# Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets

**Murtadha B. Ressan, Rehab F. Hassan**
Department of Computer science, University of Technology, Baghdad, Iraq

**Article Info**

***Article history:***

Received Dec 4, 2021
Revised Jun 30, 2022
Accepted Jul 27, 2022

***Keywords:***

Bernoulli Naïve Bayes
COVID-19
Multinomial Naïve Bayes
Naïve-Bayes
Sentiment analysis Twitter

**ABSTRACT**

This paper proposes a system to analyze the sentiments of tweeters. It is to build an accurate model to detect different emotions in a tweet. The analysis takes place through several stages (i.e., pre-processing, feature extraction, and training more than one machine learning (ML)). Naïve Bayes, Multinomial Naïve Bayes and Bernoulli Naïve Bayes were selected as supervised machine learning for sentiment analysis using a dataset of 3,057 tweets with users ranging from fear to happiness, anger, and sadness because this method is suitable for solving a problem of this type. This system was also applied to another dataset of 10,000 Tweets (5,000 positive and 5,000 negatives). This approach, consisting of three Naïve Bayes classification models, was applied to two datasets to analyze the sentiment used in them and classify each category separately. The Multinomial Naïve Bayes model outperformed the other models Where it achieved an accuracy of (91.6%) when applied to the first dataset and accuracy (87.6%) when applied to the second dataset. The researchers aim to continue this research with larger data by using other methods of sentiment analysis to predict users' thoughts about COVID-19 or any other problem and to obtain higher accuracy for the models used.

***Corresponding Author:***

Murtadha B. Ressan
Department of Computer science, University of Technology
Baghdad, Iraq
Email: cs.19.02@grad.uotechnology.edu.iq

## 1. INTRODUCTION

The World Health Organization (WHO) declared COVID-19 a global pandemic on January 30, 2020 [1]. It is considered one of the most widespread, influential, and dangerous epidemics in global health history, as it causes a severe disease that sometimes leads to death [2]. Nowadays, people can share their valuable information via powerful social media like Facebook, Twitter, and other social networking platforms. During the pandemic period, people mainly shared their experiences, thoughts, and opinions on Twitter. Twitter is a popular social networking platform with a large number of users, i.e. more than 500 million users worldwide. Twitter is a primary source of health-related information due to the diversity of information shared by individuals and official bodies [3] so that this information can be fruitfully used to study people's behavior and analyze their interaction with any therapist addressed [2], [4]. Because of this pandemic, many businesses were disrupted and many workers lost their jobs, while the economy of some industrial aspects such as the pharmaceutical industry and health protection tools recovered, this research paper discusses the impact of these repercussions by identifying and analyzing their tweets to know their feelings, as well as discussing opinion mining for randomly collected tweets It will also be detailed later. The purpose of this research is to know the feelings of the tweeters during the pandemic period and to classify the tweets to obtain a reliable and high-accuracy approach that can be adopted as a predictive approach to help find solutions to such problems [5].

The dataset used in this paper is more than 3,000 tweets taken from kaggle.com, and the different interactions of the tweets are categorized as joy, fear, anger, and sadness. Another dataset of 10,000 tweets (5,000 positives and 5,000 negatives) was also used to classify tweets, both datasets are tweets and Twitter comments in English. Supervised learning techniques such as the Naïve Bayes group's methods for sentiment classification and analysis are used. The Naïve Bayes group (normal Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes) was selected based on the type of problem to be solved (sentiment analysis, classification, opinion mining) in proportion to the type of dataset where Naïve Bayes achieved good results and a high accuracy rate in the classification of feelings in previous studies similar to this study.

For each dataset, these steps are applied separately, classification accuracy results are recorded and restricted separately, and results tables will be displayed in the Results and Discussion pane. The dataset progress through several stages starting from the processing process (most datasets are it contains a lot of noise and useless components that affect badly the results of the analysis) The preprocessing stage begins with the step of cleaning the data, tokenization, and removing stop words, then setting the part of speech and returning each word to its origin (lemmatization) after converting words to lowercase so that the results of the analysis accuracy are good. After that, the process of extracting features using the term frequency-inverse document frequency (TF-IDF) method, and then the data is divided into two parts (70% for training and 30% for examination). It is worth noting that two files were collected, the first containing (2005) positive words, and the second containing (4,781) negative words in the English language for the second dataset, where the words of the negative tweet are compared with the positive words collected in the file. Delete this word from the tweet, and if a negative word is found within the positive tweet, this word is deleted from the tweet, this step led to an increase in the accuracy of the classification.

## 2. RELATED WORK

Nowadays, researchers and stakeholders use social media as a robust statistical resource to analyze sentiment for achieving or anticipating those related outcomes. Social media is a good platform for expressing feelings, opinions, and initiatives. Twitter is one of the most popular social media platforms across the world. It presents a decent way for people to express themselves honestly. Several studies similar to this study will be reviewed, arranged in descending order depending on the value of accuracy.

Adamu *et al*. [6] six machine learning (ML) algorithms are used in this study, and a comparative analysis of their performance is conducted. The algorithms are Multinomial Naïve Bayes (MNB), support vector machine (SVM), random forest (RF), logistics regression (LR), K-nearest neighbor (KNN), and decision tree (DT). The conducted experiments reveal that the SVM outperforms the remaining classifiers with the highest accuracy of 88%.

Shofiya and Abidi [7] the research focused on analyzing the feelings of a group of Canadian tweeters towards social distancing that is instructed officially due to COVID-19 consequences for approximately thirty days, relying on Twitter's data. Authors used the SentiSt Strength Tool and SVM Classifier to carry out that analysis, resulting in 40% of neutral feelings of Canadian people with instructed distancing, while other percentages of 35% of negative feelings, however, 25% of Canadian tweets had positive feelings about it. The outcome showed an accuracy of 87%, using the SVM algorithm.

Villavicencio *et al*. in [8] studied and analyzed people's feelings towards COVID-19 vaccines in the Philippines based on their opinions i.e., positive, neutral, or negative. According to the results, it is obvious that 83% of the tweets were positively supporting the idea of vaccination, whereas 9% of them were neutral, and only 8% stated negative feelings. The data was preprocessed using various natural language processing (NLP) techniques, and a classifier model was successfully developed using the Naïve Bayes classification algorithm with an accuracy of 81.77%. Sari and Ruldeviyani [9] research was made to analyze the sentiment of the COVID-19 transmission to commuter line passengers. This research was implemented using a comparison of 2 methods, Naïve Bayes outperformed the decision tree with an accuracy of 73.59%.

## 3. METHOD

Based on the analysis of the literature, it was determined that there are existing problems that need to be resolved to perform sentiment classification. Adopting supervised learning significantly reduces computational complexity and provides accuracy at the expense of a larger volume of training data. This section discusses the proposed methodology for analyzing and categorizing sentiment for the dataset used. With the presented dataset the spyder IDE was used as an interface to work with Microsoft Windows 10. Sentiment analysis was performed in five steps as shown in: i) preprocessing, ii) feature extraction, iii) tagging of Tweets and dataset segmentation, vi) Implementation of classification algorithms, and v) comparison of evaluation results. Figure 1 illustrates these steps. Before going into details of the action steps, it is necessary to identify and classify the dataset used in sentiment analysis. The steps of the proposed approach were applied to two datasets that will be detailed in subsections.

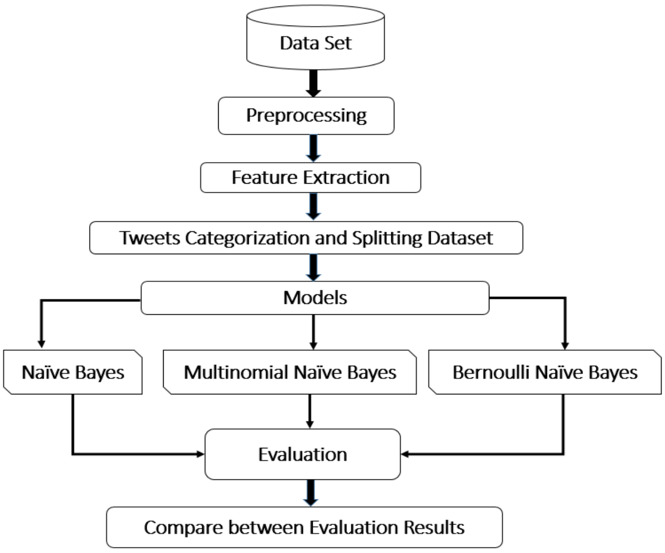

Figure 1. The general structure of the proposed system

### 3.1. Dataset of tweeters during the COVID-19 pandemic

Table 1 demonstartes extracted dataset from kaggle.com. It is a set of 3,057 tweets, which is representing four sentiments that reflect the feelings of tweeters during the COVID-19 pandemic. These sentiments are joy, fear, anger, or sadness. Which will later be categorized into four categories according to the aforementioned sentiments.

Table 1. Sentiments embedded in tweets of dataset A

| Sentiment type | Count of tweets | Sample of tweet |
|---|---|---|
| Joy | 708 | "Let us set the pace for innovation and creativity, bring out the entrepreneur in you, and break barriers of productivity 19pic twitter com/repw43b7qf." |
| Sad | 787 | 1881, Sensex closes 1 203 points lower nifty gives up 8 300 amid coronavirus crisis it financial stocks worst hit â€¦ shared via ndtv news app (android \| iPhone) |
| Fear | 801 | 1763, stone pelters need to be treated as terrorists they need to be detained under terrorist, and disruptive activities (prevention) act 1987 (tada) stone pelters are more danger than corona we need to vanish both stone corona & pelters (terrorist) from India |
| Anger | 761 | 3506,â€œif regular unemployment hits 20% black unemployment will likely be around 50% that will be a full collapse for black America â€• â€” â€œmnuchin warns senators that we could see a 20% unemployment rate due to coronavirus |

### 3.2. Dataset of tweets included in the Python libraries

The dataset consists of 10,000 tweets including one of the Corpus. It was into the NLTK library with the name (twitter_samples), and these tweets are randomly selected. It's divided into (5,000 positive tweets and 5,000 negative tweets) as shown in Table 2.

Table 2. Sentiments embedded in tweets of dataset B

| Sentiment type | Count of tweets | Sample of tweet |
|---|---|---|
| positive | 5000 | "This is a movie that refreshes the mind and spirit along with the body, so original are its content, look, and style." |
| negative | 5000 | " Stupid, naive, and undeserving of what he got " |

### 3.3. Preprocessing

Preprocessing is one of the most effective techniques to make sure of the data's correctness. Before applying the analysis algorithms. It includes (data cleaning, tweets splitting, sentences splitting, stopwords, and lemmatization performing).

#### 3.3.1. Data cleaning

Cleaning the data is an early preprocessing step to ignore any unnecessary qualities in the data to reduce processing time and focus modeling work on the data necessary pieces. The cleanup process includes removing unique hypertext markup language (HTML) entities, converting all characters in words to lowercase, removing hyperlinks, punctuation, whitespace, and special characters that appear in sentences e.g., "([0-9] +) | (#) | (@[A-Za-z0-9] +) | ([^0-9A-Za-z \t]) | (\w +: \ / \ / \ S +) " or English abbreviation e.g., don't = not or I'm = I am or OMG = Oh My God. Data cleansing is a very necessary step to initialize data for the next steps of preprocessing [10]–[13].

#### 3.2.2. Tokenization

It is the procedure of splitting raw textual data and converting it to several separated tokens that will eventually be a word or a character. The purpose of tokenization is to research phrases in a sentence or phrase [14], [15]. There are two types of tokenization: i) sentence tokenization: it is the process of dividing a textual document into a group of sentences. The objective of this operation is to convert the texts into meaningful sentences. These techniques are used to find the separation marks between two sentences such as a duration (.) and a newline character (\n) to achieve sentence tokenization [10], [11]; ii) word tokenization: it is the process of separating the sentence into words that form the targeted sentence, called tokens. These techniques are used to find segregation words such as dot (.) or separator (,) whitespace to achieve tokenization between words [12], then each word convert to lowercase.

#### 3.3.3. Remove stop-words

Stop words are parts of the natural humane language that do not make sense or have a meaning. A stop word is a group of usual frequented features that appears in all textual documents. Common features e.g., conjugations such as and, or, but, pronouns, he, and she is supposed to be removed because they have a little or no effect on the text mining process. it is difficult to understand the content of the text files that contain stop words because of the appearance of these words as they appear very frequently. Removing stop-words from textual documents helps to sort the text's appearance orderly by eliminating the less important words, decreasing the amount of processed textual data, which gains better system performance [16], [17].

#### 3.3.4. Lemmatization

Lemmatization is an important preprocessing step for several text mining applications. it is used in natural language processing, representing a useful tool for sentiment analysis and classification processes. Lemmatization converts each word to its basic form, the lemma. For instance, "good," "better," or "best" are converted to the root word “good”. It helps with determining part of speech (POS) tagging, returning the words form, with mandatory validity. The goal of lemmatization is to reduce each word's inflectional forms and derivations, returning them to their common root [18].

#### 3.3.5. Part of speech tagging

It is a way in which the part of speech specifies each word in a phrase. POS knowledge plays an essential role as any word in a sentence with each POS tag has a distinct meaning depending on the sentence situation with the essential speech components such as verb, noun, adverb, and adjective. POS tags can be used to differentiate words and articulate their speech parts. The Lemmatization process relies mainly on POS tagging [19].

### 3.4. Feature extraction

It is an important process of work. The process of decreasing inputs to analyze, process, or manage the most considerable data is called feature selection [20]. Thus, several features are extracted from the dataset. The extracted features must be in a specific format that can directly be an input to the classification algorithms. This paper used the TF-IDF method.

Its main application is to determine the significance of a particular word for a document from a specific dataset. Each word in the document is assigned a weight:

$$TF - IDF(t, d) = TF\ (t, d).IDF(t).N \quad (1)$$

$$IDF(t) = log\ N/DF(t) \quad (2)$$

where the frequency of the term 't' is represented as TF (t, d) appearing in a specific document 'd', with the total size of 'N' represented as IDF divided by the number of documents that make up the dataset D, contains the term t [16], [20], [21].

### 3.5. Tweets categorization and splitting dataset

Classify the data into specific categories, such as sadness, joy, anger, or fear, in dataset A. It is an important piece of data in machine learning procedures. The training set of our model is based on historical data with predefined target attributes (values). The process of marking the dataset must be done carefully because of the susceptibility to errors that cause inaccuracies and thus affect the quality of the dataset and the performance of the model in data analysis.

Through a process of classifying (sad, joy, anger, or fear) the training part of the dataset (tweets), the entire dataset is pre-split into two parts, 70% for training and 30% for testing. The same procedures are also applied with dataset B, but the categories in it are only two categories (negative and positive). A collection of the Naïve Bayes family has been used and each one will be explained in the next subsection.

#### 3.5.1. Naïve Bayes

The sentiment analyzer is built using the model of Naïve Bayes as a classifier. This model is learning from the labels of the training part, performing the sentiments classification. It supposes that the existence of a particular feature within a class is independent of the existence of the other features within the same class. NB theory determines the probability of a specific event to have happened based on the probabilistic related distributions of other particular events [22]-[24].

With this study, the dataset (training set) containing the labeled tweets was set as an input to each model for training on the characteristics of the tagged sentence's emotional traits.
Based on Bayes's theory, it is expressed as:

$$P(H|X) = \frac{P(X|H)X\,P(H)}{P(X)} \tag{3}$$

where:
- P(H|X) denotes the final probability of hypothesis H happening when a specific event E happens.
- P(X|H) denotes the probability of proofing the event E will influence H.
- P(H) denotes the initial probability when H happens irrespective of any proof.
- P(X) denotes the initial probability proof of E of H or other proof.

The two variables used to use Bayes's theory are aspects/features as H, and sentiments as E. A sentence consists of several words, while practically, it is not easy to out which tweet can be nominated as an aspect/feature. Thus, it is reasonable to assume every word as an aspect/feature to apply Bayes' Theorem.

$$P(C|A) = \frac{P(A|C)X\,P(C)}{P(A)} \tag{4}$$

Where:
A is a word or a feature.
C is a sentiment value or category.
Because the features of words that support one category can be many, e.g. there are features A1, A2, and A3, the Bayes theory can be developed into:

$$P(C|\text{A1,A2,A3}\ ) = \frac{P(\text{A1,A2,A3 3}|C)X\,P(C)}{P(\text{A1,A2,A3 })} \tag{5}$$

because Bayes's theory requires that the evidence (in this case is a word or feature) that exists is independent of each other, then the formula can be changed to:

$$P(C|\text{A1,A2,A3}\ ) = \frac{P(A1|C)XP(A2|C)XP(A3|C)X\,P(C)}{P(A1)XP(A2)XP(A3)} \tag{6}$$

if described in general can be formulated as shown in:

$$P(C|A) = \frac{\Pi_{i=0}^{q}X\,P(A1|C)}{P(A)} \tag{7}$$

because the fixed value of P(A) for a Sentiment value the P(A) value is determined only if the $\Pi_{i=0}^{q}X\,P(A1|C)$ is determined.

### 3.5.2. Multinomial Naïve Bayes

It is similar to Naïve Bayes considering a probabilistic technique. Multinomial NB develops the utilization of the Naïve Bayes algorithm. It uses NB for data that is partitioned multinomially, however, it is a frequency-depended model. The Multinomial Naïve Bayes algorithm operates with the definition of the term's frequency, explaining the iteration of item repetition during operation. the main difference between the classifiers of Naïve Bayes and Multinomial Naïve Bayes is that Naïve Bayes operates based on conditional probability (as conditional independence of the characteristics is considered), however, the Multinomial Naïve Bayes operates based on the multinomial distribution. In other words, Multinomial NB is considered an updated version of the NB algorithm. It effectively helps to calculate the frequency of any item [24]-[27]. Multinomial Naïve Bayes work can be illustrated by the following equation [28]-[33]:

$$\mathrm{P}(\mathrm{t\kappa}|\mathrm{c}) \; = \frac{\mathrm{Ttc}}{\sum \mathrm{t}' \epsilon \nu \mathrm{Tct}'\prime} \tag{8}$$

Where:

- $\mathrm{P}(\mathrm{t\kappa}|\mathrm{c})$: is the conditional probability of the word ($tk$) that appears in the document having class c.
- Ttc: is the number of occurrences of the word (t) in the document having class c.
- $\sum \mathrm{t}' \epsilon \nu \mathrm{Tct}''$ : is the total number of occurrences of all words in class c.

### 3.5.3. Bernoulli Naïve Bayes

It is a classifier that works efficiently on the binary concept when the items appear or not, unlike Multinomial NB, Bernoulli NB does not notify the frequency of the term. It does not manipulate the same multinomial process where the term frequencies are considered by the multinomial approach. In contrast, the Bernoulli NB approach is only beneficial in determining the presence of a term in the text under consideration. In the multivariate Bernoulli Naïve Bayes algorithm, features are distinct binary variables, explaining the appearance or absence of the term in the file under specified consideration [24], [30]. Algorithm 1 depicts the implementation steps of the proposed system. Bernoulli's Naïve Bayes work can be illustrated by (9).

$$P(tk|c) \; = \frac{1+|Trtk,c|}{2+|Trc|} \tag{9}$$

Where:

- $t$: represent term in the document.
- $tk$: represent many times terms appear in the document.
- $P(tk|c)$: represent the conditional probability of the term ($tk$) that appears in the document having class c.
- $Trtk, c$: represent the number of documents of class c that represent appear of terms ($tk$).
- $Trc$: represent the total number of documents of class c.

Algorithm 1: The implementation steps of the system

```
Input: tweets Dataset
Output: The best Prediction Method
Begin
Step 1: for each tweet in the dataset // The same steps for the entered dataset (first or
second dataset)
 call Preprocessing function
       •  Data cleaning
       •  Abbreviation processing
       •  Remove Numbers and other marks
       •  Delete website links
       end for
Step 2: for each tweet in the dataset
 call Tokenization function // Tokenize each tweet or comment into single words
 convert to lowercase.
 call Remove Stop words & punctuation function
 call Word Lemmatization function
 call part of speech tagging function
 end for
Step 6: Splitting data into training and testing. // 70% training and 30% testing
Step 7: Switch (x) // After training the models, each model is called to check its accuracy
Step 7.1: Case (1): Call NB model for classification; Break;
Step 7.2: Case (2): Call Bernoulli NB model for classification; Break;
Step 7.3: Call Multinomial NB model for classification; Break;
```

```
End Switch
Step 6: Compare among results of the models' accuracy measurement and choose the best
method.
End
```

## 4. RESULT AND DISCUSSION

The same steps were applied to the first dataset (subsection A) and the second dataset (subsection B), but the difference is that the first dataset is classified into anger, fear, sadness, or joy depending on the content of the tweet, while the second dataset is classified into negative and positive tweets depending on the characteristics of the tweet that reflects the user (writer) orientation, the clean dataset is set as input to the rating model and sentiment analysis. The results for each model are covered in Tables 3 and 4, showing the models gained accuracy for each dataset.

Table 3 shows the accuracy values for each model for the first dataset. As mentioned earlier in this paper, the Multinomial Naïve-Bayes model achieved the highest accuracy values (91.6%). It should be noted that after applying the pre-processing steps to the dataset, words showing the direction and category of the tweet appear from one of the four categories, and therefore it is extracted as a calculated category, while table 4, which represents the accuracy results of the models for classifying the second dataset. As in Table 3, the Multinomial Naïve-Bayes model has outperformed the rest of the models, as it scored higher accuracy than the rest of the models (87.6%).

Because Multinomial Naïve-Bayes depends on the principle of a frequency-depended model of the feature and because the tweets are in a group The text data uses some repeating elements (words), where after the pre-processing process and when extracting the attributes, this attribute will be repeated to a certain extent, which enables the Multinomial Naïve-Bayes model to classify tweets more accurately than the rest of the models. Compared with the related works that studied problems similar to the problem of this research, the accuracy achieved by this system is higher than the accuracy achieved by some similar works (in terms of the type of dataset and similarity of the method of extracting features) as in Table 5.

Table 3. Accuracy value for each model of dataset A

| Model | Accuracy value |
|---|---|
| Naïve-Bayes | 83.5% |
| Bernoulli Naïve-Bayes | 83.4% |
| Multinomial Naïve-Bayes | 91.6% |

Table 4. Accuracy value for each model of Dataset B

| Model | Accuracy value |
|---|---|
| Naïve-Bayes | 82.4% |
| Bernoulli Naïve-Bayes | 85.9% |
| Multinomial Naïve-Bayes | 87.6% |

Table 5. Summerize of related work

| Ref | Title and publishing year | Method used | Dataset source | Feature extraction | Highest accuracy |
|---|---|---|---|---|---|
| [6] | Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning-2021 | Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), Random Forest (RF), Logistics Regression (LR), K-Nearest Neighbor (KNN), and Decision Tree (DT) | Nigerian Local English Slang-Pidgin (NLES-P) | TF-IDF | Support Vector Machine accuracy was 88%. |
| [7] | Sentiment analysis on covid-19-related social distancing in Canada using Twitter data - 2021 | Support Vector Machine | Twitter | TF-IDF | SVM accuracy was 87%. |
| [8] | Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve Bayes-2021 | Naïve Bayes | Twitter | TF-IDF | Naïve Bayes algorithm accuracy was 81.77% |
| [9] | Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers-2020 | Naïve Bayes and Decision Tree | Twitter | (unknown) | Naïve Bayes algorithm accuracy was 73.59% |

Based on the results presented in Tables 3 and 4 and Figure 2, it is possible to adopt the results of (Multinomial NB) as the most efficient and optimal algorithm, while dispensing with the rest of the algorithms to solve a problem of this kind.

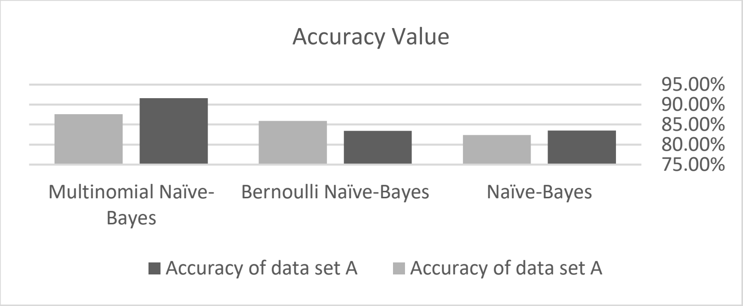

Figure 2. Accuracy value for each model and both datasets

## 5. CONCLUSION

The objective of this study is to analyze and classify the data of social media users and to know their attitudes, sentiments, and interaction with an event. The study concluded that the features used, which represent the direction of the tweet, usually, these features are repeated, so the multinomial NB model succeeded in achieving a higher classification accuracy than the rest of the used machine learning models (it achieved an accuracy of 91.6% in the first dataset and an accuracy of 87.6% for the second dataset). Despite the different topics of tweets for the two datasets, the system achieved a good accuracy value when applied to them, as indicated by the accuracy tables. From the observations of this study, it is possible to increase the classification accuracy of the approach either by using more powerful models or by using the method of extracting other features depending on the type of data and the extent to which they are free of impurities, as well as the type of problem that this approach is intended to solve. The researchers seek to do similar work to this study, which was a prediction of the users' opinion regarding a particular product or issue before it is released on the ground, and also the researchers seek to link this approach directly with Twitter through the (tweeps) library in the Python programming language to be analyzing tweets interactively directly.

## REFERENCES

[1] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: a large-scale benchmark twitter data Set for COVID-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 976–988, Aug. 2021, doi: 10.1109/TCSS.2021.3051189.

[2] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," *Knowledge-Based Systems*, vol. 228, p. 107242, Sep. 2021, doi: 10.1016/j.knosys.2021.107242.

[3] R. Kimmons, J. Rosenberg, and B. Allman, "Trends in educational technology: what Facebook, Twitter, and Scopus can tell us about current research and practice," *TechTrends*, vol. 65, no. 2, pp. 125–136, 2021, doi: 10.1007/s11528-021-00589-6.

[4] D. H. Abd, A. T. Sadiq, and A. R. Abbas, "Political Arabic articles classification based on machine learning and hybrid vector," in *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, Nov. 2020, pp. 1–7, doi: 10.1109/CITISIA50690.2020.9371791.

[5] D. V. Cruz, V. F. Cortez, A. L. Chau, and R. S. Almazán, "Does Twitter affect stock market decisions? financial sentiment analysis during pandemics: a comparative study of the H1N1 and the COVID-19 periods," *Cognitive Computation*, vol. 14, no. 1, pp. 372–387, Jan. 2022, doi: 10.1007/s12559-021-09819-8.

[6] H. Adamu, S. L. Lutfi, N. H. A. H. Malim, R. Hassan, A. Di Vaio, and A. S. A. Mohamed, "Framing Twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning," *Sustainability*, vol. 13, no. 6, p. 3497, Mar. 2021, doi: 10.3390/su13063497.

[7] C. Shofiya and S. Abidi, "Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 5993, Jun. 2021, doi: 10.3390/ijerph18115993.

[8] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," *Information*, vol. 12, no. 5, 2021, doi: 10.3390/info12050204.

[9] I. C. Sari and Y. Ruldeviyani, "Sentiment analysis of the Covid-19 virus infection in Indonesian public transportation on twitter data: a case study of commuter line passengers," in *2020 International Workshop on Big Data and Information Security (IWBIS)*, Oct. 2020, pp. 23–28, doi: 10.1109/IWBIS50925.2020.9255531.

[10] S. Vijayarani and R. Janani, "Text mining: open source tokenization tools-an analysis," *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37–47, 2016.

[11] S. Vijayarani, J. Ilamathi, and Nithya, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[12] F. M. J. M. Shamrat *et al.*, "Sentiment analysis on twitter tweets about COVID-19 vaccines usi ng NLP and supervised KNN classification algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, p. 463, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp463-470.

[13] D. H. Abd, A. R. Abbas, and A. T. Sadiq, "Analyzing sentiment system to specify polarity by lexicon-based," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 283–289, 2021, doi: 10.11591/eei.v10i1.2471.

[14] H. Zhao, L. Huang, R. Zhang, Q. Lu, and H. Xue, "SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3239–3248, doi: 10.18653/v1/2020.acl-main.296.

[15] D. Sarkar, *Text Analytics with Python - A Practitioner's Guide to Natural Language Processing*. Springer, 2019.
[16] A. I. Kadhim, "An evaluation of preprocessing techniques for text classification," *International Journal of Computer Science and Information Security*, vol. 16, no. 6, pp. 22–32, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/.
[17] S. Qaiser and R. Ali, "Text mining: Use of TF-IDF to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.
[18] V. Dunjko and H. J. Briegel, "Machine learning & artificial intelligence in the quantum domain: A review of recent progress," *Reports on Progress in Physics*, vol. 81, no. 7, p. 074001, Jul. 2018, doi: 10.1088/1361-6633/aab406.
[19] M. R. Khatun, S. I. Ayon, M. R. Hossain, and M. J. Alam, "Data mining technique to analyse and predict crime using crime categories and arrest records," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, p. 1052, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1052-1060.
[20] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and Bernoulli Naïve Bayes for text classification," in *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*, Apr. 2019, pp. 593–596, doi: 10.1109/ICACTM.2019.8776800.
[21] J. K. Kruschke and T. M. Liddell, "Bayesian data analysis for newcomers," *Psychonomic Bulletin & Review*, vol. 25, no. 1, pp. 155–177, Feb. 2018, doi: 10.3758/s13423-017-1272-1.
[22] J. A. Hatch, "Deciding to do a qualitative study," *Doing Qualitative Research in Education Settings*, pp. 1–35, 2002.
[23] V. Kalra and R. Aggarwal, "Importance of text data preprocessing &amp; Implementation in RapidMiner," in *Proceedings of the First International Conference on Information Technology and Knowledge Management*, Jan. 2018, vol. 14, pp. 71–75, doi: 10.15439/2017KM46.
[24] H. R. Arabnia, K. Daimi, R. Stahlbock, C. Soviany, L. Heilig, and K. Brüssau, "Correction to: principles of data science," in *Principles of Data Science*, Springer, 2020, pp. C1–C1.
[25] A. Naresh and P. V. Krishna, "An efficient approach for sentiment analysis using machine learning algorithm," *Evolutionary Intelligence*, vol. 14, no. 2, pp. 725–731, Jun. 2021, doi: 10.1007/s12065-020-00429-1.
[26] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, and X. Wu, "Chinese text sentiment analysis based on extended sentiment dictionary," *IEEE Access*, vol. 7, pp. 43749-43762, 2019.
[27] J. K. Alwan, A. J. Hussain, D. H. Abd, A. T. Sadiq, M. Khalaf, and P. Liatsis, "Political Arabic articles orientation using rough set theory with sentiment lexicon," IEEE Access, vol. 9, pp. 24475–24484, 2021, doi: 10.1109/ACCESS.2021.3054919.
[28] A. A. Farisi, Y. Sibaroni, and S. Al Faraby, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier," *Journal of Physics: Conference Series*, vol. 1192, no. 1, p. 012024, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012024.
[29] Q. B. Baker, F. Shatnawi, and S. Rawashdeh, "Forecasting epidemic diseases with Arabic Twitter data and WHO reports using machine learning techniques," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 739–749, Apr. 2022, doi: 10.11591/eei.v11i2.3447.
[30] P. P. M. Surya, L. V. Seetha, and B. Subbulakshmi, "Analysis of user emotions and opinion using Multinomial Naive Bayes Classifier," in *Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019*, Jun. 2019, pp. 410–415, doi: 10.1109/ICECA.2019.8822096.
[31] T. A. Almeida, A. Yamakami, and J. Almeida, "Evaluation of approaches for dimensionality reduction applied with naive bayes anti-spam filters," in *2009 International Conference on Machine Learning and Applications*, Dec. 2009, pp. 517–522, doi: 10.1109/ICMLA.2009.22.
[32] D. N. Mhawi, "Proposed hybrid correlation feature selection forest panalized attribute approach to advance IDSs," *Modern Science*, vol. 7, no. 4, p. 15, 2021.
[33] D. N. Mhawi, A. Aldallal, and S. Hassan, "Advanced feature-selection-based hybrid ensemble learning algorithms for network intrusion detection systems," *Symmetry*, vol. 14, no. 7, p. 1461, Jul. 2022, doi: 10.3390/sym14071461.

## BIOGRAPHIES OF AUTHORS

**Murtadha B. Ressan** Senior programmer in the Iraqi Federal Ministry of Construction, Housing and Municipalities, Member of the Committee for the Development and Modernization of the Governmental Human Resources System. Research interests focus on developing and modernizing the financial and administrative systems of government agencies. He has many researches published in the local journal of the University of Technology. He holds a master's degree in computer science from the University of Technology. He can be contacted at email: cs.19.02@grad.uotechnology.edu.iq.

**Rehab F. Hassan. Assist** Professor in Computer Science department at University of Technology, Iraq. Her researchs interest is in the area of Wireless Sensor Networks, Mobile Computing, Information Technology, and Intelligent Environment/IoT, and GIS. She has published more than 50 conference/journal papers. She obtained a PhD, an MSc, and a BSc, all in Computer Science from university of Technology. She can be contacted at email: Rehabf.hassan@uotechnology.edu.iq.