ISSN 2278-3091
Volume 12, No.1, January - February 2023
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse031212023.pdf
https://doi.org/10.30534/ijatcse/2023/031212023

Comparison of Tiktok User Sentiment Analysis Accuracy with Naïve Bayes and Support Vector Machine

Parasian D.P Silitonga[1], Mitra Hasibuan[2], Z Situmorang[3], Desinta Purba[4]
[1]Information Engineering Study Program, Universitas Katolik Santo Thomas, Indonesia, parasianirene@gmail.com
[2]Information Engineering Study Program, Universitas Katolik Santo Thomas, Indonesia, mitrahasibuan01@gmail.com
[3]Information System Study Program, Universitas Katolik Santo Thomas, Indonesia, zakarias65@yahoo.com
[4]Information Engineering Study Program, Universitas Katolik Santo Thomas, Indonesia, desinta.poerba@yahoo.com

Received Date December 19, 2022 Accepted Date: January 21, 2023 Published Date: February 06, 2023

ABSTRACT

This study aims to compare the accuracy of the sentiment analysis of TikTok application users using the Naïve Bayes algorithm and the Support Vector Machine. The data set in this study comes from comments from Tiktok users on Twitter social media. Comparison of the accuracy of sentiment analysis in this study was carried out through three tests. The first test was conducted on 848 tweets, the second test used 957 tweet data, and the third test used 1,925 tweet data. Testing is done by dividing the data by 70% for training data and 30% for test data. The results showed that the accuracy of the Naive algorithm was 89.35% and 94.08% using the Support Vector Machine algorithm.

Key words : Sentiment Analysis, TikTok, Naïve Bayes, Support Vector Machine, Twitter.

1. INTRODUCTION

In recent years the growth of digital data is increasing, and knowledge discovery and data mining have attracted significant attention with the need to turn that data into useful information and knowledge. The use of information and knowledge extracted from large amounts of data is beneficial for many applications, such as market analysis and business management [1]. In many applications, databases store information in text form, making text mining one of the least desirable areas for research. One implementation of data mining is sentiment analysis.

Sentiment analysis is interpreting and classifying user emotions (positive, negative, or neutral) about a subject in text data using text analysis [2]. Researchers widely use sentiment analysis as a branch of research in computer science. Social networks like Twitter are commonly used in sentiment analysis to determine public perception. Sentiment analysis can also be equated with opinion mining because it focuses on opinions that are stated positively or negatively [3]. With the help of sentiment analysis, unstructured information can be converted into more structured data, which can then be used to explain people's opinions about products, brands, services, politics, or other topics. Companies, governments, and other fields then use these data to conduct marketing analysis, product feedback, and community services [4].

Several methods can be used to perform sentiment analysis, one of which is machine learning. Machine learning is used to produce robots capable of classifying types of sentiment in textual data [5]. Machine learning is a branch of artificial intelligence that can access existing data at its command. Machine learning can study existing data, perform specific tasks, and learning algorithms and statistical models.

A social media application is a computer program made to work on and carry out specific tasks from users that can help users quickly use it. The existence of applications in this era makes it easier for us to communicate with people who are far away and will be close. TikTok is one of the most popular applications today; TikTok users in Indonesia are dominated by teenagers aged 14-24 years. Tiktok users in Indonesia reached 92.2 million users, calculated as of July 2021 [6]. The application allows users to create short music videos. TikTok is not just making videos but can send video results to social media such as Instagram, YouTube, and others that users have made, as well as being able to see the results of videos that other people have made and give likes and comments on videos that users have shared.

The comments and opinions given were positive, and some were negative. Comments and opinions, especially those contained in social media, are a source of data that can be used to measure the popularity of a program or product launched. Support or rejection of a program can be calculated based on comments and public opinion on social media [7].

An application always has its advantages and disadvantages, which can lead to various responses from application users, such as satisfaction and disappointment with the application.

Social media is a place to express user satisfaction and dissatisfaction or opinions about the application. This can be used as material for sentiment analysis for the Tiktok application. This study compares the results of sentiment analysis accuracy of the satisfaction of users of the TikTok application using the Naïve Bayes algorithm and the support vector machine.

2. METHODS

The research stages in writing this research are shown in figure 1 below :

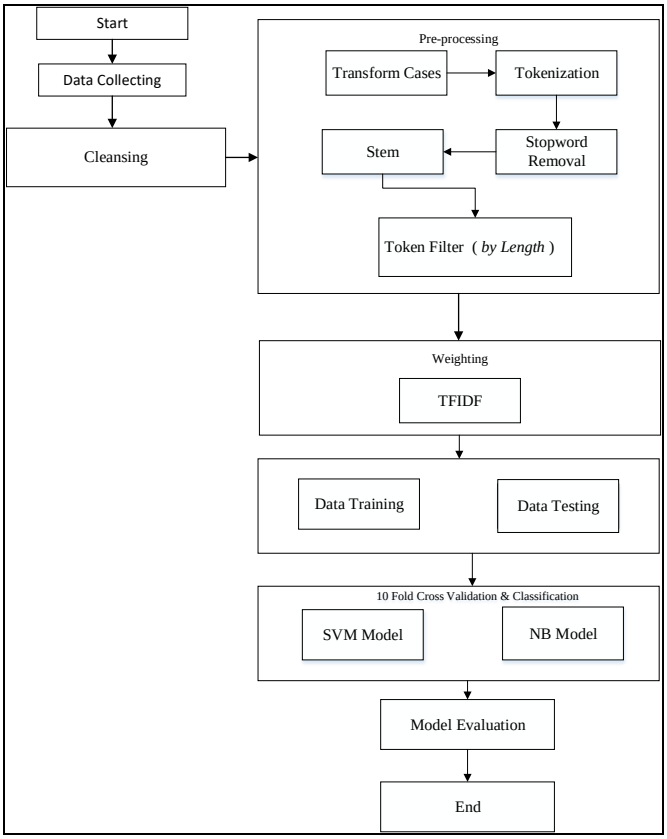

Figure 1 : Research stages

A. *Data Collecting*

This process takes reviews from the Twitter application. The input data to be used comes from @tiktok tweets. In this case, the review that will be taken is a review of the Twitter application that is entered in the CSV format. The amount of review data is total - data consisting of positive and negative comments.

B. *Data Cleansing.*

The cleansing stage is the word cleaning stage, which does not affect the sentiment classification results [8]. The tweet document component has various attributes that do not affect sentiment because almost every tweet has these attributes. Examples of unimportant attributes are mentioned starting with attributes ('@'), hashtags starting with attributes ('#'), links starting with attributes ('http','bit.ly') and symbol characters (~!@#$%^&*()_+?<>,.?:{}[]|). The attributes that have no effect will be removed from the document and replaced with a space character.

C. *TF-IDF.*

Term Frequency-Inverse Document Frequency (TF-IDF) weighting is transforming data from textual data into numeric data for each word or feature to be weighted [9]. TF-IDF is a statistical measure used to evaluate a word's importance in a document. TF is the frequency of occurrence of a word in each given document indicating how important that word is in each of those documents [10]. DF is the frequency of documents containing the word's meaning and the word's prevalence. IDF is the inverse of the DF value. The result of word weighting using TF-IDF is the multiplication result of TF multiplied by IDF [11]. The word weight is more significant if it frequently appears in a document and is smaller if it appears in many documents [12]. Figure 2 is an illustration of the TF-IDF process.

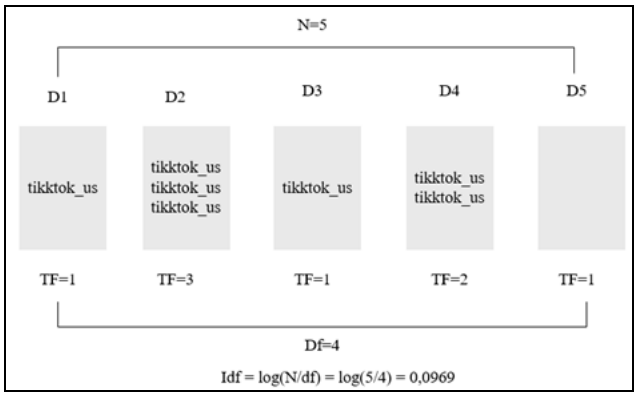

Figure 2 : Illustration of the TF-IDF Algorithm

where : D1…D5 = document
TF = the number of words in each document
N = document totals
Df = the number of documents on the word searched

The formula for TF-IDF word weighting is :

$$W_{t,d} = tf_{t,d} \; X \; idf_t = \; tf_{t,d} \; X \; \log\frac{N}{df_t} \qquad (1)$$

where : $W_{t,d}$ = TF-IDF weight
$tf_{t,d}$ = number of word frequencies
idf_t = number of document frequency inverse per word
df_t = number of document frequency per word
N = total number of documents

The result of weighting words with TF-IDF is the multiplication of the TF and IDF values, resulting in a smaller weight if the word frequently appears in each document in the collection. On the other hand, the TF-IDF weight will be greater if the word rarely appears in every document in the group. In this study, the TF-IDF weighting used is TF-IDF without normalization [13].

D. Naïve Bayes

Naïve Bayes is a simple probability classification method that applies Bayes' Theorem with a high independence assumption [14]. The use of the Naive Bayes method in this study is based on a large number of datasets used, so it requires a technique that has a fast performance in classification and high accuracy [15].

The Naive Bayes method takes two stages in the text classification process: the training and testing stages. The training process is used for the sentiment analysis model, which aims as a classification guide with testing data or different data. The calculation of the comparison between the terms in the data testing with each existing class can be done with equation 2 [16].

$$P(w_i|C) = \frac{count(w_i,c)+1}{count(C)+|V|} \quad (2)$$

where : C = class category tested
d = document
w_i = the i word
w(i,c) = the number of words wi in C
count (C) = word in C class
|V| = number of word

In this equation, there is an additional 1 in the numerator to avoid a 0 value in the probability if there is a word in the test document that has a zero value because it is not in the training document.

E. Support Vector Machine

Support Vector Machine (SVM) is a classification algorithm to find a separator function that can separate two sets of data from two different classes [17]. SVM has a hyperplane that separates the two classes of data by as wide a margin as possible, leading to good generalization accuracy on unseen data and supporting unique optimization methods that enable SVM to learn from large amounts of data [18].

The basic principle of SVM is a linear classifier. It is further developed to work on on-linear problems by incorporating the kernel trick concept in high-dimensional workspaces [17]. Support vector machine (SVM) is a classification method using machine learning (supervised learning) that predicts classes based on models or patterns from the training process results [19]. Classification is done by looking for a hyperplane or boundary line (decision boundary) that separates one class from another class, in which case the bar plays a role in separating data from different types or with the positive sentiment (labeled +) from data with a negative view (labeled -) [20].

The best-dividing hyperplane between two classes can be found by measuring the hyperplane margin and finding its maximum point. Margin is the distance between the hyperplane and the nearest data from each category. The closest subset of the training data set is called the carrier vector. Efforts to find the optimal hyperplane location form the basis of the learning process at SVM [21].

3. RESULT AND DISCUSSION

A. Data Source

The data source used in this study is text data crawled from Twitter social media by using the Twitter search attribute for those who have done the Twitter Connecting API to get an access token select attributes. The access token is used to select data that is crawled only text comments, text that has been crawled with the name 'data crawl tiktok.csv.' The implementation of the data collection process model is presented in the figure 3.

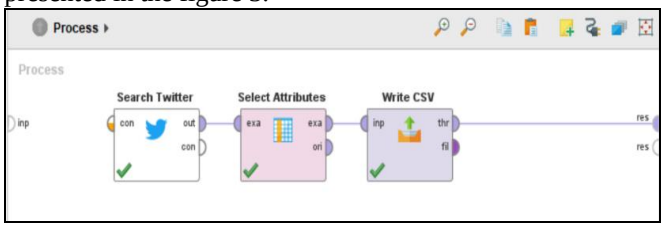

Figure 3 : Data crawling using the rapidminer tool

B. Data Cleaning

The following process is the cleansing stage; in this stage, there are several activity steps, namely cleaning the data, which will later be used in the data modeling stage. These steps include deleting expressions, removing the at sign (@), and deleting numbers, HTTPS addresses and hashtags. The cleansing process model for this cleansing process is shown in figure 4.

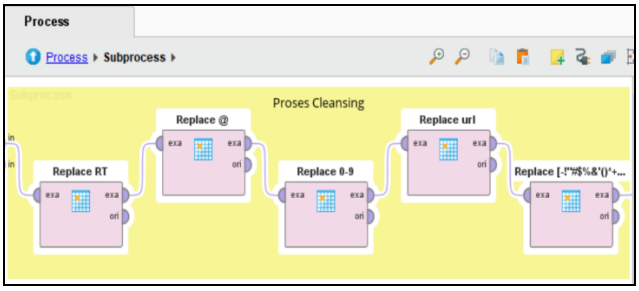

Figure 4: Data cleaning using rapidminer tool

C. Modelling

The following results of data modeling labeled positive and negative using the Naïve Bayes and Support Vector Machine algorithms are presented in figure 5.

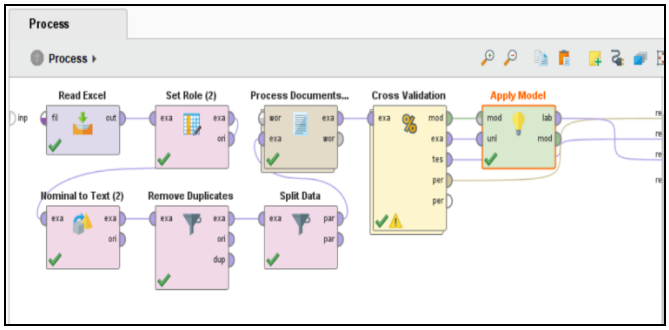

Figure 5 : Naïve Bayes and Support Vector Machine Modelling

D. Evaluation

The following process is to evaluate by comparing the two methods being analyzed, namely the Naïve Bayes and Support Vector Machine algorithms in the form of accuracy. This evaluation is carried out to find out the benefits of the model made in the previous stage. The review was carried out using a confusion matrix, namely the false positive rate (FP rate), false negative rate (FN rate), actual positive rate (TP rate) and valid negative rate (TN rate) as indicators. Based on the results of the first test of the Naïve Bayes model, the accuracy value is shown in Figure 6. The actual positive rate is 266 records, categorized as positive labels, and the false positive rate is 246, categorized as negative. Then the actual negative rate is 236 records categorized as negative labels, and the false negative rate is 100 records classified as positive. Figure 7 shows the accuracy level of the Naïve Bayes algorithm is 59.20%.

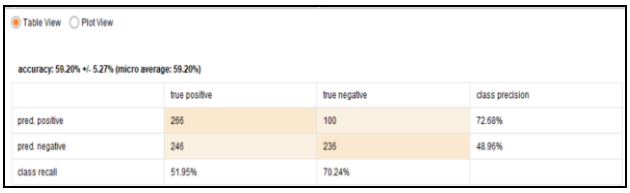
Table View Plot View

accuracy: 59.20% +/- 5.27% (micro average: 59.20%)

	true positive	true negative	class precision
pred. positive	266	100	72.68%
pred. negative	246	236	48.96%
class recall	51.95%	70.24%	

Figure 6 : The accuracy value of the Naïve Bayes algorithm in the first test

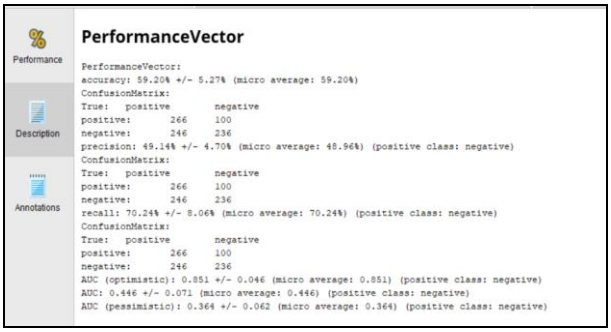

Figure 7 : Performance vector of the Naïve Bayes algorithm in the first test

The level of accuracy of sentiment analysis is measured using equation 3.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100 \qquad (3)$$

so that the accuracy level is obtained :

$$= \frac{266+236}{266+236+246+100} * 100$$
$$= \frac{502}{848} * 100$$
$$= 59{,}20\%$$

Based on the results of the first test of the Support Vector Machine model, the accuracy value is found in figure 8, showing the number of true positive rates is 492 records categorized as positive labels and false positive rates are 20 records classified as negative labels. Then the true negative rate is 89 records categorized as negative labels, and the false negative rate is 247 records categorized as positive labels. Figure 9 shows the accuracy level of the SVM algorithm is 68.51%.

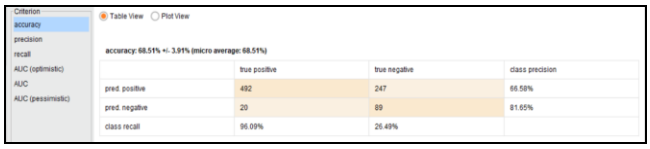
Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)

Table View Plot View

accuracy: 68.51% +/- 3.91% (micro average: 68.51%)

	true positive	true negative	class precision
pred. positive	492	247	66.58%
pred. negative	20	89	81.65%
class recall	96.09%	26.49%	

Figure 8 : Accuracy value of the Support Vector Machine algorithm in the first test

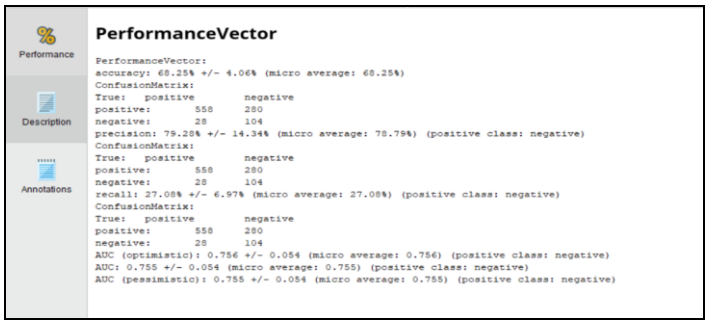

Figure 9 : Performance vector of the Support Vector Machine Algorithm in the first test

By equation 3, the level of accuracy obtained in the first test is equal to :

$$\text{Accuracy} = \frac{492+89}{494+81+18+255} * 100$$
$$= \frac{581}{848} * 100$$
$$= 68{,}51\%$$

After testing three times for each algorithm with a different amount of data, the accuracy results are obtained, as shown in table 1.

Table 1 : Results of model testing accuracy

Testing	Amount of Data	Level of accuracy	
		Naïve Bayes	Support Vector Machine
1st	848 tweets	59,20%	68,51%
2nd	957 tweets	66,57%	74,50%
3rd	1.925 tweets	89,35%	94,08%

5. CONCLUSION

Based on the results of crawling the satisfaction of Twitter social media users with the TikTok application, there were 3,730 tweets with positive sentiments of 2,977 tweets and negative emotions of 797 tweets which were divided into three tests, with each test divided into two types of data with a ratio of 70% for 30 training data. % for test data with Naïve Bayes and Support Vector Machine algorithms. Through the preprocessing and modeling process and the evaluation process for categorizing positive and negative labeled reviews, it can be seen that the final result of the Naïve Bayes accuracy value in the first test was 59.20%. The Support Vector Machine was 68.51%, and the Naïve Bayes accuracy in the second test was 66.57 %. The Support Vector Machine algorithm has a better level of accuracy when compared to Naïve Bayes, with an advantage rate of 9.31% in the first test, 7.93% in the second test and 4.73% in the third test.

REFERENCES

[1] B. Chen, H. He, and J. Guo, "Language feature mining for document subjectivity analysis," in *Proceedings of the 1st International Symposium on Data, Privacy, and E-Commerce, ISDPE 2007*, 2007, pp. 62–67. doi: 10.1109/ISDPE.2007.105.

[2] M. A. Sghaier and M. Zrigui, "Sentiment analysis for Arabic e-commerce websites," 2016. doi: 10.1109/ICEMIS.2016.7745323.

[3] H. U. Khan and D. Peacock, "Possible effects of emoticon and emoji on sentiment analysis web services of work organisations," *Int. J. Work Organ. Emot.*, vol. 10, no. 2, pp. 130–161, 2019, doi: 10.1504/IJWOE.2019.104297.

[4] A. Y. L. Chong, B. Li, E. W. T. Ngai, E. Ch'ng, and F. Lee, "Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach," *Int. J. Oper. Prod. Manag.*, vol. 36, no. 4, pp. 358–383, 2016, doi: 10.1108/IJOPM-03-2015-0151.

[5] J. R. Saura, A. Reyes-Menendez, and P. Palos-Sanchez, "A feeling analysis in Twitter with machine learning: Capturing sentiment from #BlackFriday offers ," *Espacios*, vol. 39, no. 42, 2018, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055057928&partnerID=40&md5=830513edad8b62fe8adf0de560d53a07

[6] M. Yao, "Examination of Underlying Factors in Success of TikTok," 2021. doi: 10.2991/aebmr.k.210601.051.

[7] W. Kaswidjanti, H. Himawan, and P. D. P. Silitonga, "The accuracy comparison of social media sentiment analysis using lexicon based and support vector machine on souvenir recommendations," *Test Eng. Manag.*, vol. 82, no. 3–4, pp. 3953–3961, 2020.

[8] H. Sharma, A. Tandon, P. K. Kapur, and A. G. Aggarwal, "Ranking hotels using aspect ratings based sentiment classification and interval-valued neutrosophic TOPSIS," *Int. J. Syst. Assur. Eng. Manag.*, vol. 10, no. 5, pp. 973–983, 2019, doi: 10.1007/s13198-019-00827-4.

[9] M. Al Omari, M. Al-Hajj, N. Hammami, and A. Sabra, "Sentiment classifier: Logistic regression for Arabic services' reviews in Lebanon," 2019. doi: 10.1109/ICCISci.2019.8716394.

[10] X. Liu and Y. Wang, "Experiment and comparision on classification of chinese car reviews," *Frontiers in Artificial Intelligence and Applications*, vol. 303. School of Information Management and Engineering, Shanghai University of Finance and Economics, China, pp. 810–821, 2018. doi: 10.3233/978-1-61499-900-3-810.

[11] J. Hu, X. Zhang, Y. Yang, Y. Liu, and X. Chen, "New doctors ranking system based on VIKOR method," *Int. Trans. Oper. Res.*, vol. 27, no. 2, pp. 1236–1261, 2020, doi: 10.1111/itor.12569.

[12] M. A. Razzaqe and T. Basak, "Text mining in unstructured text : techniques , methods and analysis," vol. 174, no. October, pp. 68–84, 2022.

[13] S. S. and P. K.V., "Sentiment analysis of malayalam tweets using machine learning techniques," *ICT Express*, no. xxxx, pp. 2–7, 2020, doi: 10.1016/j.icte.2020.04.003.

[14] V. Reshma and A. John, "Aspect based summarization of reviews using naïve Bayesian classifier and fuzzy logic," in *2015 International Conference on Control, Communication and Computing India, ICCC 2015*, 2016, pp. 617–621. doi: 10.1109/ICCC.2015.7432970.

[15] R. Sanda, Z. K. A. Baizal, and F. Nhita, "Opinion mining feature-level using Naive Bayes and feature extraction based analysis dependencies," in *AIP Conference Proceedings*, 2015, vol. 1692. doi: 10.1063/1.4936448.

[16] I. Perikos and I. Hatzilygeroudis, "Aspect based sentiment analysis in social media with classifier ensembles," in *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, 2017, pp. 273–278. doi: 10.1109/ICIS.2017.7960005.

[17] X. Wen, J. Chen, and L. Zhao, "Application of cost-sensitive hybrid-kernel support vector machine (Svm) to sentiment analysis," *Int. J. Simul. Syst. Sci. Technol.*, vol. 17, no. 42, pp. 59.1-59.7, 2016, doi: 10.5013/IJSSST.a.17.42.59.

[18] J. Gawade and L. Parthiban, "Opinion mining of amazon product data by hybrid svm," *J. Adv. Res. Dyn. Control Syst.*, vol. 10, no. 13, pp. 2034–2041, 2018, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067431221&partnerID=40&md5=3fa09cc3e16df8bbea00e10ab3648c61

[19] O. Abdelwahab, M. Bahgat, C. J. Lowrance, and A. Elmaghraby, "Effect of training set size on SVM and Naïve Bayes for Twitter sentiment analysis," in *2015 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2015*, 2016, pp. 46–51. doi: 10.1109/ISSPIT.2015.7394379.

[20] J. Gawade and L. Parthiban, "Opinion mining on product data using modified SVM," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 3, pp. 1829–1837, 2019, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067514619&partnerID=40&md5=090e10f7fcffe80d38fcaa3bc4b87fed

[21] F. Noor, M. Bakhtyar, and J. Baber, "Sentiment Analysis in E-commerce Using SVM on Roman Urdu Text," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 285. Department of Computer Science and Information Technology, University of Balochistan, Quetta, Pakistan, pp. 213–222, 2019. doi: 10.1007/978-3-030-23943-5_16.