PERFORMANCE OF SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE ON SUPPORT VECTOR MACHINE AND K-NEAREST NEIGHBOR FOR SENTIMENT ANALYSIS OF METAVERSE IN INDONESIA

[1]Roy Antonio, [2]Hironimus Leong
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Katolik Soegijapranata
[2]marlon.leong@unika.ac.id

ABSTRACT

The metaverse is one of the most discussed things on social media, Twitter in Indonesia. This view can be both positive and negative in Indonesian society, hence the need for sentiment analysis. However, creating a sentiment classification model with unbalanced data will reduce performance. For this reason, Synthetic Minority Oversampling is needed in Support Vector Machine and K-Nearest Neighbor algorithms. The results of Synthetic Minority Oversampling can improve the accuracy of the Support Vector Machine and K-Nearest Neighbor algorithms.

Keywords: Metaverse, Synthetic Minority Oversampling, Support Vector Machine, K-Nearest Neighbor

INTRODUCTION

Metaverse is one of the things that is discussed on social media, especially on Twitter. Based on Tunca et al. in TWITTER ANALYSIS FOR METAVERSE LITERACY, Indonesia is number two for countries that use tweets with #Metaverse. Based on this research, it can be seen that many Indonesians are following the development of the Metaverse. So it raises the question of what influence is obtained on Indonesian society, how Indonesian people respond to the metaverse, and to answer these questions requires sentiment analysis from social media, one of which is from Twitter

Responses from Indonesian people need to be classified into positive and negative responses using sentiment analysis. Sentiment analysis is the process of understanding, extracting and processing textual data automatically to obtain information (Pang & Lee, 2008). In the analysis, classification is needed to classify which responses have positive sentiment and negative sentiment. There are several algorithms for classifying text including Support Vector Machine (SVM) and K-nearest Neighbor (K-NN). The author wants to compare the sentiment analysis algorithm of Indonesian people's responses to the metaverse on Twitter social media with SVM and K-NN classification algorithms and perform the Synthetic Minority Oversampling Technique (SMOTE) method to handle cases of unbalanced data and measure the quality of analysis results using several parameters such as AUC value, accuracy, precision and recall.

LITERATURE STUDY

Data obtained from social media platforms is a popular subject today. This research [1] provides important information about people's thoughts about an event, situation or concept. For this purpose, several studies have been conducted with different methods in the literature. Twitter is very popular and useful for research that analyzes tweets because many people express feelings and ideas about a topic.

Ahmad & Gata [2] from Nusa Mandiri University conducted sentiment analysis research on Metaverse with the aim to see the response of Indonesian people to Twitter social media regarding the emergence of metaverse technology. This research uses the CRISP-DM method and uses the Support Vector Machine algorithm and is compared with the Tree algorithm, using the R language with the Rstudio application. However, there is a gap if using R Studio the file size stored is relatively small. Researchers use four stages of preprocessing: case folding, cleaning, stopword-removal, and stemming.

Hadma et al. [3] from Gajah Mada University, conducted research on Sentiment Analysis on Twitter. The research conducted was the data collection and labeling stage. Retrieving data from Twitter is quite easy to do because Twitter already provides an API (Application Programming Interface). After the data is collected into a dataset, the next stage is labeling. The number of sentiment classes that are widely used are two and three classes, namely negative, neutral, and positive. The next stage is preprocessing, including tokenization, cleansing, and filtering. After that, word weighting is a mechanism to assign a value to the occurrence of a word in a text document. One of the popular methods to perform word weighting is TF-IDF. Advanced sentiment analysis by comparing two methods, namely Naive Bayes Classifier and Support Vector Machine.

Harjanta [4] from PGRI University Semarang, conducted research on text preprocessing in the text mining process, which is expected to reduce by removing unnecessary or meaningless words or text from text databases or documents. Preprocessing text in this article contains transform case, stop word filter, and tokenize filter. from the article can help me to choose what preprocessing will be used in my journal later.

Muhidin & Wibowo [5] from Budi Luhur University conducted sentiment analysis research on New Normal. The author takes data directly from Twitter with the search keywords New Normal and #NewNormal. Next, perform the Text Preprocesing stage and determine positive and negative. Followed by evaluation using the K-NN and SVM algorithms. From the author's research, I can conclude that the KNN algorithm will be superior to SVM if using the 10-fold K-fold cross-validation evaluation from this journal.

Nasution & Hayaty [6] from Amikom University Yogyakarta conducted research on machine learning and methods used to analyze sentiment by comparing supervised learning classifications, namely the K-Nearest Neighbor algorithm and Support Vector Machine. The author conducted a comparison to find out which one is better in terms of accuracy and processing time. According to the author, the K-NN algorithm is easy to implement with a high level of effectiveness and is suitable for various problems related to classification. While Support Vector Machine is one of the appropriate algorithms used for text classification.

Pamungkas & Kharisudin [7] from Semarang State University. conducted research on analyzing the responses of Indonesian people to the Covid-19 pandemic on Twitter social media. Researchers conducted text mining which is used for the process of mining text to find useful information in a collection of text documents so that patterns, trends, or interrelationships between texts are obtained. The results of text mining are in the form of text which generally has high noise, so it is necessary to do text preprocessing such as case folding, spelling normalization, tokenizing, filtering, and stemming. Next is sentiment analysis with the aim of analyzing and extracting textual data in the form of opinions, evaluations, attitudes, emotions, judgments, and sentiments of a person towards an item, person, organization, and problem.

Pertiwi [8] from Information Systems of Nusa Mandiri University, conducted research on 2019 Youth Transportation and Suggestions using sentiment analysis. The author uses several methods, namely classification methods using K-NN, SVM, Naïve Bayes, and Neural Network. Then testing and training on the dataset so that accuracy and AUC are obtained. This article proves that K-NN has the highest accuracy and AUC, in contrast to the previous article which proved that SVM is the best.

The word Metaverse first appeared in the early nineties [9]. However, after Facebook changed its brand name to Meta in October 2021, it suddenly became the most searched brand on the world agenda. Although many people spend hours on the Metaverse platform, which is one of the latest versions of virtual reality and augmented reality, studies in this area are still few. With this research, the author aims to contribute to the small number of studies and literacy of the Metaverse. The site "socialbearing.com", which uses the Natural Language Processing Method, a branch of Artificial Intelligence, has been set as the information interface in this study.

Zulfa & Winarko [10] from FMIPA UGM, Yogyakarta conducted sentiment analysis with computational research from textually expressed opinions and emotions. The authors conducted sentiment analysis using the Deep Belief Network (DBN) method which is one of the machine learning methods included in the Deep Learning method. From this article, it is known that the Deep Belief method is better than Naive Bayes and Support Vector Machine.

RESEARCH METHODOLOGY

In this research, there are several stages of methods that aim to support research such as data crawling, data preprocessing, implementation and evaluation.

Data Crawling

Data used in the study amounted to 1500. The form is in the form of tweets collected using tweepy which is connected to the twitter API.

Table 1. Annotation Tweet

No	Tweet	Annotation
0	Selagi mencari product-market fit, BRI juga terus mempersiapkan amunisi yang kami percaya dapat membantu perusahaan untuk relevan di metaverse. #KinerjaBRIQ3	Positive
1	Pasar Masih Kacau, Vacuum Pacu Ekspansi Metaverse ke Korea dan Amerika https://t.co/djbHGKlMAT	Negative

From this data (Table 1) an annotation is given to the sentiment of the tweet. Annotations are divided into 2 namely, positive and negative. Positive means tweets that are good, supportive and have a positive impact on others. While Negative means tweets that are bad, disastrous and have a negative impact on other people.

Preprocesing Data

Data preprocessing is changing the data to the format needed in future processes (Table 2). The collected tweets data is usually less qualified, such as there are abbreviations, less standard words, emoticons and inconsistencies so that data preprocessing is needed. Data preprocessing used such as Remove URL, Tokenization, Indonesian stemming, Transform not (Negative), Stopword removal and Case folding.

Table 2. Preprocessing Tweet

No	Tweets after tokenization, stopword removal and case folding	Annotation
0	selagi mencari productmarket fit bri amunisi percaya membantu perusahaan relevan metaverse kinerjabriq	Positive
1	pasar masih kacau vacuum pacu ekspansi metaverse korea amerika	Negative

Indonesian Stemming, which returns the word to its base word or removes affixes (Table 3).

Table 3. Tweet Stemming

No	Tweet after Stemming	Annotation
0	selagi cari productmarket fit bri amunisi percaya bantu usaha relevan metaverse kinerjabriq	Postive
1	pasar masih kacau vacuum pacu ekspansi metaverse korea amerika	Negative

TF-IDF

After going through the preprocessing stage, it must be numeric. To convert the data into numeric, namely using the TF-IDF weighting method. The method used determines how far the word is connected to the document by giving the weight of each word, namely Term Frequency Invers Document Frequency.

Calculation of the weight of a data using Tf-IDF, is done by calculating the TF number of each word with the weight of each word is 1. While the IDF value is calculated using the function equation $IDF(word) = log\frac{td}{df}$.

$$IDF(word) = log\frac{td}{df} \tag{1}$$

IDF (word) is the IDF value of each word to be searched, td is the total number of documents, df is the number of occurrences of words in all documents.

Implementation

The modeling process by comparing Support Vector Machine with K-Nearest Neighbor used for classification algorithms, the results of cleansing and polarity datasets are split into two data, namely training data and test data with a ratio of 70:30 with 70% training data and 30% test data. Followed by the stages of the Synthetic Minority Oversampling Technique (SMOTE) method.

Support Vector Machine

Support Vector Machine The algorithm classifies by dividing the data into two classes using a vector array called a hyperplane. The hyperplane serves to separate the classes with the maximum margin between the data points of the two classes.

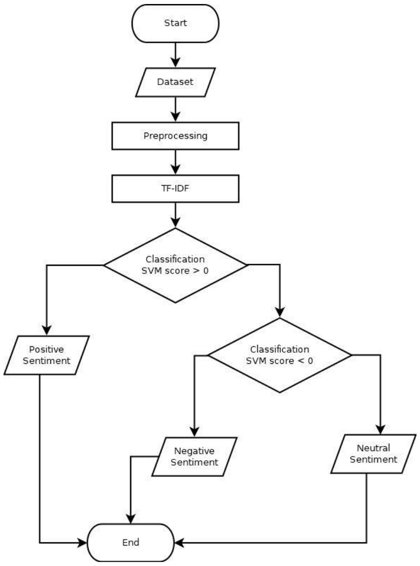

Figure 1. SVM Flowchart

K-Nearest Neighbor

K-Nearest Neighbors algorithm is an example-based classification determinant that does not build an explicit, declarative representation of categories, but relies on category labels attached to training documents similar to test documents. The K-Nearest Neighbors algorithm is a method of classifying objects based on the closest training data. Calculate the similarity between documents using cosine similarity in function equation $cos(\theta_{ij}) = \frac{\Sigma_k(d_{ik}d_{jk})}{\sqrt{\Sigma_k d_{ik}^2}\sqrt{\Sigma_k d_{jk}^2}}$.

$$cos(\theta_{ij}) = \frac{\Sigma_k\left(d_{ik}d_{jk}\right)}{\sqrt{\Sigma_k d_{ik}^2}\sqrt{\Sigma_k d_{jk}^2}} \tag{2}$$

Synthetic Minority Oversampling Technique

Accuracy is very important in machine learning classification. The more accurate the dataset and classes, the better the output. Classification can experience imbalance class data where each class does not have the same portion in the data set. Imbalanced data can cause inaccurate classification results.

SMOTE balances the data by adding synthetic data to the minority data by connecting the minority class data and then adding data along the randomly connected lines so that the amount of minority data can match the majority class.

First randomize 2 data from the minority class. Then the distance is calculated using the Euclidian distance formula and then multiplied by a value from 0 to 1 to produce a new point between the two points. This step is done continuously until the amount of data is balanced with the majority class.

For example N1(x1,y1) & N2(x2,y2) are randomly picked from the data. Then the distance is calculated and one of the points N3(x3,y3) through the trajectory is taken randomly. So as to produce function equation

$$(x_3, Y_3) = (x_1 + x_2.R, Y_1 + Y_2.R), 0 < R < 1 \tag{3}$$

.

$$(x_3, Y_3) = (x_1 + x_2.R, Y_1 + Y_2.R), 0 < R < 1 \tag{3}$$

Evalution

As a result of the comparison of SVM and K-NN used as a classification algorithm, the data will be refined using the Synthetic Minority Oversampling Technique (SMOTE) method and the data will be labeled into two training data and test data. This is followed by testing with parameters of accuracy, precision, recall to determine which algorithm performance is better to be used in sentiment analysis on metaverse.

RESULTS

From 3000 data taken from November to December 2022, it can be seen that there is more positive data than negative data. Annotating data on its own can lead to biased data. This is after data cleaning (neutral sentiment). So, 619 positive tweets and 202 negative tweets are obtained. This means, Indonesian people views the metaverse as a positive sentiment.

In this study there are 5 scenarios to get the best kernel and k. Balance data has a difference of less than 50 while imbalanced data has a data difference of more than 50. Dataset 1 is the data resulting from the preprocessing and data cleaning process which amounts to 619 and 202 for positive and negative tweets (imbalanced data), this dataset without using the SMOTE method. Dataset 2 is the data that is carried out the SMOTE process, the amount of data is 619 and 619 for positive and negative tweets (balanced data). Dataset 3 is the amount of data 202 and 619 for positive and negative tweets (imbalanced data), negative tweets are generated from the SMOTE method and positive tweets are taken randomly. Dataset 4 is the amount of data 400 and 300 for positive and negative tweets (imbalanced data), negative and negative tweets are taken randomly. Dataset 5 is the amount of data 300 and 400 for positive and negative tweets (imbalanced data).

The performance of SVM and KNN models without SMOTE shows approximately the same performance at 70% accuracy. While the performance of SVM and KNN is higher using the SMOTE method, namely SVM 94-96% and KNN 91%. The performance of the model can be seen on Figure 2 to Figure 7. Specifically, SVM Model performance can be seen on Figure 2, Figure 3, and Figure 4. The best kernel for SVM is RBF.

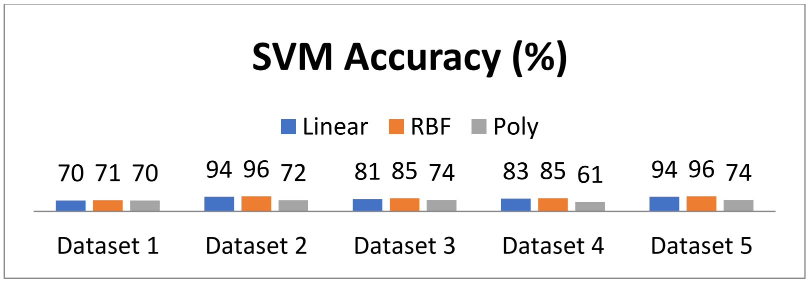

Figure 2. SVM Accuracy Graph

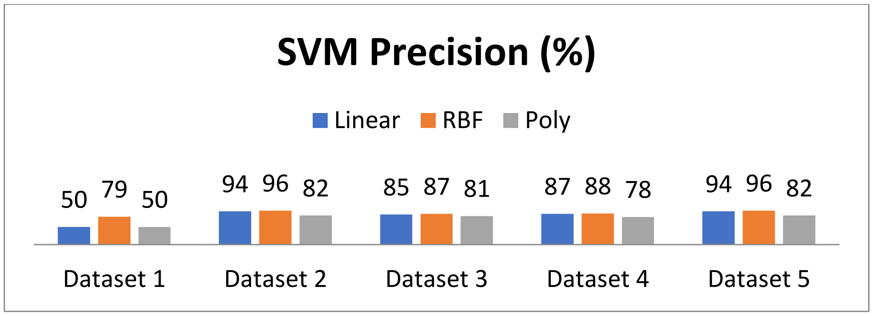

Figure 3. SVM Precision Graph

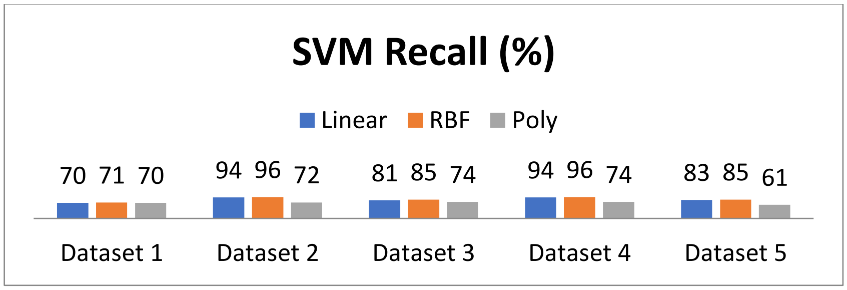

Figure 4. SVM Recall Graph

On the other hand, K-NN model performance can be seen on Figure 5, Figure 6, and Figure 7. The performance of K-NN is higher when using unbalanced datasets, namely datasets 3 and 5. The k for K-NN is around 5 or 6.

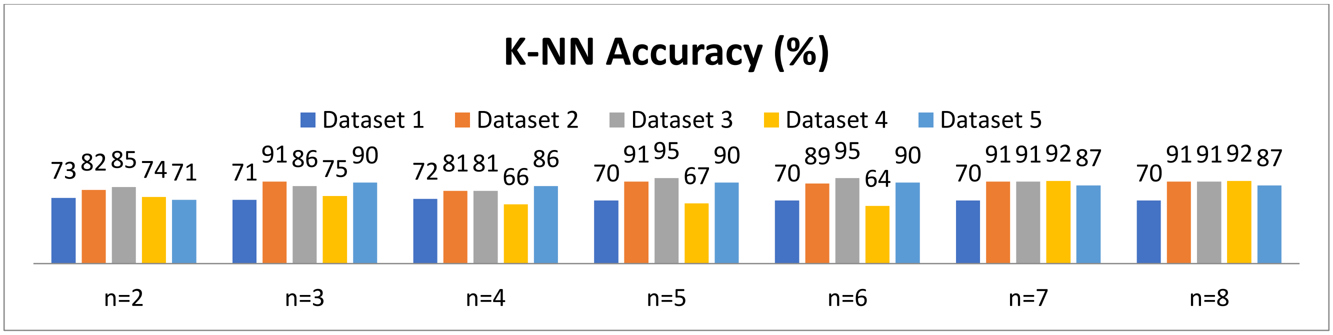

Figure 5. K-NN Accuracy Graph

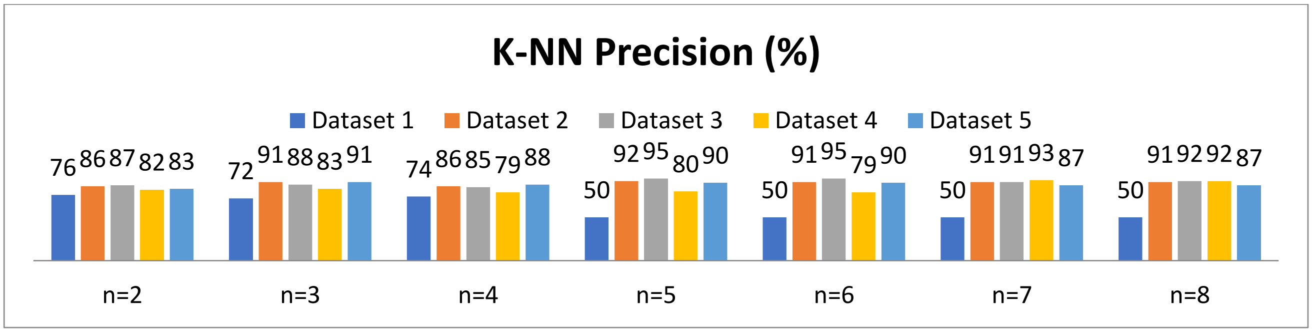

Figure 6. K-NN Precision Graph

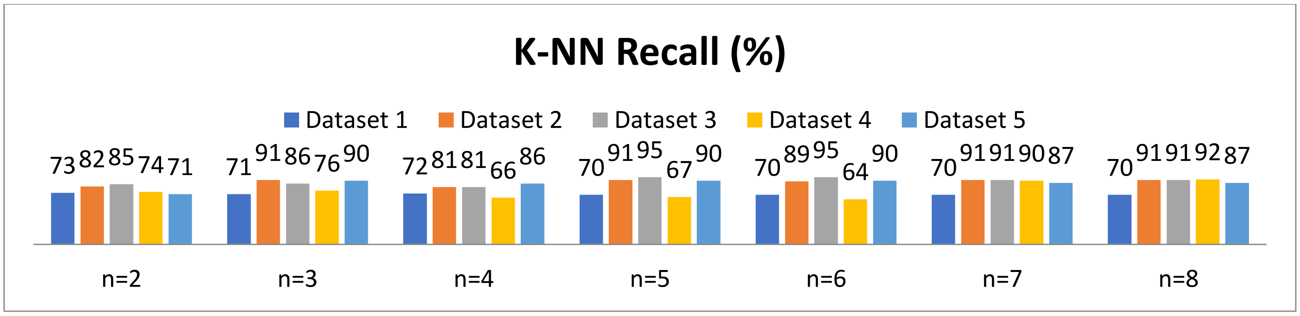

Figure 7. K-NN Recall Graph

The results of the above algorithm can be compared to Figure 8 and Figure 9

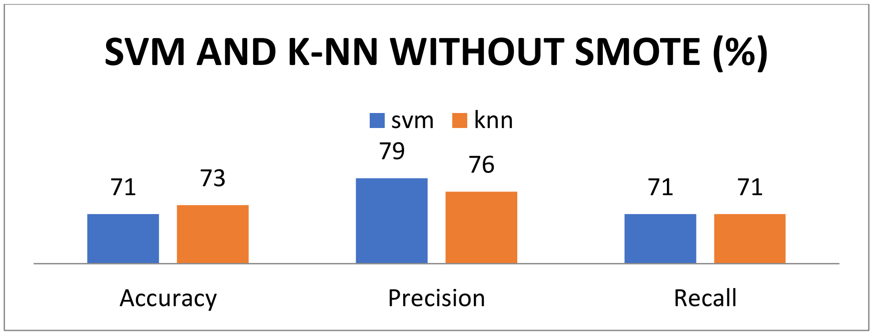

Figure 8. Comparison between SVM and K-NN using SMOTE

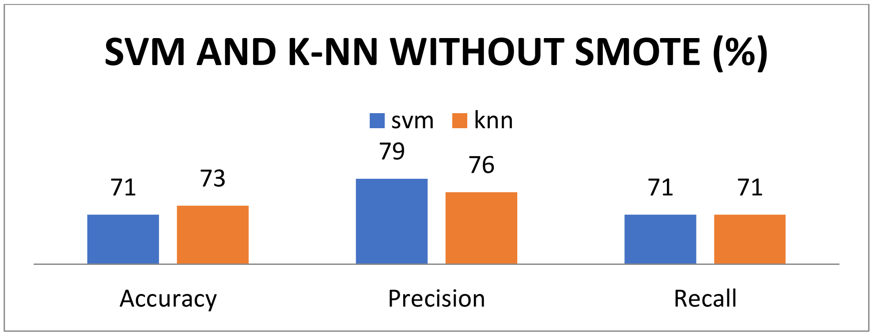

Figure 9. Comparison of SVM and K-NN without using SMOTE

The SMOTE method is higher because the data processed has data that is balanced against each other. Synthetic data to add data with minor classes allows the model to know the characteristics of each class. This is evident from the evaluation results of the SVM and K-NN modes which are higher using SMOTE. To be seen from its accuracy increases by 20%.

From this, it can be seen that the balanced data from the SMOTE method increases the amount of data available. Therefore, the lines created by the SVM algorithm can more perfectly divide the regions of each cluster. That is why the SVM algorithm is superior when the SMOTE method is used.

This is different when not using the SMOTE method where the cluster separation is not very clear because the data is less dense or there is a considerable distance between the data. Therefore, separators such as the SVM algorithm reduce the effectiveness in processing data, so the K-NN algorithm benefits more because it sees the similarity of the surrounding neighbors.

CONCLUSION

In this study the author examined sentiment analysis of the metaverse in Indonesia using the K-NN algorithm and SVM, and also SMOTE. The result of this research is that Indonesian people give positive sentiment to metaverse. In addition, the performance of SVM model and K-NN model are not much different. SVM algorithm performs superior on balanced data while K-NN algortihm perform superior on imbalanced data. The affect of SMOTE in this study is very effective to improve the performance of SVM and K-NN algorithm, because it increases up to 20% of the model performance.

The limitation of this research is that the data taken is not too much only 821 data after cleaning from 3000 data obtained in November to December 2022. This is due to many tweets that contain metaverse but are not related to metaverse. Suggestions for future research reproduce the data by increasing the time span, so that the machine learning model gets more data to reach a balanced point between bias and variance. Annotations carried out in this research are done alone, which can cause bias because the data can be manipulated and the data may not be classified correctly. To classify data, it is necessary to annotate with several people, then the data can be compared from the results of several people's annotations to determine the class of each tweet

REFERENCES

[1] Ö. AĞRALI and Ö. AYDIN, “Tweet Classification and Sentiment Analysis on Metaverse Related Messages,” Journal of Metaverse, vol. 1, no. 1, pp. 25–30, 2021.

[2] A. Ahmad and W. Gata, “Sentimen Analisis Masyarakat Indonesia di Twitter Terkait Metaverse dengan Algoritma Support Vector Machine,” jtik, vol. 6, no. 4, pp. 548–555, Mar. 2022, doi: 10.35870/jtik.v6i4.569.

[3] N. M. S. Hadna, P. I. Santosa, and W. W. Winarno, “Studi literatur tentang perbandingan metode untuk proses analisis sentimen di Twitter,” Semin. Nas. Teknol. Inf. dan Komun, vol. 2016, pp. 57–64, 2016.

[4] A. T. J. Harjanta, “Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining,” Jurnal Informatika Upgris, vol. 1, no. 1 Juni, 2015, [Online]. Available: http://journal.upgris.ac.id/index.php/JIU/article/view/804

[5] D. Muhidin and A. Wibowo, “Perbandingan Kinerja Algoritma Support Vector Machine dan K-Nearest Neighbor Terhadap Analisis Sentimen Kebijakan New Normal,” STRING, vol. 5, no. 2, p. 153, Dec. 2020, doi: 10.30998/string.v5i2.6715.

[6] M. R. A. Nasution and M. Hayaty, “Perbandingan Akurasi dan Waktu Proses Algoritma KNN dan SVM dalam Analisis Sentimen Twitter,” Jurnal Informatika, vol. 6, no. 2, pp. 226–235, 2019.

[7] F. S. Pamungkas and I. Kharisudin, “Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter,” in PRISMA, Prosiding Seminar Nasional Matematika, 2021, vol. 4, pp. 628–634. [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/article/view/45038

[8] M. W. Pertiwi, “Analisis sentimen opini publik mengenai sarana dan transportasi mudik tahun 2019 pada twitter menggunakan algoritma naïve bayes, neural network, KNN dan SVM,” Inti Nusa Mandiri, vol. 14, no. 1, pp. 27–32, 2019.

[9] S. Tunca, B. SEZEN, and Y. S. BALCIOĞLU, “TWITTER ANALYSIS FOR METAVERSE LITERACY’, 4,” in INTERNATIONAL NEW YORK ACADEMIC RESEARCH CONGRESS, 2022. [Online]. Available: https://www.researchgate.net/profile/Sezai-Tunca2/publication/358045545_TWITTER_ANALYSIS_FOR_METAVERSE_LITERACY/link s/61ee6aed8d338833e38f33f5/TWITTER-ANALYSIS-FOR-METAVERSELITERACY.pdf

[10] I. Zulfa and E. Winarko, “Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network,” IJCCS, vol. 11, no. 2, p. 187, Jul. 2017, doi: 10.22146/ijccs.24716.