Indonesian Journal of Data and Science

Volume 5 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i2.82

Research Article

Comparative Analysis of Machine Learning Algorithm Variations in Classifying Body Shaming Topics on Social Media X

Sarah Fila Nurul Fitri H [1,*]; Farniwati Fattah [2]; Huzain Azis [3]

[1] Universitas Muslim Indonesia, Jalan Urip Sumoharjo, Makassar, 90231, Indonesia, sarahfilanurulfitrih@gmail.com
[2] Universitas Muslim Indonesia, Jalan Urip Sumoharjo, Makassar, 90231, Indonesia, farniwati.fattah@umi.ac.id
[3] Universitas Muslim Indonesia, Jalan Urip Sumoharjo, Makassar, 90231, Indonesia, huzain.azis@umi.ac.id

Correspondence should be addressed to Sarah Fila Nurul Fitri H; sarahfilanurulfitrih@gmail.com

Received 30 April 2024; Accepted 03 June 2024 2023; Published 31 July 2024

Abstract:

Machine learning is an approach in computer science where systems or models can learn from data and experience to improve performance or perform specific tasks. There are several popular machine learning algorithms, such as naïve bayes, decision tree, K-NN, and SVM. This study aims to compare the performance of accuracy, precision, recall, and F-1 score in sentiment analysis of body shaming topics on Social Media X (formerly known as Twitter) by applying decision tree, K-NN, and SVM methods and identifying the most effective algorithm in classifying the data. Based on the classification performance testing results, it can be concluded that the classification method using the trigram feature model provides the best performance compared to other methods. The trigram model is able to achieve high recall, particularly in recognizing positive classes, without significantly compromising accuracy.

Keywords: Machine Learning, Body Shaming, Decision Tree, K-Nearest Neighbor, Support Vector Machine.
Dataset link: -

1. Introduction

Machine learning is a branch of Computer Science related to Artificial Intelligence, primarily focusing on the development and study of systems designed to learn from data they receive. Machine learning cannot operate without data mining [1]. In recent studies, it has been found that machine learning has three main categories: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Supervised learning is further divided into two types of problems: classification and regression. Classification is used when the output variable is categorical, such as red or blue, or disease and no disease. Meanwhile, regression is used when the output variable is a real value, such as dollars or weight. Several popular algorithms in supervised learning include Naïve Bayes, Decision Tree, Support Vector Machine, and K-Nearest Neighbor [2].

The Naïve Bayes Classifier algorithm has been widely used in previous research for sentiment analysis classification [3]. The advantage of using the Naïve Bayes classifier is that this method only requires a small amount of training data to determine the parameter estimates needed for the classification process [4]. Research conducted by St. Fajriah Fattah and Purnawansyah (2022) on "Sentiment Analysis of Body Shaming on Twitter Using the Naïve Bayes Classifier Method" found that the accuracy result was 68%, precision 66%, recall 68%, and f-measure 66% [5].

Decision Tree is a well-known and effective method for prediction and classification [6]. Besides its relatively fast construction process, the resulting model is also easy to understand [4]. Research conducted by A. Q. Surbakti et al., titled "Analysis of Responses to PSBB in Indonesia Using the Decision Tree Algorithm on Twitter," found that testing results using the decision tree algorithm showed an accuracy value of 84.78%, precision of 84.78%, and recall of 100% [7].

K-Nearest Neighbor (K-NN or KNN) is a method for classifying objects by using learning data (neighbors) that are closest in distance to the object being classified [8]. K-NN's performance as an algorithm for classifying data is quite good. In research conducted by J. A. Septian et al. on "Sentiment Analysis of Twitter Users Towards the Polemic of Indonesian Football Using TF-IDF Weighting and K-Nearest Neighbor," experiments were conducted using 2000 tweet data to find the K-NN model with the highest accuracy within the k=1 to k=30 odd values range. It was found that optimal accuracy occurred at k=23, reaching an accuracy level of 79.99% [9].

Support Vector Machine (SVM) is a set of learning techniques that analyze data and identify patterns [10]. The SVM method is one of the most superior methods in classification, prediction, and regression analysis. In research conducted by Yoel Julianto et al. on "Sentiment Analysis of Restaurant Reviews Using the Support Vector Machine Method," testing results using the SVM method found an accuracy of 93% and an F-1 score of 93% [11].

In this study, we will compare the classification performance of tweets on the topic of body shaming on the social media platform X (formerly Twitter) using three different algorithms: decision tree, K-NN, and SVM. This study aims to evaluate and compare the performance of the three algorithms in solving similar problems. To achieve this goal, the authors will use a confusion matrix to obtain values for accuracy, precision, recall, and F-1 score. This research is a continuation of previous research conducted by St. Fajriah Fattah and Purnawansyah, who had previously used the Naïve Bayes classifier method.

2. Method:

a. Research Stages

The steps in this research involve a series of phases that will be undertaken throughout the research process, with these steps based on the research methods previously explained.

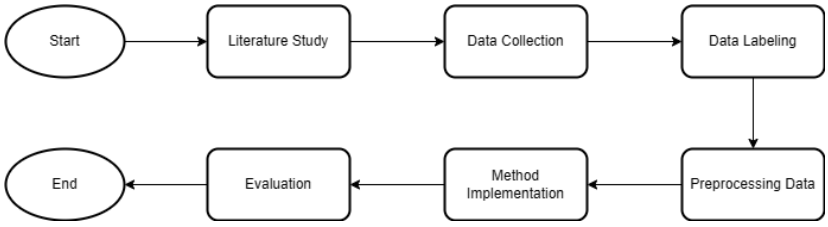

Figure 1. Research Stages

b. Data Collection

In this study, data were collected from tweets on the social media platform X. The data collection process was conducted by searching for keywords related to body shaming such as "*cungkring*" (skinny), "*botak*" (bald), "*tepos*" (flat), and similar words. The dataset used is secondary data, where the dataset was directly collected by students of the Informatics Engineering program under the name St. Fajriah Fattah and stored through the Kaggle Repository, with a total of 907 body shaming tweets.

c. Data Labelling

The data labelling process was carried out manually by reading each tweet individually, then assigning labels.

d. Pre-processing

Text pre-processing is a process that involves preparing raw text for further processing [12]. The steps in pre-processing are as follows:

i. Cleaning

Cleaning is the process of removing unnecessary attributes from the text. The goal is to reduce noise in the dataset. Examples of characters that are removed include punctuation marks such as periods (.), commas (,), and other punctuation marks [13].

ii. Case Folding

Case folding involves converting all letters in the document to lowercase. Only the letters 'a' to 'z' are accepted. Characters other than letters are considered delimiters.

iii. Tokenizing

Tokenizing involves splitting the input string based on the words that compose it. It is the process of breaking sentences into words.

iv. Stemming

Stemming is the process of reducing filtered words to their root form. For example, the word "tuju" (goal) with the prefix "di-" and the suffix "-kan" will be reduced to its base form.

v. Stopwords

Stopwords are common words that frequently appear in the text and are not considered to have important meaning. Stopwords removal is the process of eliminating these words. Examples of stopwords include "dari" (from), "di-" (in/on/at), "yang" (which), and so on.

e. Feature Model

Feature Model is a method for extracting data and generating relevant features from the tweets. These features consist of a set of words that will be used as a reference in the subsequent sentiment analysis classification process. In this study, TF-IDF and trigram are applied. The TF-IDF model works by assigning each feature a weight that represents the frequency of the word in the text relative to the entire document. Meanwhile, the trigram model breaks down each sentence in the tweets into trigrams. In this case, n=3 indicates that each n-gram consists of three words.

f. Implements Method Decision Tree, K-NN dan SVM

At this stage, the classification process is carried out using three algorithms: decision tree, K-NN, and SVM through the implementation of Cross Validation, where the dataset will be divided into two parts: training data and testing data. The method used is K-fold cross validation with k=4, which means there will be four iterations for training and testing. Each iteration consists of 75% training data and 25% testing data. During each iteration, the testing data portion will alternate so that all parts of the dataset will serve as testing data.

g. Research Design

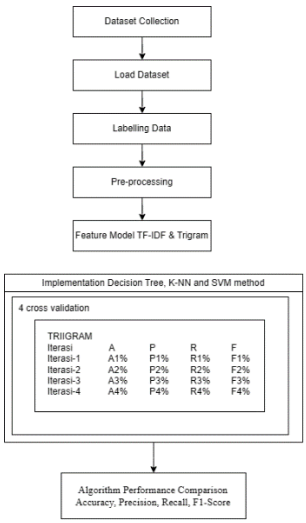

Figure 2. Research Results

3. Results and Discussion

a. Results

1) Data Collection

The data was collected by previous researchers through the website http://www.kaggle.com/ and stored in .csv format. This dataset contains tweet data from the social media platform Twitter. The data was collected based on words related to body shaming, such as "*cungkring*" (skinny), "*botak*" (bald), "*burik*" (blemished), "*buncit*" (pot-bellied), "*gendutan*" (gaining weight), "*jelek*" (ugly), "*pesek*" (flat-nosed), and others. The list of tweets is shown in **Table 1**.

Table 1. Body Shaming Tweets Data

Id	Tweets
0	*Dh panjang eh rambut kau saddiq, drpd kau botak licin sampai rambut dh panjang pon kita masih pkp lagi. Nnti kau dh ada ub…*
1	*jdi kek botak*
3	*fleet gk ada jdi aneh gitu deh, botak tl nya plss AHAHHAHA*
4	*Si botak lagi ngigau ?? https://t.co/25H4o8q8OC*

2) Data Labelling

After collecting the data, labelling was performed to separate the data into positive and negative classes. The labelling process was done manually by reading each tweet individually and assigning a label to each piece of data. The labelled tweet data is shown in **Table 2**.

Table 2. Tweets with Manual Labelling

Id	Tweets	Class	Label
0	*Dh panjang eh rambut kau saddiq, drpd kau botak licin sampai rambut dh panjang pon kita masih pkp lagi. Nnti kau dh ada ub…*	1	Positif
1	*jdi kek botak*	0	Negatif
3	*fleet gk ada jdi aneh gitu deh, botak tl nya plss AHAHHAHA*	1	Positif
4	*Si botak lagi ngigau ?? https://t.co/25H4o8q8OC*	1	Positif

3) Data Pre-processing Implementation

a) Cleaning

Table 3. Tweets After Cleaning

No	Tweets
1	*Dh panjang eh rambut kau saddiq drpd kau botak licin sampai rambut dh panjang pon kita masih pkp lagi Nnti kau dh ada ub*
2	*jdi kek botak*
3	*fleet gk ada jdi aneh gitu deh botak tl nya plss AHAHHAHA*
4	*Si botak lagi ngigau*

b) Case Folding

Case Folding is the process of converting text to lowercase, removing unnecessary elements such as links, emoticons, numbers, and punctuation, and eliminating white spaces or excessive spacing. The tweet data after the case folding stage is shown in **Table 4**.

Table 4. Tweets After Case Folding

No	Tweets
1	*Dh panjang eh rambut kau saddiq drpd kau botak licin sampai rambut dh panjang pon kita masih pkp lagi Nnti kau dh ada ub*
2	*jdi kek botak*
3	*fleet gk ada jdi aneh gitu deh botak tl nya plss ahahhaha*
4	*si botak lagi ngigau*

c) Tokenizing

Tokenizing is the process of separating words from the original sentence, which is then referred to as tokens or terms. The tokenizing used in this study is trigram, where each token consists of only three words. The tweet data after the tokenizing stage is shown in **Table 5**.

Table 5. Tweets Setelah Tokenizing

No	Tweets
1	*['dh', 'panjang', 'eh', 'rambut', 'kau', 'saddiq', 'drpd', 'kau','botak', 'licin', 'sampai', 'rambut', 'dh', 'panjang', 'pon', 'kita', 'masih', 'pkp', 'lagi', 'nnti', 'kau', 'dh', 'ada', 'ub']*
2	*['jdi', 'kek', 'botak']*
3	*['fleet', 'gk', 'ada', 'jdi', 'aneh', 'gitu', 'deh', 'botak', 'tl', 'nya', 'plss', 'ahahhaha']*
4	*['si', 'botak', 'ngigau']*

d) Stopwords

Stopwords are a set of words that frequently appear in the text but have little relevance to the text's context. These words are removed because they do not have specific meaning and do not significantly affect system analysis. The tweet data after the stopwords removal stage is shown in **Table 6**.

Table 6. Tweets After Stopwords Removal

No	Tweets
1	*['dh', 'eh', 'rambut', 'kau', 'saddiq', 'drpd', 'kau', 'botak', 'licin', 'rambut', 'dh', 'pon', 'pkp', 'nnti', 'kau', 'dh', 'ub']*
2	*['jdi', 'kek', 'botak']*
3	*['fleet', 'gk', 'jdi', 'aneh', 'gitu', 'deh', 'botak', 'tl', 'nya', 'plss', 'ahahhaha']*
4	*['si', 'botak', 'ngigau']*

e) Stemming

Stemming is the process of reducing words with affixes to their root form. The tweet data after the stemming stage is shown in **Table 7**.

Table 7. Tweets After Stemming

No	Tweets
1	*['dh', 'rambut', 'saddiq', 'drpd', 'botak', 'licin', 'rambut', 'dh', 'pkp', 'nnti', 'dh', 'ub']*
2	*['jdi', 'kek', 'botak']*
3	*['gk', 'jdi', 'aneh', 'gitu', 'deh', 'botak', 'tl', 'nya', 'plss', 'ahahhaha']*
4	*['botak', 'ngigau']*

4) Feature Models

The feature modelling stage in text classification serves to transform unstructured text format into structured format, enabling machine learning algorithms to classify text into predetermined classes.

a) TF-IDF

The TF-IDF process aims to convert the unstructured text representation into a structured representation with numerical values. The feature model based on TF-IDF is shown in **Table 8**.

Table 8. Tweet Data Resulting from TF-IDF

Dokumen	***Abang***	***Abis***	***Aki***	***…***	***Wajah***	***Wkwk***	***Yaudah***
Dokumen 3	0.0	0.0	0.0	…	0.480447	0.0	0.0
Dokumen 5	0.0	0.0	0.0	…	0.0	0.268778	0.0
Dokumen 20	0.240185	0.0	0.0	…	0.0	0.0	0.0
Dokumen 52	0.0	0.284769	0.0	…	0.0	0.0	0.0

b) Trigram

Trigram aims to improve text representation by considering the combination of three consecutive words as a feature for more complex text analysis. The feature model based on trigram is shown in **Table 9**.

Table 9. Tweet Data After Trigram

Tweets	***Frequency***
milik gigi tonggos	4
pake louis vuitton	3
louis vuitton perut	3
vuitton perut buncit	3

5) Classification Algorithm Implementation

Based on the performance testing of the three classification methods, namely decision tree, K-NN, and SVM, on body shaming tweet data, and testing the accuracy, precision, recall, and F-1 score performance, the testing results will be clearly explained as follows:

a) Decision Tree

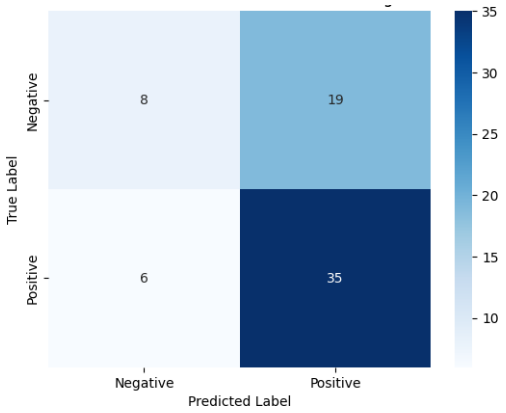

Figure 3. Confusion Matrix of Decision Tree with TF-IDF Feature Model

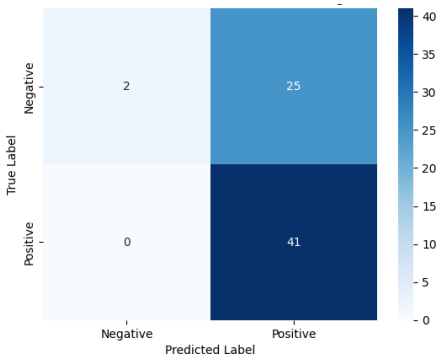

Figure 4. Confusion Matrix of Decision Tree with Trigram Feature Model

b) K-NN

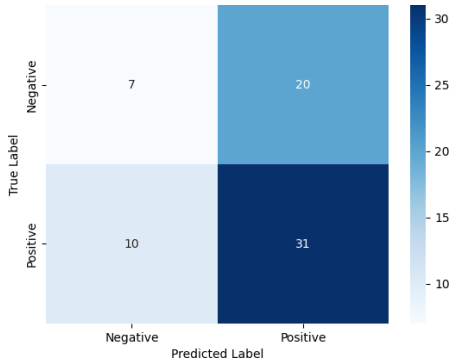

Figure 5. Confusion Matrix of Decision Tree with TF-IDF Feature Model

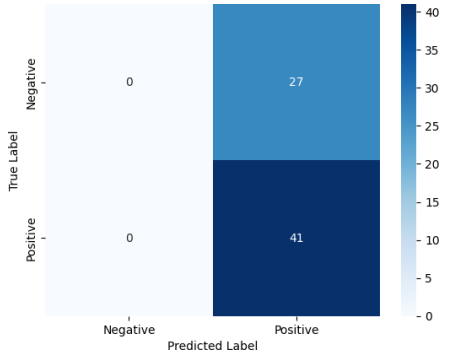

Figure 6. Confusion Matrix of K-NN with Trigram Feature Model

c) SVM

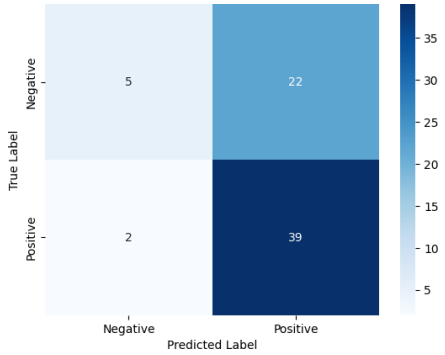

Figure 7. Confusion Matrix of SVM with TF-IDF Feature Model

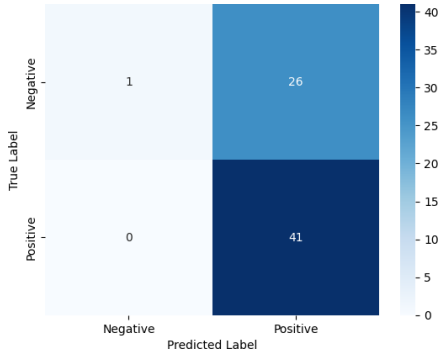

Figure 8. Confusion Matrix of SVM with Trigram Feature Model

6) Cross Validation Implementation

a) Decision tree

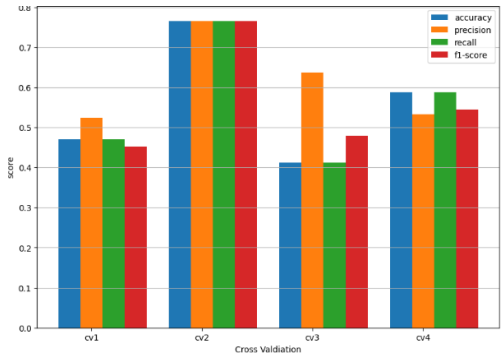

Figure 9. Cross Validation Graph Testing for Decision Tree with TF-IDF

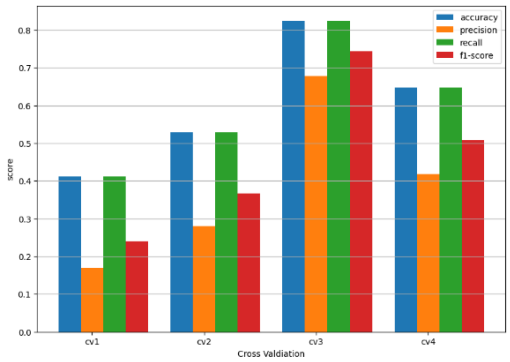

Figure 10. Cross Validation Graph Testing for Decision Tree with Trigram

b) K-NN

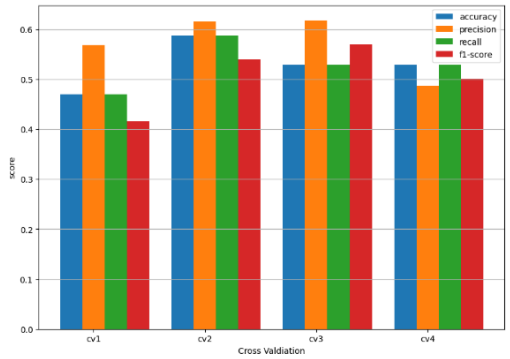

Figure 11. Cross Validation Graph Testing for K-NN with TF-IDF

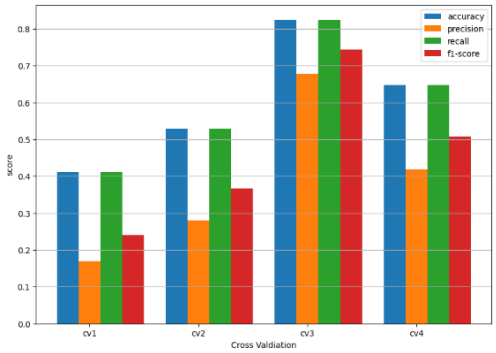

Figure 12. Cross Validation Graph Testing for K-NN with Trigram

c) SVM

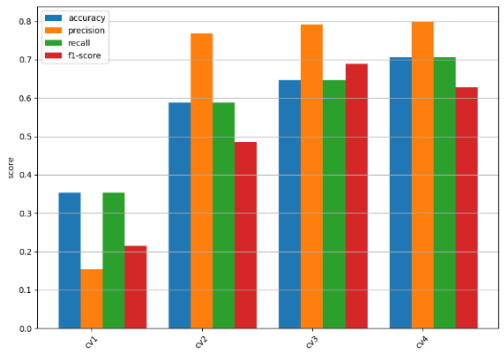

Figure 13. Cross Validation Graph Testing for SVM with TF-IDF

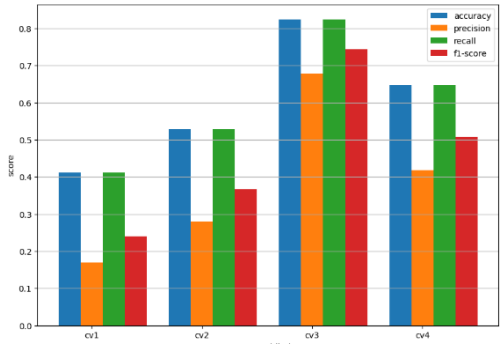

Figure 14. Cross Validation Graph Testing for SVM with Trigram

7) Algorithm Performance Comparison

Table 10. Algorithm Performance Comparison

Model	*Feature*	*Accuracy*	*Precision*	*Recall*	*F-1 Score*
Decision Tree	TF-IDF	63%	65%	85%	74%
	Trigram	63%	62%	100%	77%
K-NN	TF-IDF	56%	61%	76%	67%
	Trigram	60%	60%	100%	75%
SVM	TF-IDF	65%	64%	95%	76%
	Trigram	62%	61%	100%	76%

b. Discussion

The dataset used in this research was obtained from previous researchers and stored on Kaggle. The dataset consisted of 907 samples, but after going through the pre-processing stage, the number was reduced to 678 due to the removal of duplicate tweet data.

The research results showed that the accuracy of the classification methods (decision tree, K-NN, and SVM) was low. This was due to the limited amount of data (only 678 samples after pre-processing), poor data quality due to typos and slang (English), and class imbalance, with the sentiment distribution percentages being 65.1% (positive) and 34.9% (negative)

4. Conclusion

Based on the performance testing results of classification using the decision tree, K-NN, and SVM methods with two types of feature models, namely TF-IDF and trigram, several conclusions were obtained:

1. In the decision tree model, using the TF-IDF feature model achieved an accuracy of 63%, precision of 65%, recall of 85%, and an F-1 score of 74%. Meanwhile, with trigram, the accuracy remained the same (63%), but precision increased to 62% and recall reached 100%, resulting in an F-1 score of 77%. This indicates that using trigram improved the model's ability to recognize the positive class (recall) while slightly sacrificing precision.
2. The K-NN model with TF-IDF produced an accuracy of 56%, precision of 61%, recall of 76%, and an F-1 score of 67%. Meanwhile, with trigram, accuracy slightly increased to 60%, precision remained at 60%, but recall reached 100%, resulting in an F-1 score of 75%. Using trigram in K-NN also improved recall without sacrificing precision.
3. The SVM model with TF-IDF obtained an accuracy of 65%, precision of 64%, recall of 95%, and an F-1 score of 76%. Meanwhile, with trigram, accuracy dropped to 62%, precision to 61%, but recall increased to 100%, keeping the F-1 score at 76%. Using trigram in SVM also enhanced the model's ability to recognize the positive class without significantly compromising accuracy.

Compared to previous research, the use of the trigram feature model with the Naïve Bayes classifier method on the body shaming tweets dataset resulted in an accuracy of 68%, precision of 66%, recall of 68%, and an F-measure of 66%. These results indicate that the use of trigram consistently improves the model's ability to recognize the positive class but may affect precision in some classification methods.

Acknowledgments:

We sincerely thank the North-West University (NWU), Vanderbijlpark, for their outstanding resources, support, and financing for study. Their help has been essential to propelling our research to new heights. Without their assistance, the achievements and developments in this study would not have been feasible. We are appreciative of NWU for the opportunities they have provided us with as well as their dedication to advancing innovation and knowledge.

References:

[1] A. Rahmansyah, O. Dewi, P. Andini, T. Hastuti, P. Ningrum, and M. E. Suryana, "Membandingkan Pengaruh Feature Selection Terhadap Algoritma Naïve Bayes dan Support Vector Machine," 2018.

[2] A. Roihan, P. Abas Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," Tangerang, Apr. 2020.

[3] D. F. Zhafira, B. Rahayudi, and Indriati, "Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes dan Pembobotan TF-IDF Berdasarkan Komentar pada Youtube," Malang, Aug. 2021.

[4] M. K. Anam, B. N. Pikir, and M. B. Firdaus, "Penerapan Naive Bayes Classifier, K-Nearest Neighbor (KNN) dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah," MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 21, no. 1, pp. 139– 150, Nov. 2021, doi: 10.30812/matrik.v21i1.1092.

[5] St. F. Fattah and Purnawansyah, "Analisis Sentimen Terhadap Body Shaming Pada Twitter Menggunakan Metode Naïve Bayes Classifier," Indonesian Journal of Data and Science (IJODAS), vol. 3, no. 2, pp. 61–71, 2022.

[6] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiasari, "Analisis sentimen pada rating aplikasi Shopee menggunakan metode Decision Tree berbasis SMOTE," AITI: Jurnal Teknologi Informasi, vol. 18, no. Agustus, pp. 173–184, 2021.

[7] A. Q. Surbakti, R. Hayami, and J. Al Amien, "Analisa Tanggapan Terhadap PSBB Di Indonesia Dengan Algoritma Decision Tree Pada Twitter," Jurnal CoSciTech (Computer Science and Information Technology), vol. 2, no. 2, pp. 91–97, Dec. 2021, doi: 10.37859/coscitech.v2i2.2851.

[8] M. M. Baharuddin, H. Azis, and T. Hasanuddin, "Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca," ILKOM Jurnal Ilmiah, vol. 11, no. 3, pp. 269–274, Dec. 2019, doi: 10.33096/ilkom.v11i3.489.269-274.

[9] T. M. F. A. N. J. A. Septian, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," Surabaya, Aug. 2019.

[10] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, "Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi," Jurnal Teknoinfo, vol. 14, no. 2, p. 115, Jul. 2020, doi: 10.33365/jti.v14i2.679.

[11] Y. Julianto, D. H. Setiabudi, and S. Rostianingsih, "Analisis Sentimen Ulasan Restoran Menggunakan Metode Support Vector Machine," Surabaya, 2022.

[12] M. I. Hasan, "Information Retrieval System Artikel Kesehatan Menggunakan Pembobotan TF-IDF dan Latent Semantic Indexing," 2018.

[13] M. Syarifuddin, "Analisis Sentimen Opini Publik Terhadap Efek PSBB Pada Twitter dengan Algoritma Decision Tree, KNN, dan Naive Bayes," INTI Nusa Mandiri, vol. 15, no. 1, pp. 87–94, Aug. 2020, doi: 10.33480/inti.v15i1.1433.