculated each time, and this accuracy is then arithmetic averaged across all folds to obtain a reliable estimate of the model's performance on previously invisible data. The general accuracy of the model expressed using the arithmetic mean of the individual measurements of each fold, as shown in equation4.

$$CVA = \frac{1}{n}\sum_{1}^{n} A_i \qquad (4)$$

Where CVA stands for Cross Validation Accuracy, n is the number of folds, and A is the accuracy measure for each fold (Alqaryouti, et al., 2020). Suppose n=4 so, the training set constitutes 60% of the total dataset, with the remaining portion allocated for testing, dataset is partitioned into two subsets: the training set and the testing set. The training set comprises 60% of the total dataset and is utilized to train the model, enabling it to learn underlying patterns and relationships within the data. The remaining 40% of the dataset constitutes the testing set, which is employed to evaluate the model's performance and generalization capability on unseen data. This division ensures that the model is assessed on its ability to make accurate predictions on new, previously unobserved example. The fundamental criterion for performance assessment is the confusion matrix. Table 5 shows a confusion matrix for a two class classification problem. The **True Positive** and **True Negative** are true classifications for each class. **False Positive** means a class B sample is predicted as a class A sample, while a **False Negative** means a class A sample is wrongly assigned as a class B sample (.M & Reddy, 2011) (Srujan, et al., 2018) .

Table 5 - Confusion matrix for a two class.

Class	A predicted	B predicted
A factual	True positive	False negative
B factual	False positive	True negative

3.5. The classification Phase

It is significant purpose is to determine the identical properties between the points in the dataset, then use the classification model to classify the points into separate classes (Srujan, et al., 2018). In supervised learning algorithms, the parameters are unknown and are not determined from the beginning. Therefore, it is important to provide a set of training data to determine these unknown parameters with the structure of the algorithm so that it provides expectations based on the set of training data that it learns from the parameters with the input data (Alzubaidi, et al., 2021). Thus, an easy example to understand this task is possible to classify animal species using physical features into two categories: 'mammals' and 'birds', according to features such as weight, length, number of legs, and presence of feathers. Classification tasks in machine learning according to the problem domain can be Bilateral classification: This involves classifying data into two distinct categories. Multi-Class Classification: Assigns each instance to one of three categories such as classifying animal images into categories such as "cat" "dog" or "rabbit".Multi-tag classification: where each instance belongs to multiple classes at once. Such as classifying his essay as "social" and "economic". Unbalanced classification: occurs when certain categories are underrepresented. To train the learning models it is necessary to divide the input data set into three subsets: **Training**: used to train the model, allowing it to learn basic patterns in the data. **Verification group**: used to adjust model parameters and make decisions about model configurations. **Testing**: provides an unbiased assessment of the performance of the final model, ensuring that it generalizes well to unseen data These groups are useful for knowing the speed of the model used in finding expected patterns and the efficiency of the model used (Hossain, et al., 2021) (Anisur, 2019) , Figure 2 **presents** the input dataset partitioning strategy. Therefore, Supervised learning offers greater adaptability compared to unsupervised learning, making it more suitable for predicting outcomes in test datasets that share similar distributional assumptions with the training data (Chen, et al., 2015) . The initial algorithm utilized in this study is the Decision Tree. This classification algorithm operates by systematically exploring the training samples to identify the optimal separation node (or characteristic), which forms the foundation for categorizing different types of decision trees.

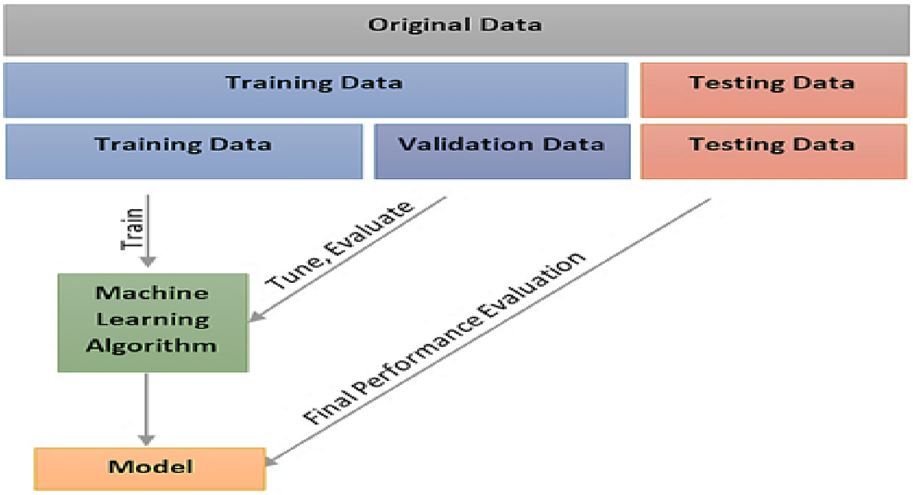


Fig.2 -Input dataset partitioning process.

Similarly**,** Classification and Regression Tree (CART) (uses gini gain) **.** The CART approach constructs a binary tree, where each internal node denotes a condition on a feature. Each of the two branches corresponds a conditional outcome (true and false), and each leaf node denotes a class label. Let C be the class attribute with $(c1, c2, \ldots, cn)$ values, n the number of classes, and A represent the feature which can be used to split training dataset D into K subsets, such that $(d1, d2, \ldots, dn)$. These partitions mean the branches of the node. Thus, the gini index is determined using Equation 5 (Polaka & Borisov, 2010):

$$gini(c) = 1 - \sum_{i=1}^{n} p(c_i) \quad \mathit{(5)}$$

Whereas, $P(ci)$ is the relative frequency of class value . Gini split information, which measures the $gini$ index for all feature values, is determined according to Equation 6 (Polaka & Borisov, 2010) **:**

$$gini_{split(A)} = \sum_{j=1}^{k} p(d_j)gini(d_j) \qquad (6)$$

Where j represents the $j-\ th$ feature value, for **CART**, $j =\ (1{,}2)$ in a binary split. The minimum $gini$ index mean the maximum impurity reduction which is denoted as gini gain, that is calculated by Equation 7 (Polaka & Borisov, 2010):

$$\boldsymbol{ginigain = gini(c) - gini_{spilt(A)}} \; . \; . \; . \; . \; .7$$

The CART method is improved by calculating the $gini$ index for just the values lying between two consecutive sorted values with distinct classes as separation points. As a consequence, the computations are simplified. The split point of a continuous feature will be the location with the lowest $gini$ index. The $gini$ index is a measure of the dataset impurity, if all instances of a subset have the same class label, the $gini$ index is 0. The $gini$, on the other hand, will be 1 if each instance has a distinct class (Khan, 2023)**.**

3.6. Clustering Phase

Clustering is an unsupervised learning technique aimed at grouping data points into clusters, ensuring that items within the same group are more similar to each other than to those in different groups. This method assigns each data point to a specific cluster, facilitating the identification of patterns and structures within the data. In our study, we employed the k-means clustering algorithm to partition the dataset into distinct clusters based on feature similarity (Dey, et al., 2016), K-means method is one of the simplest and most widely used clustering techniques, as it finds cluster centers in specific areas of data in two steps, the first is to assign each data point to the nearest center in the cluster, and then each cluster center assigns the data points assigned to it and completes when there is no mapping of instances to clusters (Hossain, et al., 2019). Figure 3 shows the algorithm on a virtual data set as an illustrative example, where the data points appear as circles and the group centers appear as '+', then select search for Three groups. Therefore, the algorithm was initialized by randomly searching for three data points as cluster centers, as shown in Figure 3-a. After that, each data point is assigned to the nearest cluster center based on the Euclidean distance, as shown in Figure 3-b.After assigning to the appropriate clusters, the cluster centers are updated, where the new center for each cluster is calculated as an average of all the points that were assigned, as shown in Figure 3-c,it continues to be repeated in these two steps until stability occurs, meaning that the mapping of points to clusters and clustering centers does not change between iterations so that the algorithm reaches a state of stability and stops (Alfina, et al., 2017) (Bayhaqy, et al., 2018).

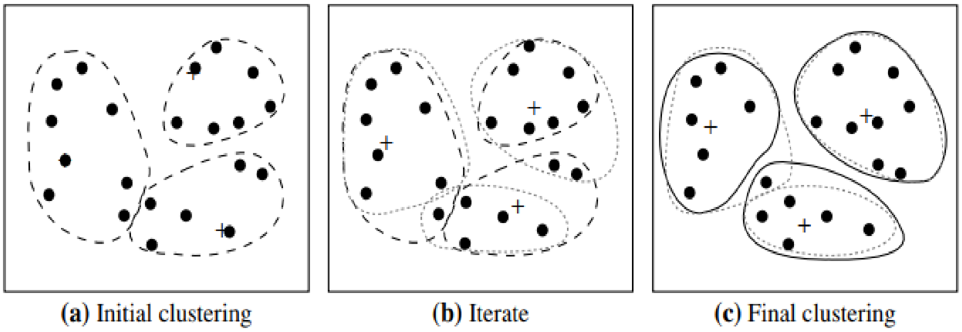


Fig.3 - Clustering of a set of objects using the k-means method.

4. Results and discussion

In this section evaluates the effort and work of the methods used to implement for SA the proposed algorithms were tested and evaluated. Trying hard to make the results as best as possible, and to display all the results from the stages as much as possible. Firstly it shows the results of the pre-processing stages until configuration for training and test files. Finally, present ts the results DM models' were explained for the feature selection models Bag of words and TF-IDF results. Similar, the metrics for the experimental results for models evaluation for each class's samples such as (precision, recall, F1-score) are described in tables. The work is based on a comparative study of the methods used our experiment was performed using processor intel core $i7 - 8^{th}$ Gen, Quad-Core 1.8 GH_z Clock speed, the installed memory of 8 GB RAM, and the system type is the 64-bit operating system,x64 based processor.

4.1. Preprocessing Results

Initial preparation of data in any analysis is necessary, especially for dealing with social media content. The data set for this work contained non-textual elements such as emoji, punctuation marks, repetition of letters and words, in addition to spelling errors. Therefore, it was initially processed to convert it into an analyzable form, where only the content of the reviews was extracted. With its negative and positive ratings, removing information and recurring reviews, the following processing steps were carried out:

1- To unify writing, the entire text was converted into small letters.

2- Remove unnecessary punctuation marks, taking into account that some marks may carry an emotional connotation.

3-Correcting spelling errors .

4- Remove common words that do not carry a strong semantic meaning, such as in on,

here the data is transformed into a unified format to ensure the accuracy and effectiveness of the analytical models.

4.2. Experimental Results for Techniques used.

In the conducted work, classified tweets into two class as positive and negative comments as shown in Figure 4. The nature of the data selected for people’s tweets about the product offered tends to be positive, which achieved the results mentioned below.

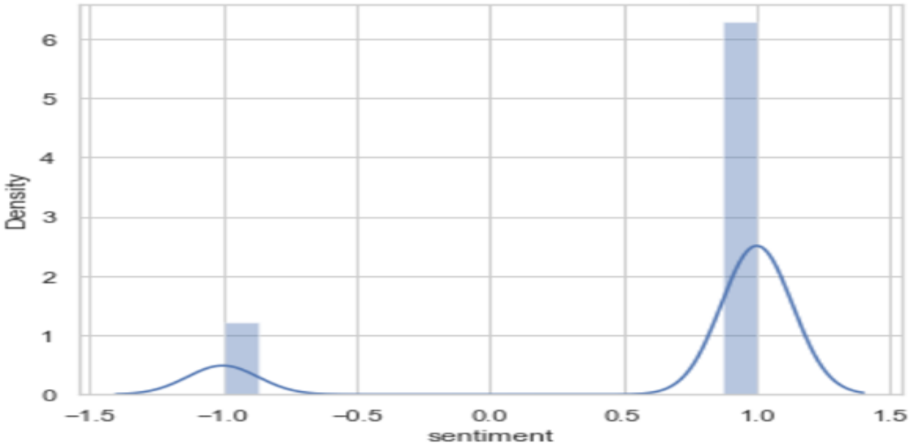


Fig.4 - Classified sentiment classes as negative and positive one.

4.2.1. Comparing results among decision-making methods

The level of understanding of decision tree algorithms is very easy compared to other classification algorithms where they try to solve the problem using decision tree. Each internal tree contract corresponds to its noun and each circumstantial node corresponds to it is noun, Where the sum of measurements obtained is shown in Table 6 by using Bag of word feature extraction for Decision Tree and in Table 7 by using TF-IDF feature extraction for Decision Tree According to the previously mentioned standards, either confusion matrix for Decision Tree as shown in Figures 5(a) and 5(b), Which explains that high accuracy was by TF-IDF feature extraction as shown in Figure 6.

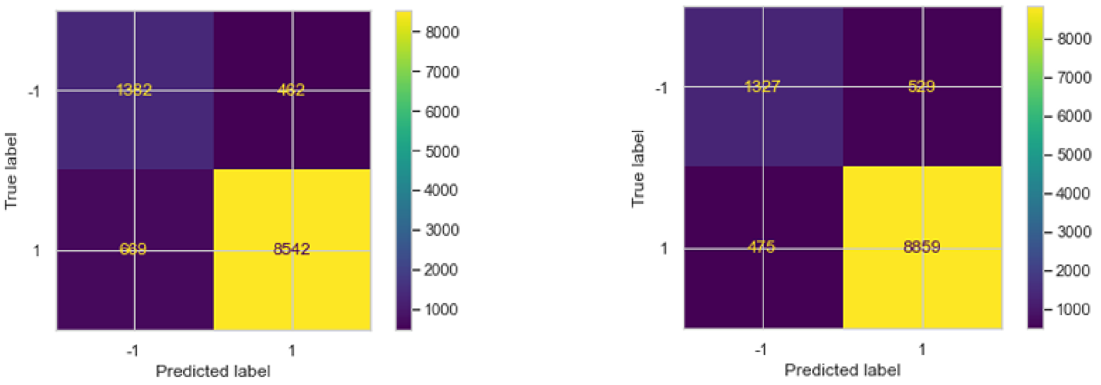
Table 6 - Classification report of Bag of word for Decision Tree.

Classes	Precision	Recall	F1-score	support
Negative	0.67	0.75	0.71	1844
Positive	0.95	0.93	0.94	9211
Accuracy			0.90	11055
Macro avg	0.81	0.84	0.82	11055
Weighted avg	0.90	0.90	0.90	11055

Table 7 - Classification report of TF-IDF for Decision Tree.

Classes	Precision	Recall	F1-score	Support
Negative	0.74	0.71	0.73	1856
Positive	0.94	0.95	0.95	9334
Accuracy			0.91	11190
Macro avg	0.84	0.83	0.84	11190
Weighted avg	0.91	0.91	0.91	11190

The primary challenge is to apply the decision tree to identify the features that need to be considered. As the root contract and each level of the tree takes into consideration a measure of choice for the best poison is which is used by CART. The sum of measurements obtained is shown in Table 8 by using Bag of word feature extraction.



(a) **(b)**

Fig.5- Confusion matrix of DT by (a) the Bag of word (b) the TF-IDF.

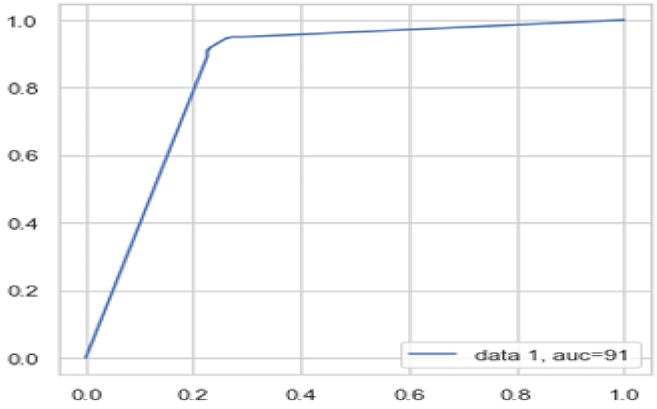


Fig.6 - Accuracy of DT by the TF-IDF.

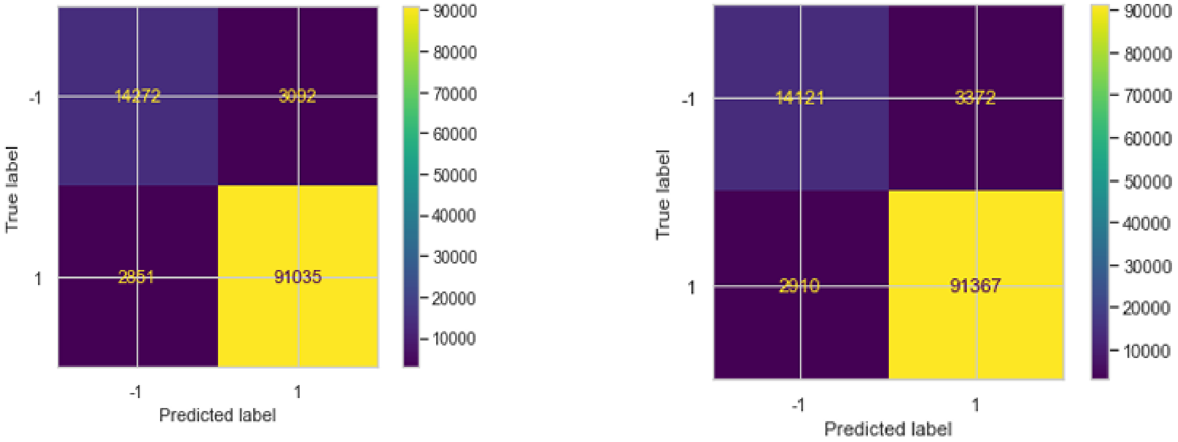
and in Table 9 by using TF-IDF feature extraction. Confusion matrix for CART is shown in Figure 7(a) and Figure 7(b). So, the high accuracy is obtained by CART as shown in Figure 8 by using bag of word feature extraction and the time spent running out was between 5 and 6 seconds. System process by using training-testing by 80% from entire dataset and 20% for test,and high accuracy result by CART. Because of the choice of the most efficient characteristics. In addition to, this is due to the unbalanced data and the system's bias towards positive closures.

Table 8 - Classification report of Bag of word for CART.

Classes	Precision	Recall	F1-score	Support
Negative	0.82	0.83	0.83	17123
Positive	0.97	0.97	0.97	94127
Accuracy			0.95	11125
Macro avg	0.90	0.90	0.90	11125
Weighted avg	0.95	0.95	0.95	11127

Table 9 - Classification report of TF-IDF for CART.

Classes	**Precision**	**Recall**	**F1-score**	**Support**
Negative	0.83	0.81	0.82	17493
Positive	0.96	0.97	0.97	94277
Accuracy			0.94	111770
Macro avg	0.90	0.89	0.89	11177
Weighted avg	0.94	0.94	0.94	111775



(a) **(b)**

Fig.7- Confusion matrix of CART by (a) the Bag of word (b) the TF-IDF.

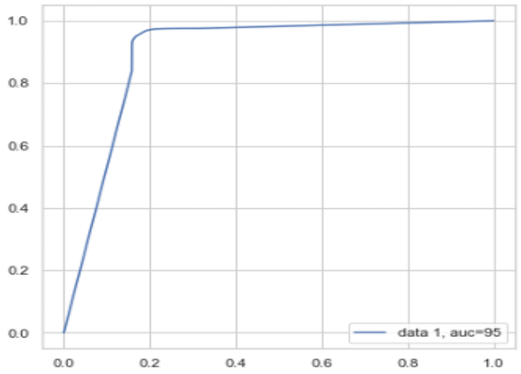


Fig.8 - Accuracy for CART by the Bag of word.

4.2.2. Experimental Results for K-means.

K-means clustering has been applied on 45952 instances from total data collected after feature extraction step has been done .The k-means clustering consists of option for specify the number of clusters, which obtain ten after several attempts, and the most in retesting parameter was comment, summary for each person in data set, To ensure robustness in result aggregation, scikit-learn performs the k-means algorithm multiple times using different random initializations. This approach takes advantage of the fact that the number of clusters remains constant and is not affected by the initialization phases, thus enhancing the stability and reliability of the clustering results. The obtained results are shown in Table 10 for TF-IDF feature extraction and Table 11 for Bag of word. The random configuration of k-means makes its accuracy less effective, which means that the result of its implementation depends on a random seed. In addition, other aspects of k-means are the relative assumptions that are made in the form of groups and the requirement to specify the number of groups you are looking for, which may be unknown in real applications. In addition, in our area of research and the use of available data, the data need to be classified into specific categories to identify it and define a person's opinion more clearly; especially. Those resulting groups were ignored as shown by confusion matrix in figures 9(a) and 9(b) .it reaches almost identical totals.

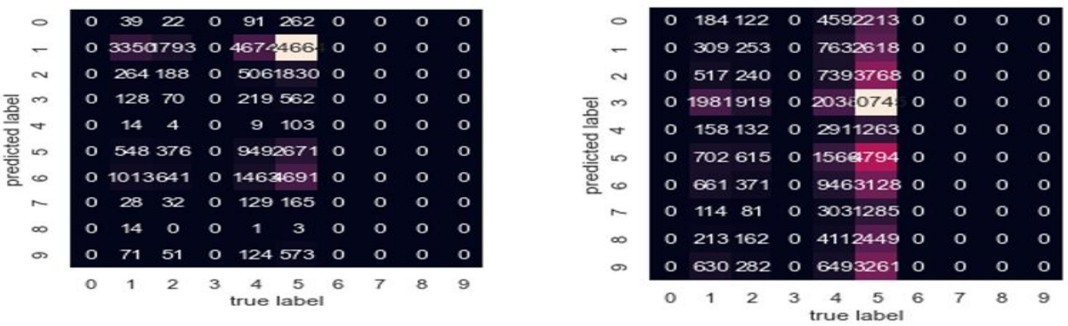
Table 10 - Evaluate report of K-means for TF-IDF.

Cluster	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	0
1	0.06	0.02	0.03	4721
2	0.08	0.14	0.10	2814
3	0.00	0.00	0.00	0
4	0.15	0.48	0.23	7288
5	0.70	0.08	0.14	31130
6	0.00	0.00	0.00	0
7	0.00	0.00	0.00	0
8	0.00	0.00	0.00	0
9	0.00	0.00	0.00	0
Accuracy			0.14	45953
macro	0.10	0.07	0.05	45953
weighted	0.51	0.14	0.14	45953

Table 11 - Evaluate report of K-means for Bag of word.

Cluster	Precision	Recall	F1-score	Support
0	0.00	0.00	0.00	0
1	0.10	0.61	0.17	5469
2	0.07	0.06	0.06	3177
3	0.00	0.00	0.00	0
4	0.07	0.00	0.00	8165
5	0.59	0.06	0.13	35524
6	0.00	0.00	0.00	0
7	0.00	0.00	0.00	0
8	0.00	0.00	0.00	0
9	0.00	0.00	0.00	0
Accuracy			0.12	52335
macro	0.08	0.07	0.04	52335
weighted	0.42	0.12	0.11	52335

The fact that the algorithm's results are weak is due to its reliance on random initial values with the selection of an inappropriate number, and also the abnormal data has an impact, which reduces the accuracy of the results.



(a) (b)

Fig.9 -confusion matrix of k-means by (a) Bag of word (b) TF-IDF.

5.Conclusion

Recently, sentiment analysis technology has witnessed wide spread because of the analysis capabilities it provides that help support people and institutions in making more accurate decisions. It is considered an important tool for understanding customer opinions and assessing the extent of their satisfaction with products or services, which helps institutions improve their performance and develop their products. As for individuals, it will It provides an effective way to discover a person's impression about a particular topic, which leads to saving time and effort and improving the quality of decisions based on these analyses. As a summary of this work, a sentiment analysis model was presented that aims to determine the positive or negative text of a group of reviews about the quality of food obtained from the Amazon website. Creating this model includes mining text data from useless elements and errors and converting it into a form that can be understood by machine learning algorithms. Using feature extraction methods, it is converted into a numerical representation, which contributes to improving performance. Three machine-learning classifiers were involved in this work; Decision tree and CART classifiers. The CART classifier received the highest accuracy of 95% by selecting the feature selection (Bag of words). In fact, if both accuracy and performance are taken into account as two focuses. As for the unsupervised learning algorithm K-means, no satisfactory results were obtained.

References

[1] Alam, S., Mrida, S. H. & Rahman, A., 2025. SENTIMENT ANALYSIS IN SOCIAL MEDIA: HOW DATA SCIENCE IMPACTS PUBLIC OPINION KNOWLEDGE INTEGRATES NATURAL LANGUAGE PROCESSING (NLP) WITH ARTIFICIAL INTELLIGENCE (AI). *American Journal of Scholarly Research and Innovation,* 04(01), pp. 63-100.

[2] Alfina, I., Sigmawaty, D., Nurhidayati, F. & Hidayanto, A. N., 2017. Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain. *Proceedings of the 9th international conference on machine learning and computing,* pp. 43-47.

[3] Alqaryouti, O., Siyam, N., Monem, A. A. & Shaalan, K., 2020. Aspect-based sentiment analysis. *Applied Computing and Informatics,* 20(2), pp. 142-161.

[4] Alzubaidi, L. et al., 2021. Review of deep learning: concepts, CNN. *Journal of big Data,* 8(1), p. 53.

[5] Araque, O., Zhu, G. & Iglesias, C. A., 2019. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems,* Volume 165, pp. 346-359.

[6] Bayhaqy, A., Sfenrianto, S., Nainggolan, K. & Kaburuan, E. R., 2018. Sentiment Analysis about E-Commerce from. *nternational conference on orange technologies (ICOT),* pp. 1-6.

[7] Chen, F. et al., 2015. Data Mining for the Internet of Things:. *International Journal of Distributed Sensor Networks,* p. 14.

[8] Cichosz, . P., 2024. Active Learning for Biomedical Article Classification with Bag of Words and FastText Embeddings. *Applied Sciences,* 14(17), p. 7945.

[9] Dey, L. et al., 2016. Sentiment Analysis of Review Datasets using. *arXiv preprint arXiv.*

[10] Fadel, F. H. & Behadili, S. F., 2022. A Comparative Study for Supervised Learning Algorithms to Analyze Sentiment Tweets. *Iraqi Journal of Science,* 63(6), pp. 2712-2724.

[11] Hassan, F. and Behadili, S. F., 2022. Modeling social networks using data mining approaches – review. Iraqi Journal of Science, 63(4), pp. 1313–1338

[12] Hossain, A., Karimuzzaman, M., Moyazzem Hossain, M. & Rahman, A., 2021. Text Mining and Sentiment Analysis of Newspaper Headlines. *Information,* 12(10), p. 414.

[13] Hossain, M. Z., Akhtar, M. N., Ahmad, R. B. & Rahman, M., 2019. A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical engineering and computer science,* 13(2), pp. 521-526.

[14] Khan, M. Z., 2023. AN INSIGHT ON MACHINE LEARNING ALGORITHMS AND. *European Chemical Bulletin,* pp. 6029-6034.

[15] Medhat, W., Yousef, A. H. & Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal,* 5(4), pp. 1093-1113.

[16] Polaka, I. & Borisov, A., 2010. CLUSTERING-BASED DECISION TREE CLASSIFIER CONSTRUCTION. *Technological and economic development OF ECONOMY,* 16(4), p. 765–781.

[17] Riaz, S., Fatima, M., Kamran, M. & Nisar, M. W., 2019. Opinion mining on large scale data using sentiment analysis. *Cluster Computing,* 22(3), pp. 7149-7164.

[18] Rustam, F. et al., 2021. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *Plos one,* 16(2), p. 0245909.

[19] Younis, E. M., 2015. Sentiment Analysis and Text Mining for Social Media. *International Journal of Computer Applications,* 112(5), p. 0975 – 8887.

[20] .M, V. & Reddy, L. C., 2011. A Review on Data mining from Past to the Future. *International Journal of Computer Applications,* 15(7), p. (0975 – 8887.

[21] Alatabi, H. & Abbas, . A., 2020. Sentiment Analysis in Social Media using Machine Learning Techniques. *Iraqi Journal of Science,* 61(1), pp. 193-201.

[22] Anisur, R., 2019. Rare sequential pattern mining of critical infrastructure control logs for anomaly detection. *Doctoral dissertation, Queensland University of Technology.*

[23] Cai , T. & Zhang , X., 2023. Imbalanced Text Sentiment Classification Based on Multi-Channel BLTCN-BLSTM Self-Attention. *Sensors,* 23(4), pp. 1-15.

[24] Martinovic, M., Domovic, J. & Piric, V., 2018. PERCEPTION OF HEALTH PRODUCTS AND TRENDS IN ONLINE SHOPPING AND PROMOTION OF HEALTH PRODUCTS IN CROATIA. *Вісник Національної академії керівних кадрів культури і мистецтв,* Issue 1, pp. 1446-1452.

[25] Nalini, C. et al., 2021. Text mining with sentiment analysis on seafarers’ medical documents. *International Journal of Information Management Data,* 1(1).

[26] Srujan, K. S. et al., 2018. Classification of Amazon Book Reviews Based. *Information Systems Design and Intelligent Applications,* pp. 401-411.

[27] Zhang, H., Sun, S., Hu, Y. & Liu, u., 2020. Sentiment classification for Chinese text based on interactive multitask learning. *IEEE Access,* Volume 8, pp. 129626-129635.