Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 5, pp. 943–958
DOI : 10.47974/JSMS-1194

Comparing the accuracy of support vector machine and the Naïve Bayes algorithm : Twitter sentiment analysis during the COVID-19 pandemic

Aseel Almansour
Department of Statistics, Science College
King Abdulaziz University
P. O. Box. 80200
Jeddah 21589
Saudi Arabia

Abstract

The COVID-19 pandemic has imposed substantial challenges globally, and many have voiced their thoughts on social media about the unfolding crisis. This study performed a sentiment analysis (SA) on Twitter posts, to assess the general public's perspective on the pandemic. Our primary objective was to compare the performance of two frequently used algorithms, support vector machine (SVM) and Naive Bayes classification (NBC), in performing SA, which is a novel direction. The results showed that the more sophisticated SVM algorithm obtained 81% accuracy compared with the 74% achieved by NBC. Consequently, SVM may be used to assess and acquire improved public SA from Twitter.

Subject Classification: *62-08 Computational methods for problems pertaining to statistics.*

Keywords: *Sentiment analysis, Naive bayes classification, Support vector machine, Machine learning, COVID-19.*

1. Introduction

COVID-19, the "pandemic of the century," has resulted in significant global challenges, such as controlling the virus's spread, safeguarding public health, and implementing effective governance measures (World Health Organization, 2020) [4]. The lack of data on authentic public sentiments is a major contributor to these issues. Since early 2020, as the world has endured multiple cycles of COVID-19, the public has been subject to lockdowns and restrictions, resulting in an increased reliance on Twitter and other social media platforms for expressing opinions Agarwal

E-mail: aseel_almansour@yahoo.com

et al., (2011).[7]. These opinions can aid governing bodies and administrations in identifying trends and implementing policy changes more efficiently by providing valuable insights into the public's perspective.

This study aims to collect and analyze information regarding public sentiments during the COVID-19 pandemic. Using data mining techniques such as Naive Bayes classification (NBC) and support vector machine (SVM) to classify user perspectives, this study builds on previous research on sentiment analysis using social media data, including Twitter Singh & Dass et al., (2018) [3]. The NBC method calculates the posterior probability for each class based on the frequency of the text, while the SVM algorithm generates hyperplane equations by classifying the data Bhatnagar et al., (2020) [1]. This study seeks to determine the most efficient technique for analysing Twitter data for sentiment during the pandemic by comparing the performance of these two algorithms.

The descriptive analysis of COVID-19 patients in India performed by Bhatnagar et al. (2020)[1] offers crucial insights into the factors that may influence the virus's spread in the country. Age was not a significant factor in determining susceptibility to the disease, but there was a correlation between gender and transmission type. Incorporating these findings into the current research, we plan to examine how the COVID-19 pandemic may have affected various aspects of society, with a focus on the role of gender and transmission type in the spread of the virus and its effects.

During the COVID-19 pandemic, Twitter and other social media platforms have emerged as indispensable information sources. Munjal et al. (2018)[2] examined the use of Twitter sentiments for predicting trends in the film industry. Using lexical analysis, their research presented an efficient visualization-based framework for identifying opinions in tweets. This methodology can be used to investigate the public's attitudes and opinions regarding the COVID-19 pandemic and its effects on various sectors of society in the context of the present study. By analyzing the sentiments expressed on Twitter, we can obtain insight into the public's concerns, fears, and perceptions of the pandemic, as well as how these factors may have impacted their behavior and decision-making during this challenging time.

Combining the findings of Bhatnagar et al. (2020)[1] and Munjal et al. (2018)[2], the current study seeks to examine the role of gender and transmission type in the spread of COVID-19 and its repercussions on various sectors of society. In addition, by utilising the Twitter sentiment analysis framework, we expect to comprehend the public's perspective on

the pandemic and how it has influenced their behavior and decisions. This exhaustive analysis will aid in the development of targeted strategies and policies to mitigate the multifaceted effects of the COVID-19 pandemic and promote societal resilience.

Previous research has examined the use of Twitter for sentiment analysis in various contexts, such as predicting film industry trends based on user opinions Mousavizadeh et al., (2020)[6]. Bennett et al. (1998)[8] developed a visualization-based framework that detects opinions from tweets using a lexical analysis approach, thereby providing industry stakeholders with valuable insights. Ben et al. (2020)[5] performed a descriptive analysis of COVID-19 patients in India, focusing on characteristics such as age, gender, and travel history. This study highlights the potential of data mining techniques for analyzing public sentiment during a pandemic.

This study contributes by assessing the effectiveness of the SVM and NBC algorithms in determining Twitter sentiments, particularly during the COVID-19 pandemic. This study seeks to contribute to data-driven decision-making during public health crises by providing a quick and accurate method for understanding public opinion on a widely used social media platform such as Twitter. In the end, the results demonstrate that the more advanced SVM algorithm obtained 81% accuracy compared to NBC's 74% Bhatnagar et al., (2020)[1], indicating that SVM may be better suited for evaluating and acquiring improved public sentiment analysis from Twitter during the COVID-19 pandemic.

Combining the methodologies and findings of prior research Borah et al., (2021)[9]; Chen et al., (2020)[10]; Davidson et al., (2020)[11], this study provides a comprehensive understanding of the role of gender and transmission type in the spread of COVID-19 and its effects on various sectors of society. Moreover, it demonstrates the potential for Twitter sentiment analysis to monitor public opinion and identify trends that can inform policy decisions during a public health emergency. This study's exhaustive methodology not only adds to the existing body of knowledge on COVID-19, but also lays the groundwork for future research on the role of social media data in comprehending public sentiment during global crises.

2. Literature Review

SA using SVM and NBC, which is the primary objective of the current work, has been investigated by several studies. Using this strategy,

researchers have conducted statistical analyses of elements such as movie reviews and e-commerce product reviews to determine the efficacy of their techniques.

Zhang et al. (2011)[13] developed compositional sentiment rules for calculating textual sentiments. The algorithm generated a straightforward form that did not require machine learning (ML). Using SVM, Dey et al., (2012)[12] extracted a feature map from a dataset; nevertheless, the system faced difficulties when a sarcastic statement was run through their linear regression model. Pang et al. categorized movie reviews using SVM and NBC. The accuracy of the Naive Bayes model increased to 81%, while that of the SVM model increased from 72% to 81%. Khorsheed et al., (2013)[14] suggested a method for selecting features based on the value of the chi-square statistic, thereby assisting in decreasing dimensionality and classifying the dataset using an SVM. Nevertheless, no comparison evaluations were conducted. Giachanou et al., (2016)[15] studied whether a more sophisticated information-retrieval feature weighting scheme may increase classification accuracy. However, they only analyzed a thousand characteristics for SA. Using machine learning techniques, Go A. et al., 2009 [16] explored the polarization of attitudes on Twitter datasets. Eventually, their precision should be enhanced by selecting new characteristics and utilizing a more effective preprocessing method. This research aimed to solve these limitations and establish a more efficient SA procedure by comparing the NBC and SVM algorithms during SA using first-hand data on COVID-19 from Twitter.

2.1 *Problem statement and objectives*

In the context of microblogging, SA is a relatively recent area of study. Prior research has concentrated on the SA of user evaluations, articles, online blogs, and publications, as well as the SA of generic phrases. However, few studies have examined the efficacy of ML algorithms such as NBC and SVM. Utilizing an ML algorithm is a precondition for performing SA. The Naive Bayes algorithm and SVM are two important ML algorithms; Bayes theorem is the foundation of the NBC approaches. The Naive Bayes classifier provides outstanding results when analyzing text data. Natural language processing is one example.

The Naive Bayes method calculates a probability based on the provided dataset (Twitter data regarding COVID-19), and the Naive Bayes classifier can be employed as a probabilistic classifier. Classifiers are implemented using the concept of a mixed model. By contrast, the input

and output formats of SVM have been standardized. SVM is a general-purpose learner whose output is either positive or negative, and its input is a vector space. Text papers are not conducive to learning; therefore, texts are grouped and transformed into a format that matches the input of an ML algorithm. After calculating the scores of the texts, the scores are input into the SVM. SVM has been demonstrated to be a practical text categorization learning system.

During the pandemic, several studies explored Twitter attitudes; this study fills a research gap by comparing the performance of the Naive Bayes and SVM ML algorithms in developing an efficient strategy. Accuracy, precision, recall, and F1 measure were used to distinguish both algorithms and evaluate their effectiveness.

The Naive Bayes methodology is faster than that adopted by SVM because the NBC algorithm assesses models and considers them independent. By contrast, the SVM algorithm evaluates the relationships between models to a certain degree. Thus, SVM has a greater likelihood of appropriately recognizing emotions. This study also investigated the contradicting NBC and SVM positions.

Because Twitter is a prominent microblogging network where individuals express their thoughts, we selected this social network for conducting our research. The purpose of this study was to collect, assess, and analyze people's responses to Tweets during a specific period. In addition, we compared the SVM and NBC algorithms to identify which algorithm is optimal for SA.

3. Methodology

This study attempted to establish the general public's opinions on the efficacy of the response to the coronavirus pandemic. We used Python, as it has straightforward data collection, processing, and visualization tools. All the data were gathered using the Tweepy package, and a statistical analysis was conducted in Google Collab utilizing the TextBlob Python tool. We considered R a backup option if the Python library failed to accomplish our goal. Using the Twitter API, tweets with hashtags #coronavirus and #COVID-19 were collected. SA was conducted on the resulting dataset.

SA based on Twitter data is a novel application area of natural language artificial intelligence. The Tweepy python package and Twitter API provided by the Twitter Developer were used to collect data.

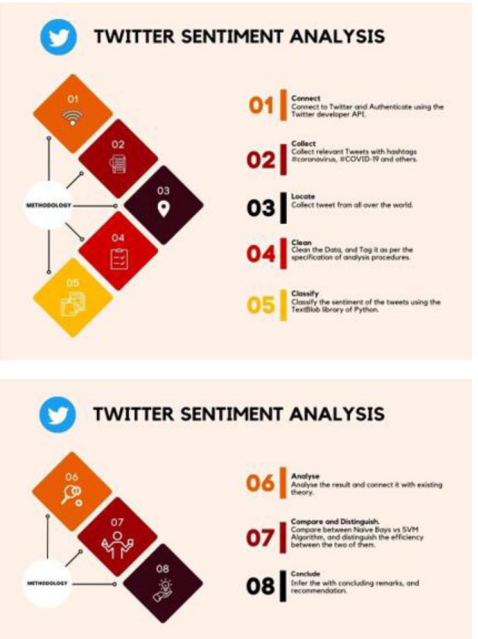

Figure 1

Methodology of the study

Numerous fields from this API were captured via web scraping, including text, source, retweets, language, user, and location.

After collecting twitter data in text format, the TextBlob library was used to identify the text's positive, negative, or neutral polarity. SA is a systematic strategy beginning with data collection to identify extracted data and related properties. For simplicity, only English texts were evaluated for our study. Furthermore, sentiment categorization was performed, and finally, an analysis was conducted on that categorization. The entire procedure is depicted in detail in Figure (1).

3.1 *Basic definitions and theoretical framework*

3.1.1 Support Vector Machine Algorithm

SVM is a well-known supervised ML model for classifying and predicting unknown data. Several researchers have shown that SVM is an accurate technique for text Jakkula et al., (2006) [17]. Furthermore, SVM is widely used in SA. In particular, it is a powerful modeling approach that can analyze and categorize unknown data into categories included in the

training set. SVM is a linear learning algorithm that finds the best hyperplane to differentiate between two classes. As a supervised classification model, it maximizes the distance between the closest training point and the class to improve the classification performance on the test data.

The classification process of SVM includes the following steps:

(a) SVM takes the labeled data sample and draws a line separating the two classes, referred to as the decision boundary. Only training data points extremely close to the decision boundary are used in the solution, and the data points are referred to as support vectors.

(b) When classifying new data, they are assigned to either the left or right side of the decision boundary. Data are classified following the side they enter.

To classify our data with the best precision, we specified categories such that the data are evenly spread on both sides of the decision boundary. For simplicity, we considered a set of tweets in a .csv file as x_i, where $x_i = x_{ia,} x_{ib,} x_{ic}$. Here, our set of tweets had a specified pattern, which could be grouped and termed positive and negative classes. Therefore, we termed each word positive or negative by denoting them as either +1 or -1. Subsequently, our tweets were specified in pairs as $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$, and then data that were symmetrically separated by the p dimensional separator function, also known as the hyperplane H_o,could be considered. If we have n amount of data in our set, we can define them as follows:

$$w.x_i + b = 0, \tag{1}$$

where b and w represent a scalar and weight vector, respectively. The parts described above are illustrated in Figure 2, which also shows the support vectors and the hyperplane.

In the hyperplane equation, H_o, we can calculate the largest margin value as follows:

$$\min \frac{1}{2} \| w \|^2 . \tag{2}$$

We can separate the tweets into two classes only when they are linear, limiting the classification of the above linear equation. However, when they become nonlinear, we can modify the SVM equation by adding a slack variable s_i. In other words, we have to add $(s_i \geq 0, \ \forall_i : s_i = 0)$ such

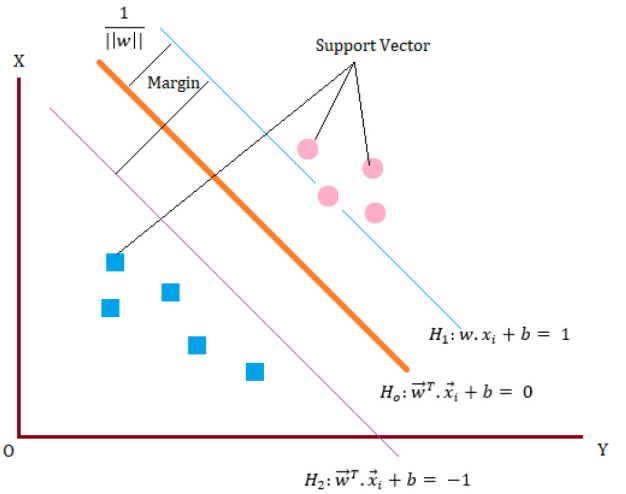

Figure 2

Hyperplane in SVM

that $y_i(w^T.x_i + b) \geq 1 - s_i$ is obtained. Thus, by rewriting Equation (2), we can obtain the following equation:

$$\min \frac{1}{2}\| w \|^2 + C(\sum_{i=1}^{n} s_i). \quad (3)$$

In Equation (3), C plays a crucial role in minimizing errors; however, it is not sufficient, and the problem is further optimized by including the following Lagrange function:

$$\min L_p(w,b,s,\alpha) = \frac{1}{2}\| w \|^2 + C(\sum_{i=1}^{n} s_i) + \sum_{i=1}^{n} \alpha_i (1 - (y_i(w^T \cdot x_i + b) - s_i)). \quad (4)$$

In Equation (4), $a_i \geq 0$, and the main objective of a_i is to minimize L_p, such that it has less impact on the weight vector w and scalar b. Moreover, we can determine the efficiency of SVM in this particular Twitter dataset by using the partial derivative of L_p upon w and b as follows:

$$L_d = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j). \quad (5)$$

3.1.2 Naive Bayes Classification Algorithm

Naive Bayes is a probabilistic learning method based on the assumption that terms occur independently. This learning method is a

classification algorithm for binary and multiclass problems. The technique is easiest to understand when stated with binary or categorical input values. The probabilities are used to represent Naive Bayes; a list of probabilities for a learned Naive Bayes model is saved in a file that includes the following:

Class Probabilities: The chances that each class in the training dataset occurs.

Conditional Probabilities: The conditional probabilities of each input value with each class value.

NBC is a probabilistic classification supervised learning approach based on Bayes theorem and intermediate naive assumptions. The Bayes rules implemented in this study can be expressed as follows:

$$P(B_j \mid A) = \frac{P(B_j)P(A|B_j)}{P(A)} \tag{6}$$

We consider $P(c_i \mid x_1,\ldots,x_n)$ as the conditional probability of class c_i using Equation (6); the word feature is $(x_1,\ldots,x_n)$, and therefore we can rewrite Equation (1) as

$$P(c_i \mid x_1,\ldots,x_n) = \frac{P(x_1,\ldots,x_n|c_i)P(c_i)}{P(x_1,\ldots,x_n)}, \tag{7}$$

where $P(x_1,\ldots,x_n \mid c_i)$ represents the conditional probability that the word features will be $(x_1,\ldots,x_n)$ on the specific class c_i. Furthermore, x_i states the number of words that appear among the collected raw data of the scrapped Twitter on COVID-19. The "Multinomial Naïve Bayes" model was used for the sentiment analysis of the Twitter data in this study. In other words, $(x_1,\ldots,x_n)$ was considered as $(x_1,\ldots,x_n) = 1$. Thus, Equation (7) can now be rewritten as follows:

$$P(c_i \mid x_1,\ldots,x_n) = P(x_1,\ldots,x_n \mid c_i)P(c_i) \tag{8}$$

The posterior probabilities can be calculated based on the likelihood of each class (prior probabilities) and the probability of each word (conditional probabilities) in the training data by simplifying Equation (8), that is,

$$P(c_i \mid x_1,\ldots,x_n) = \prod_{j=1}^{j=n} P(x_j \mid c_i)P(c_i), \tag{9}$$

where $n, i, P(c_i \mid x_1,\ldots,x_n)$, $P(c_i)$, $P(x_j \mid c_i)$, and $x_1,\ldots,x_n$ represent the number of words in the dataset, positive and negative classes, posterior probabilities, prior probabilities, conditional probabilities, and features of the words in the specific Twitter dataset.

We used Equation (4) as the Naïve Bayes model. The testing data from the collected tweets were classified into class (c_i), and these data were specific max posteriori denoted by c_{MAX} as follows:

$$c_{MAX} = \arg\max_{c_i \in C} P(c_i) \prod_{j=1}^{j=n} P(x_j \mid c_i) \tag{10}$$

Prior probability is determined using $P(c_i) = \frac{N_{ci}}{N}$ and the conditional probability using $P(x_j \mid c_i) = \frac{n_j}{n}$. Laplace smoothing was adopted by adding +1 to the word frequency when words from the Tweets fall into neither of the classes. Therefore, the value cannot be absolute zero.

3.2 *Research Methods*

A confusion matrix was implemented to determine the efficiency between the NBC and SVM algorithms in addition to using Equations (1)–(10) as the theoretical basis of the calculation. The confusion matrix is given in Table 1.

Table 1

Confusion matrix

	Actual (True/ False)	
Predicted (Positive/ Negative)	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

Sentiment classification was performed to test the classification results by analyzing the performance of the created system. The evaluation metrics included accuracy, precision, and recall, which are calculated using the confusion matrix table as follows:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \times 100\% \tag{11}$$

$$Precision_{(positive)} = \frac{True\ Positive}{True\ Positive + False\ Positive} \times 100\% \tag{12}$$

$$Recall_{(positive)} = \frac{True\ Positive}{True\ Positive + False\ Negative} \times 100\% \tag{13}$$

$$Precision_{(negative)} = \frac{True\ Negative}{True\ Negative + False\ Negative} \times 100\% \tag{14}$$

$$Recall_{(negative)} = \frac{True\ Positive}{True\ Negative + False\ Poitive} \times 100\% \tag{15}$$

In addition to the theoretical framework and the confusion matrix described above, a descriptive analysis was performed. In this regard, the descriptive analysis conducted in the study is presented using various diagrams, including word cloud, bar diagram, and histograms.

A total of 2000 tweets were collected, cleaned, and stored in a spreadsheet in .csv format. This spreadsheet was imported into Python and R to conduct the descriptive analysis.

4. Results, Analysis, and Inference

4.1 *Connecting, collecting, and tagging data*

We applied and received a developer account from Twitter along with a secret key, consumer key, and OAuth token. An application (python script) was designed to extract data from Twitter. The Python application was created under the title "Python Tweet Sentiment Analyzer." The source code is entirely open-source and custom-made for this study.

Then, through web scrapping, Tweets with specific hashtags "#coronavirus" and "#COVID19" were collected. The number of Tweets is presented in Table 2.

Table 2

Number of tweets

Hashtags	The number of tweets collected
#coronavirus	1000
#COVID19	1000

Including irrelevant characters, such as emojis, URLs, and Twitter usernames, in the aggregated data produced unanticipated outcomes. Before performing the SA, these were cleaned, removed, and labeled because they had no inherent or extrinsic relevance to the study. Only positive and negative data were examined when measuring the efficiency of the NBC and SVM algorithms. The descriptive analysis analyzed eight attitudes to better comprehend public opinion. The principal findings, shown in Figure 3, of this study are as follows:

4.2 *Trust was the dominant sentiment among the public*

During the pandemic, lifestyles were compromised, and with the economy at a standstill and the imposed lockdowns, many people lead unhappy lives. However, in the dataset of Tweets collected in this study,

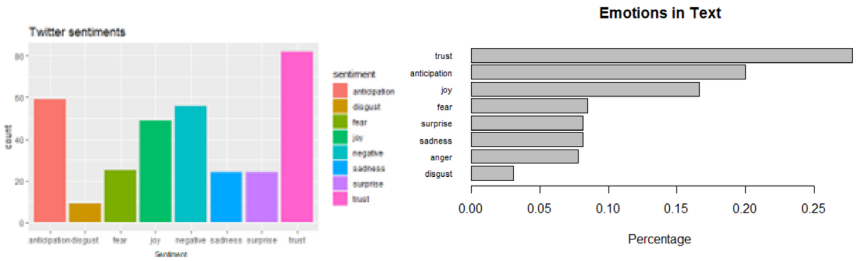

Figure 3

Descriptive analysis of eight attitudes.

trust was the most common emotion among Twitter users, and the least-counted sentiment was *disgust*; in other words, most users viewed innovations like vaccines and supported others with a positive attitude. *Anticipation* was the second-most-common sentiment among Twitter users, indicating the urgency that people wanted a cure for this pandemic and the desperation for the lockdown to end.

4.3 *Most common words used*

Figure 4 shows the list of the most common words used by users. In particular, the most commonly used words in the raw dataset for SA are similar to the first finding. Here, the most common words include *love, live, and life.* This, in turn, indicates a positive sentiment among Twitter users as they attempted to spread positivity among their circles. Some other common words include *travel* and *science,* implying that people were having relevant discussions about travel bans or were desperate to travel. Instances of the word *science* indicate that users talked about vaccination attempts. Furthermore, news about variants was also relevant to users. All these words in their root forms were used while conducting the SA, which helped us determine which algorithm performs better.

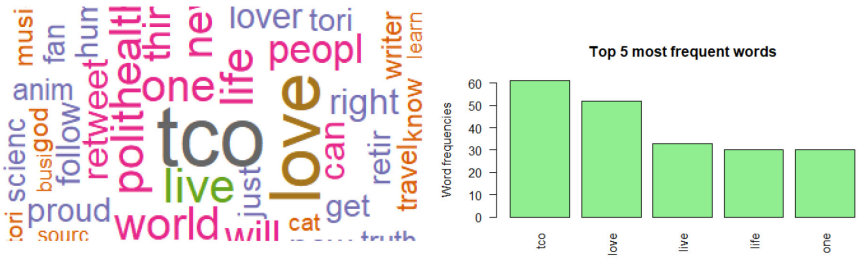

Figure 4

List of most common words.

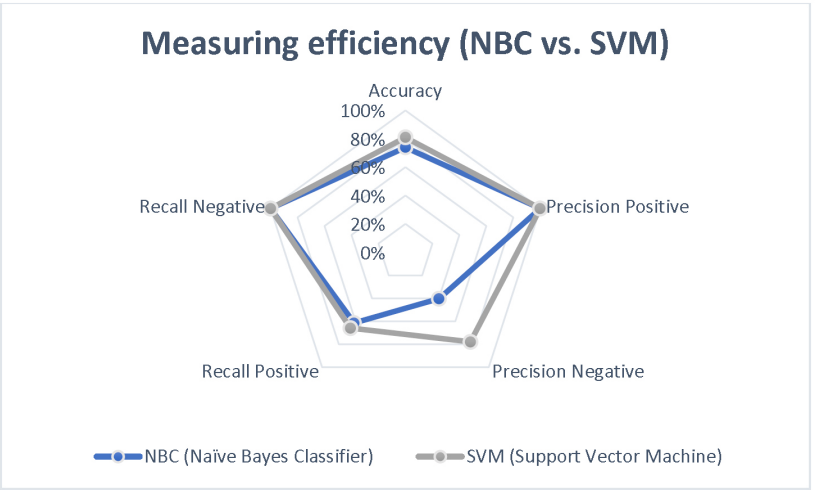

Figure 5

NBC vs. SVM efficiency.

4.4 *Determining Efficiency*

This study's main objective was to compare the efficiencies of NBC and SVM. Thus, among the Tweets, the raw data were cleaned, and for the sake of simplicity, the words were tagged into two distinct classes (*positive and negative*). Subsequently, for evaluation purposes, a confusion matrix was built as indicated by Equations (11)–(15).

Figure 5 shows that the SVM algorithm outperforms the NBC algorithm when dealing with a large amount of data, according to the results in the matrix.

5. Conclusion

SA can be utilized to provide governing bodies and policymakers with an accurate, real-time reflection of public attitude regarding a particular topic. However, errors in interpreting public sentiment during a health crisis could lead to disastrous decision-making. Thus, planners and data analysts should constantly evaluate the most efficient course of action. This study examined the Covid-19 pandemic and determined that the SVM algorithm is superior to the NBC method with the data at hand.

Nevertheless, the NBC method is relatively rapid and relies on rapid conditional probabilities to generate and evaluate. Therefore, an iterative strategy is not required. NBC enables binary and multinomial classification

and considers the independence of characteristics, even though this is not always the case. NBC is effective when applied to brief texts like tweets. NBC might outperform other classifiers by using feature selection on particular datasets.

SVM is a binary model by design; however, it can classify several classes and provide outstanding results. SVM can handle nonlinear classification difficulties. It generalizes well in high-dimensional spaces, such as those corresponding to texts. Moreover, SVM is successful with more dimensions than samples. However, it is effective when the classes are appropriately separated. For massive datasets, the training cost of SVM is a limitation. Training on enormous datasets is time-consuming, and hyperparameter tweaking is challenging. Therefore, theoretically, SVM is more enticing.

Abbreviations

SA- Sentiment Analysis

NBC- Naive Bayes Classification

SVM- Support Vector Machine

References

[1] Bhatnagar, V., Poonia, R. C., Nagar, P., Kumar, S., Singh, V., Raja, L., & Dass, P. Descriptive analysis of COVID-19 patients in the context of India. *Journal of Interdisciplinary Mathematics* (2020), DOI: 10.1080/09720502.2020.1761635. https://doi.org/10.1080/09720502.2020.1761635.

[2] Munjal, P., Narula, M., Kumar, S., & Banati, H. Twitter sentiments based suggestive framework to predict trends. *Journal of Statistics and Management Systems*, 21(4), 685-693 (2018). DOI: 10.1080/09720510.2018.1475079. https://doi.org/10.1080/09720510.2018.1475079.

[3] Singh, D. K., & Dass, P. On a functional equation related to some entropies in information theory. *Journal of Discrete Mathematical Sciences and Cryptography*, 21(3), 713-726 (2018). DOI: 10.1080/09720529.2018.1445809. https://doi.org/10.1080/09720529.2018.1445809.

[4] World Health Organization. (2020). Coronavirus disease (COVID-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[5] Ben, I. (2020). A simple statistical analysis for the spread of COVID-19. https://www.timesofisrael.com/israeli-mathematician-predicts-coronavirus-peaking-soon/.

[6] Mousavizadeh, L., & Ghasemi, S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. *Journal of Microbiology, Immunology, and Infection*, 53(3), 409-412 (2020). DOI: 10.1016/j.jmii.2020.03.022. https://doi.org/10.1016/j.jmii.2020.03.022.

[7] Agarwal A., Xie B., Vovsha I., Rambow O., & Passonneau R. J. Sentiment analysis of Twitter data Proceedings of the workshop on language in social media, p 30–38 (2011).

[8] Bennett K. P., & Blue J. A. *IEEE International Joint Conference on Neural Networks*. A support vector machine approach to decision trees, *3*, 2396–2401. (Cat No. 98CH36227) (1998).

[9] Borah P., Deb P. K., Deka S., Venugopala K. N., Singh V., Mailavaram R. P., Kalia K., & Tekade R. K. (2021). Current scenario and future prospect in the management of COVID-19. *Current Medicinal Chemistry* 27 (September), 28(2), 284–307. https://doi.org/10.2174/0929867327666200908113642

[10] Chen S., Webb G. I., Liu L., & Ma X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*, 105361. https://doi.org/10.1016/j.knosys.2019.105361

[11] Davidson H. (2020). First Covid-19 case happened in November. China government records show–report. *The Guardian*, *13.*

[12] Dey S., Wasif S., Tonmoy D. S., Sultana S., Sarkar J., & Dey M. A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. *IEEE International Conference on Contemporary Computing and Applications (IC3A)*, p 217–220 (2020).

[13] Zhang K., Cheng Y., Liao W. K., & Choudhary A. Mining millions of reviews: A technique to rank products based on importance of reviews *Proceedings of the 13th international conference on* electronics *and* communications, p 1–8 (2011).

[14] Khorsheed M. S., & Al-Thubaity A. O. Comparative evaluation of text classification techniques using a large diverse Arabic dataset.

Language Resources and Evaluation, 47(2), 513–538 (2013). https://doi.org/10.1007/s10579-013-9221-8

[15] Giachanou A., & Crestani F. Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys*, 49(2), 1–41 (2016). https://doi.org/10.1145/2938640.

[16] Go A., Huang L., & Bhayani R. Twitter sentiment analysis. *Entropy*, *17*, 252 (2009).

[17] Jakkula V. Tutorial on support vector machine (SVM). School of EECS, *37*. Washington State University, (p 2.5) : 3 (2006).

Received November, 2022