Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Optimization of the K-Nearest Neighbors (KNN) Algorithm in Imbalanced Dataset Classification Using the SMOTE Technique

Abi Fajar Ahmad Fauzi[1], Ahmad Faqih[2], Kaslani[3]

[1,2] Informatics Engineering, STMIK IKMI Cirebon
[3] Computerized Accounting, STMIK IKMI Cirebon
Fauzi3209127@gmail.com[1]*

Abstract

The naturalization of players for Indonesia's national football team has sparked diverse reactions on Twitter, ranging from support to opposition. This situation poses challenges for sentiment analysis, particularly in interpreting public opinion on the policy. A significant challenge arises from the imbalance in sentiment classes, with neutral sentiments outweighing positive and negative ones. This research investigates the effect of class imbalance on sentiment analysis accuracy by employing the KNN algorithm enhanced with the SMOTE technique. A quantitative approach is used, adopting an experimental method aligned with the KDD process stages. The findings reveal that the KNN algorithm without SMOTE achieved an accuracy of 54.77%, with a Precision of 0.65, Recall of 0.57, and F1-Score of 0.44. However, integrating SMOTE with the KNN algorithm significantly improved the outcomes, boosting accuracy to 81.49%, with a Precision of 0.87, Recall of 0.80, and F1-Score of 0.80. These results demonstrate that oversampling techniques like SMOTE are highly effective in mitigating class imbalance and enhancing classification performance, especially for underrepresented classes. This study underscores the efficacy of SMOTE as a solution for addressing class imbalance in sentiment analysis tasks.

Keywords: *KNN, SMOTE, sentiment analysis, Twitter, naturalized players*

1. Introduction

The rapid development in the field of informatics has brought significant impacts in various sectors of life, including technology, business, and education. One aspect that is highly considered in the world of technology is the application of machine learning algorithms to solve various complex problems, especially in social media about the "Naturalization of Indonesian National Team Players" program has triggered various pro and contra reactions. In this context, the K-Nearest Neighbors (KNN) algorithm has become one of the popular methods for data classification due to its simplicity and ability to handle data with various patterns. However, in real-world applications, it is often the case that the available data is not balanced. To balance the data, synthetic minority oversampling technique (SMOTE) is used on unbalanced datasets[1]. The SMOTE method generates a new synthetic data pattern by performing linear interpolation between the minority class sample and its K nearest neighbors[2]. In addition, the K-Nearest Neighbor algorithm with TF-IDF weighting is feasible for the sentiment analysis process[3]. A comparative analysis of the discussion is presented when dealing with various levels of unbalanced data[4]. This research will be tested with model evaluation such as accuracy, confusion matrix. As in previous studies using the confusion matrix model evaluation[5]. If this research achieves the expected results, it will not only enrich the literature on data analysis and technology learning but also help government policy makers to better understand the public's reaction to player naturalization policies, which helps them make better policies.

2. Research methods

In this study using several stages by applying the KNN algorithm in sentiment analysis of the national team naturalization players. As well as using the KDD method to group the results of tweets[6]. KDD (Knowledge Discovery in Databases) is a systematic process for extracting useful information from large amounts of data. The following is a picture of the research method stage:

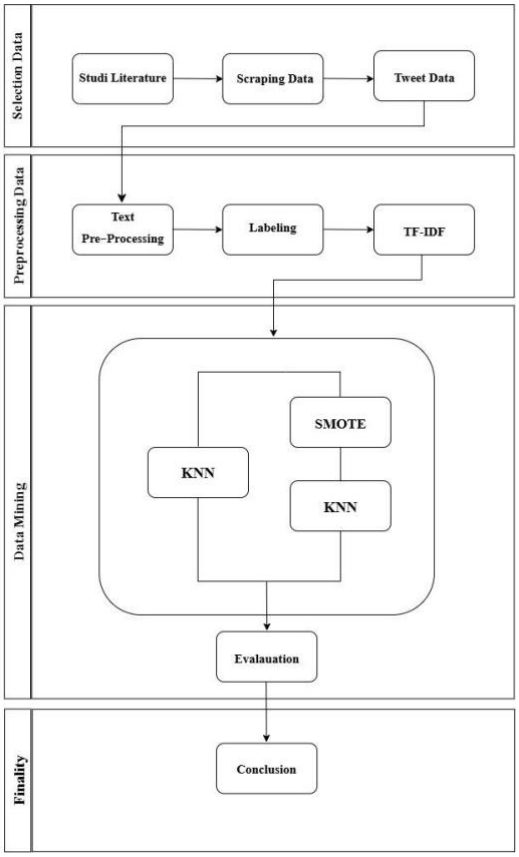

Fig. 1: Stage of The Research Method

The first stage is data selection, where data is collected to support the research process. The data used in this research is a collection of tweets from the Twitter platform, which is obtained using the Twitter Platform and the Snscrape library on Python to pull data. Next, preprocessing and labeling stages are carried out to prepare the data to be more structured. After that, model building is done with two approaches, namely KNN modeling and KNN modeling with SMOTE. The last stage is the presentation of research results summarized in the form of conclusions.

2.2. Data selection

At this stage, the data scraping process is carried out on the twitter platform to obtain data related to naturalized players, with the keywords "Naturalized Player", "National Team Naturralization", in the search process 2,800 naturalized tweets have been obtained. From these results, it is processed for selection so as to get 2,387 relevant tweets according to the discussion and there are 15 attribute columns such as conversation_id_str, created_at, favorite_count. full_text, id_str, image_url, in_reply_to_screen_name, lang, location, quote_count, reply_count, retweet_count, tweet_url, user_id_str, and username. The following is the result of data selection:

Table 1: Data Selection Result

…….	favorite_count	full_text	id_str	…….
	0	Timnas Indonesia sebenarnya hanya tim lemah tanpa 13 pemain naturalisasi seperti beberapa pertandingan terakhir. Itulah yang menyebabkan PSSI agak berhati-hati menyebut target di Piala AFF tulis Soha vn Lanjut.. https://t.co/CDG7UGrzwS #Beritabola #Football https://t.co/2ldRUkBu2b	1850520391237951671	
	0	Pemain keturunan Indonesia Pascal Struijk menegaskan dirinya mempertimbangkan untuk membela Timnas Indonesia. Namun ia lebih memperioritaskan tawaran ke Timnas Belanda. #pascalstruijk #naturalisasi #wni #timnasindonesia #belanda #belgia #leedsunited #infobola #beritabola https://t.co/bjPgP6kaNh	1850439029474631785	
	0	Blak-blakan Pertanyakan Kualitas Pemain Naturalisasi Eliano Reijnders Sampai Sindir Shin Tae-yong Begini ...: Pengamat sepak bola nasional Bung Towel memberikan komentar atas performa pemain naturalisasi terbaru timnas Indonesia Eliano https://t.co/V9NBg5JY5x https://t.co/2r9JZpdhwmt.co/bjPgP6kaNh	1850432552634655105	

2.3. Text pre-processing

The process of Cleansing and preparing text data for analysis is important to improve the efficiency and accuracy of the model. By adding samples from the minority class or removing samples from the majority class, resampling makes the data more balanced. In this case, there

are two main methods: oversampling and undersampling[7]**.** There are several stages used in this research such as cleansing, casefolding, tokenization, stopword removal, and stemming as in the research[8].

2.3.1. Cleansing

The data Cleansing process is a data cleansing process aimed at eliminating null values, incorrect data entries, irrelevant data, duplicate data, and inconsistent data, as their presence can reduce the quality or accuracy of the data mining results. Data cleansing will also affect the performance of the data mining system because the amount and complexity of the data to be handled will be reduced[9]. Here is the cleansing results table:

Table 2 : Cleansing Result

Before	After
Timnas Indonesia sebenarnya hanya tim lemah tanpa 13 pemain naturalisasi seperti beberapa pertandingan terakhir. Itulah yang menyebabkan PSSI agak berhati-hati menyebut target di Piala AFF tulis Soha vn Lanjut.. https://t.co/CDG7UGrzwS #Beritabola #Football https://t.co/2ldRUkBu2b	Timnas Indonesia sebenarnya hanya tim lemah tanpa pemain naturalisasi seperti beberapa pertandingan terakhir Itulah yang menyebabkan PSSI agak berhatihati menyebut target di Piala AFF tulis Soha vn Lanjut Beritabola Football

2.3.2. Casefolding
2.3.3.

This process converts all letters in the text to lowercase, this process is important in text analysis because it can reduce redundancy and improve the accuracy of subsequent processing. Here is the table of results from casefolding:

Table 3 : Casefolding Result

Before	After
Timnas Indonesia sebenarnya hanya tim lemah tanpa pemain naturalisasi seperti beberapa pertandingan terakhir Itulah yang menyebabkan PSSI agak berhatihati menyebut target di Piala AFF tulis Soha vn Lanjut Beritabola Football	timnas indonesia sebenarnya hanya tim lemah tanpa pemain naturalisasi seperti beberapa pertandingan terakhir itulah yang menyebabkan pssi agak berhatihati menyebut target di piala aff tulis soha vn lanjut beritabola football

2.3.4. Tokenization
2.3.5.

This process transforms a sentence into small units, from a sentence to words. In previous research, it was also stated that tokenization is the process of breaking down a complete sentence into smaller, more structured units. Here is the table of the tokenization results:

Table 4 : Tokenization Result

Before	After
timnas indonesia sebenarnya hanya tim lemah tanpa pemain naturalisasi seperti beberapa pertandingan terakhir itulah yang menyebabkan pssi agak berhatihati menyebut target di piala aff tulis soha vn lanjut beritabola football	timnas,indonesia,sebenarnya,hanya,tim,lemah,tanpa,pemain,naturalisasi,seperti,beberapa,pertandingan,terakhir,itulah,yang,menyebabkan,pssi,agak,berhatihati,menyebut,target,

2.3.6. Stopword removal

This process removes words that are considered to not provide much information, such as "and", "or", "that", "in", "from", and others. Here is the table of the stopword removal results:

Table 5: Stopword Removal Result

Before	After
timnas,indonesia,sebenarnya,hanya,tim,lemah,tanpa,pemain,naturalisasi,seperti,beberapa,pertandingan,terakhir,itulah,yang,menyebabkan,pssi,agak,berhatihati,menyebut,target,	timnas,indonesia,tim,lemah,pemain,naturalisasi,pertandingan,menyebabkan,pssi,berhatihati,menyebut,target,piala,aff,tulis,soha,vn,beritabola,footb all

2.3.7. Stemming

This process is to filter words that contain conjunctions, pronouns, prepositions, into root words by removing prefixes or suffixes[10]. Here is the table of the stemming results:

Table 6: Stemming Result

Before	After
timnas,indonesia,tim,lemah,pemain,naturalisasi,pertandingan,menyebabkan,pssi,berhatihati,menyebut,target,piala,aff,tulis,soha,vn,beritabola,footb all	timnas indonesia benar hanya tim lemah tanpa main naturalisasi seperti beberapa tanding akhir itu yang sebab pssi agak berhatihati sebut target di piala aff tulis soha vn lanjut beritabola football

2.3. Labeling

After completing the data preprocessing stage, we obtained cleaned data, followed by the labeling processing stage. In this stage, we used the IndoBERT technique to label the tweets. The purpose of this stage is to understand whether the sentiment is positive, negative, or neutral. Here is the labeling result table:

Table 7*: Labeling Result*

index	Stemming	Label
0	timnas indonesia benar hanya tim lemah tanpa main naturalisasi seperti beberapa tanding akhir itu yang sebab pssi agak berhatihati sebut target di piala aff tulis soha vn lanjut beritabola football	NETRAL
1	main turun indonesia pascal struijk tegas diri timbang untuk bela timnas indonesia namun ia lebih memperioritaskan tawar ke timnas belanda pascalstruijk naturalisasi wni timnasindonesia belanda belgia leedsunited infobola beritabola	POSITIF
2	bung towel blakblakan tanya kualitas main naturalisasi eliano reijnders sampai sindir shin taeyong begini amat sepak bola nasional bung towel beri komentar atas performa main naturalisasi baru timnas indonesia eliano	POSITIF

2.4. TF-IDF

TF-IDF (term frequency-inverse document frequency) is a weighting scheme commonly used in natural language processing to assign weights to terms in a document based on their importance or informativeness. It is calculated as the product of term frequency (TF) and inverse document frequency (IDF (TF) Term Frequency is the frequency of a word's occurrence in a text document, while (IDF) Inverse Document Frequency is the frequency of a term's occurrence across all text documents. Terms that rarely appear in the entire set of text documents have a higher Inverse Document Frequency value compared to terms that frequently appear[12]. So, TF-IDF helps to find important words in a document without being influenced by frequently occurring words.

2.5. Data mining

Data mining is the process of finding patterns from a large dataset that can subsequently uncover valuable and interesting information from the data[13]. In this process, the KDD stages are used, which aim to transform unstructured text data into organized information. this process is divided into two modeling stages, first the KNN algorithm modeling and second the KNN algorithm modeling with SMOTE.

```
from sklearn.neighbors import KNeighborsClassifier
# Memisahkan data latih dan uji setelah SMOTE
X_train, X_test, y_train, y_test = train_test_split(X_smote, y_smote, test_size=0.2, random_state=42)

# Inisialisasi model K-Nearest Neighbors
knn = KNeighborsClassifier(n_neighbors=5, weights='distance', metric='manhattan')

# Latih model dengan data latih
knn.fit(X_train, y_train)

# Lakukan prediksi pada data uji
y_pred = knn.predict(X_test)

print(knn)
```

Fig. 2 : KNN Modelling

The modeling described uses the K-Nearest Neighbors (KNN) algorithm for data classification purposes. First, KNeighborsClassifier is imported from the `sklearn.neighbors` library. KNN is an algorithm that operates with an instance-based learning approach, where predictions are made based on the majority label of the nearest neighbors in the feature space. This KNN model is initialized with several important parameters, such as n_neighbors=5, which determines the number of nearest neighbors used in the prediction, namely five neighbors. weights='distance' means that the weight of each neighbor is calculated based on distance, so closer neighbors have a greater influence on the prediction result. metric='manhattan' is used to measure the distance between points by calculating the sum of absolute differences between each dimension of the data. After the model is initialized, the next step is to train the model using the training data with the fit(X_train, y_train) method. Here, X_train contains the features used to train the model, while y_train is the label corresponding to that data. After the model is trained, predictions are made on the test data using predict(X_test), resulting in predictions based on the majority label of the five nearest neighbors for each test data. Finally, information about the KNN model is displayed with print(knn), which shows the parameters used in the model. Overall, KNN relies on training data to make predictions without requiring complex model training like other algorithms. The use of Manhattan distance and distance-based weights allows the model to give more influence to closer neighbors. The main advantages of KNN are its ease of use and simplicity, although this algorithm can become slow when applied to large datasets.

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from imblearn.over_sampling import SMOTE

# Memastikan DataFrame 'data' sudah diinisialisasi
# Inisialisasi TF-IDF Vectorizer dan transformasikan data teks menjadi vektor numerik
vectorizer = TfidfVectorizer(ngram_range=(1, 3), max_features=7000, max_df=0.8, min_df=3)
X = vectorizer.fit_transform(data['stemming']).toarray()
y = data['label']  # Label

# Terapkan SMOTE untuk menangani data tidak seimbang
smote = SMOTE(random_state=42)
X_smote, y_smote = smote.fit_resample(X, y)

print(smote)
```

Fig. 3 : SMOTE Modelling

In this modeling, the text in the 'stemming' column of the `data` dataset is converted into a numerical representation using the TF-IDF Vectorizer so that it can be processed by machine learning algorithms. Several parameters applied in the `TfidfVectorizer` include ngram_range=(1, 3) which generates unigrams to trigrams, max_features=7000 which limits the number of features to the 7000 most frequent features, max_df=0.8 to ignore words that appear in more than 80% of the documents, and min_df=3 to only consider words that

appear in at least 3 documents. The result of this process is the feature matrix X, which represents the text in numerical form, and y, which contains the category labels of the data.To address the issue of class imbalance in the dataset, the SMOTE (Synthetic Minority Over-sampling Technique) technique is applied. SMOTE works by creating new samples for the less frequent class based on the existing data, aiming to balance the class distribution. By using SMOTE(random_state=42), this technique ensures consistent results when applied. The fit_resample(X, y) function is used to generate a more balanced dataset, namely X_smote and y_smote. Finally, print(smote) is used to display information about the applied SMOTE object. This step aims to enable the model to learn patterns more effectively, especially in the underrepresented class in the dataset.

label	
POSITIF	1263
NETRAL	1263
NEGATIF	1263

dtype: int64

Fig. 4 : SMOTE Modelling Result

Figure 4 shows that the values of positive, negative, and neutral sentiments are balanced. With the data that has already used the SMOTE method, it can be used for KNN algorithm modeling with SMOTE to determine how effective the influence of SMOTE is on KNN.

```
from sklearn.neighbors import KNeighborsClassifier
# Memisahkan data latih dan uji setelah SMOTE
X_train, X_test, y_train, y_test = train_test_split(X_smote, y_smote, test_size=0.2, random_state=42)

# Inisialisasi model K-Nearest Neighbors
knn = KNeighborsClassifier(n_neighbors=5, weights='distance', metric='manhattan')

# Latih model dengan data latih
knn.fit(X_train, y_train)

# Lakukan prediksi pada data uji
y_pred = knn.predict(X_test)

print(knn)
```

Fig. 5: KNN Modelling with SMOTE

The first stage uses pandas to manage data in DataFrame format, which facilitates efficient data storage and processing. Meanwhile, TfidfVectorizer converts text into a numerical representation by calculating the TF-IDF value, which measures how relevant a word is in a document compared to the entire collection of documents. Some parameters used in TF-IDF include `ngram_range=(1, 3)`, which allows the use of unigrams, bigrams, and trigrams, as well as `max_features=7000` to limit the number of features used. Additionally, the parameters `max_df=0.8` and `min_df=3` are used to ignore words that appear too frequently or too rarely, so that only relevant words are considered. After the text is converted into numerical form, SMOTE (Synthetic Minority Oversampling Technique) is applied to address class imbalance in the dataset. SMOTE works by generating synthetic data for the minority class, which helps balance the class distribution. In this implementation, `random_state=42` is used to ensure consistent resampling results. The result of applying SMOTE is two variables: `X_smote` which contains the features after the resampling process and `y_smote` which contains the corresponding labels. To evaluate the effectiveness of SMOTE, the class distribution before and after the application of this technique can be compared to see the changes that occur in the data after SMOTE is applied.

2.6. Evaluation

Model evaluation is an important process in assessing the performance of a model or algorithm in data processing. Model evaluation can provide insights into how well or poorly the model performs in completing the given tasks, such as classification. In this study, the evaluation model uses accuracy, precision, recall, and f-1 score to provide a comprehensive picture of how well the model can correctly classify the sentiment of reviews[14]. Evaluation:
In this study, the performance of the KNN (K-Nearest Neighbors) model was compared using four main metrics: accuracy, precision, recall, and F1-score, both before and after the application of the SMOTE method (Synthetic Minority Oversampling Technique). Here are the evaluation results for each metric:

2.6.1. Accuracy

Before using SMOTE, the accuracy of the KNN model was only 54.77%, indicating that the model struggled to handle a dataset with an imbalanced class distribution, where the majority class was more dominant. After applying SMOTE, the accuracy significantly increased to 81.49%. This shows that SMOTE successfully balanced the class distribution and helped the model perform better in classification.

2.6.2. Precision

Precision measures how many true positive predictions there are compared to the total number of positive predictions. Before SMOTE, the model's precision was recorded at 0.65, which means the model often made mistakes in identifying the positive class. After SMOTE

was applied, precision increased to 0.87, indicating that the model was more accurate in recognizing the positive class and reducing prediction errors.

2.6.3. Recall

Recall measures the model's ability to detect the actual positive class. Before the application of SMOTE, the model's recall was only 0.57, indicating that the model struggled to detect the minority class. However, after SMOTE was applied, the recall increased to 0.80, showing that the model is now more effective in detecting the minority class and is not affected by the dominance of the majority class.

2.6.4. F1-Score

The F1-score is a metric that combines precision and recall to provide a comprehensive overview of the model's performance. Before SMOTE, the model's F1-score was only 0.44, reflecting an imbalance between precision and recall. After SMOTE was applied, the F1-score increased to 0.80, indicating that the model is now more balanced in achieving good results between precision and recall.

Overall, the application of SMOTE significantly improved all evaluation metrics: accuracy, precision, recall, and F1-score. This proves that SMOTE is an effective technique in addressing data imbalance and enhancing the performance of the KNN model, especially in sentiment analysis on imbalanced datasets.

3. Result and Discussion

The application of SMOTE on the KNN algorithm has proven to significantly improve classification accuracy on datasets with imbalanced class distributions. This research uses a dataset containing Twitter user comments about the naturalized players of the Indonesian national team, where positive, negative, and neutral sentiments have an uneven distribution. This imbalance causes the KNN algorithm to lean towards the majority class, which consists of 2,387 tweets used in the study, where the labeling results show 757 tweets as positive, 338 tweets as negative, and 1,263 tweets as neutral. The labeling results indicate that the neutral labeling value tends to be higher than the positive and negative values, resulting in low classification performance for the minority class. With SMOTE, synthetic data for the minority class is generated, making the data distribution more balanced. The research results show that the accuracy of the KNN algorithm increased from 54.77%, precision 0.65, recall 0.57, and f1-score 0.44 after using SMOTE to 81.49%, precision 0.87, recall 0.80, and f1-score 0.80. This occurs because SMOTE can enhance the representation of the minority class by creating synthetic samples based on the original data patterns. With a more even data distribution, KNN can calculate the proximity between data more accurately without being influenced by the dominance of the majority class. As in the research discussing the impact of SMOTE in this journal, the application of the SMOTE method to balance an imbalanced dataset shows an increase in classification model accuracy. For example, the accuracy for the Decision Tree (DT) and K-Nearest Neighbors (KNN) algorithms increased from (82.81% and 82.15%) to (89.57% and 84.05%) after the application of SMOTE and Z-score normalization. Additionally, the Recall rate also increased, indicating that SMOTE helps improve the model's performance in classifying the minority class[15]. This shows that oversampling techniques like SMOTE are an effective approach to handling data imbalance in sentiment analysis, resulting in a more reliable model capable of capturing patterns from all classes. This research highlights the importance of balanced data distribution to improve the performance of machine learning algorithms in various contexts, including social sentiment analysis. The results can serve as a reference for future research in addressing data imbalance issues in similar situations.

Table 8: Model Evaluation Result

	KNN	KNN with SMOTE	Difference
Accuracy	54.77%	81.49%	26.72%
Precision	0.65	0.87	0.22
Recall	0.57	0.80	0.23
F1-Score	0.44	0.80	0.36

4. Conclusion

This research shows that the application of the SMOTE (Synthetic Minority Oversampling Technique) on the KNN (K-Nearest Neighbors) algorithm significantly improves accuracy in sentiment analysis classification, especially on datasets with class imbalance. By using SMOTE, the data distribution becomes more balanced through the creation of synthetic samples for the minority class, allowing the KNN algorithm to more effectively recognize patterns across all classes. The research results show an increase in accuracy from 54.77% to 81.49%, as well as improvements in other metrics such as Precision, Recall, and F1-score. This proves that SMOTE is an effective technique for addressing data imbalance and improving model performance in sentiment analysis. This research provides important insights into the significance of balanced data distribution to improve the accuracy of machine learning models and can serve as a reference for future studies facing similar issues.

Referensi

[1] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, And G. S. Choi, "Impact Of SMOTE On Imbalanced Text Features For Toxic Comments Classification Using RVVC Model," *IEEE Access*, Vol. 9, Pp. 78621–78634, 2021, Doi: 10.1109/ACCESS.2021.3083638.

[2] D. Elreedy, A. F. Atiya, And F. Kamalov, "A Theoretical Distribution Analysis Of Synthetic Minority Oversampling Technique (SMOTE) For Imbalanced Learning," *Mach. Learn.*, Vol. 113, No. 7, Pp. 4903–4923, 2024, Doi: 10.1007/S10994-022-06296-4.

[3] S. Rahayu, Y. MZ, J. E. Bororing, And R. Hadiyat, "Implementasi Metode K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial FLIP," *Edumatic Jurnal. Pendidikan. Informatika.*, Vol. 6, No. 1, Pp. 98–106, 2022, Doi: 10.29408/Edumatic.V6i1.5433.

[4] E. F. Swana, W. Doorsamy, And P. Bokoro, "Tomek Link And SMOTE Approaches For Machine Fault Classification With An Imbalanced Dataset," *Sensors,* Vol. 22, No. 9, Pp. 1–21, 2022, Doi: 10.3390/S22093246.

[5] K. Pramayasa, I. M. D. Maysanjaya, And I. G. A. A. D. Indradewi, "Analisis Sentimen Program Mbkm Pada Media Sosial Twitter Menggunakan KNN Dan SMOTE," *SINTECH (Science Information. Technolog.) Jurnal.*, Vol. 6, No. 2, Pp. 89–98, 2023, Doi: 10.31598/Sintechjournal.V6i2.1372.

[6] S. W. Pebrianti, R. Astuti, And F. M. Basysyar, "Penerapan Algoritma K-Nearest Neighbor Dalam Klasifikasi Status Stunting Balita Di Desa Bojongemas," *JATI (Jurnal Mahasiswa. Teknik. Informatika.)*, Vol. 8, No. 2, Pp. 2479–2488, 2024, Doi: 10.36040/Jati.V8i2.8448.

[7] J. H. Joloudari, A. Marefat, M. A. Nemathollahi, S. S. Oyelere, And S. Hussain, "Effective Class-Imbalance Learning Based On SMOTE And Convolutional Neural Networks," *Applied. Sciences.*, Vol. 13, No. 6, Pp. 1–34, 2023, Doi: 10.3390/App13064006.

[8] J. W. Iskandar And Y. Nataliani, "Perbandingan Naïve Bayes, SVM, Dan K-NN Untuk Analisis Sentimen Gadget Berbasis Aspek," *Jural. RESTI (Rekayasa Sistem. Dan Teknologi. Informasi)*, Vol. 5, No. 6, Pp. 1120–1126, 2021, Doi: 10.29207/Resti.V5i6.3588.

[9] H. W. Azizah, O. Nurdiawan, G. Dwilestari, Kaslani, And E. Tohidi, "Klasifikasi Pemberian Bantuan UMKM Cirebon Dengan Menggunakan Algoritma K-Nearest Neighbor," *Journal. Computer. System. Informatics*, Vol. 3, No. 3, Pp. 110–115, 2022, Doi: 10.47065/Josyc.V3i3.1392.

[10] J. Supriyanto, D. Alita, And A. R. Isnain, "Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring," *Jurnal. Informatika. Dan Rekayasa Perangkat Lunak*, Vol. 4, No. 1, Pp. 74–80, 2023, Doi: 10.33365/Jatika.V4i1.2468.

[11] S. D. Prasetyo, S. S. Hilabi, And F. Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes Dan KNN," *Jurnal. Komtekinfo*, Vol. 10, No. 1, Pp. 1–7, 2023, Doi: 10.35134/Komtekinfo.V10i1.330.

[12] W. E. Nurjanah, R. Setya Perdana, And M. A. Fauzi, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat Pada Media Sosial Twitter Menggunakan Metode K-Kearest Neighbor Dan Pembobotan Jumlah Retweet," *Jurnal. Pengembangan. Teknologi. Informasi. Dan Ilmu Komputer.*, Vol. 1, No. 12, Pp. 1750–1757, 2017.

[13] N. Amalia, T. Suprapti, And G. Dwilestari, "Analisis Sentimen Pengguna Twitter Terhadap Pelaksanaan Kurikulum Mbkm," *E-Link Jurnal. Teknik. Elektro Dan Informatika.*, Vol. 18, No. 1, P. 57, 2023, Doi: 10.30587/E-Link.V18i1.5335.

[14] D. Nurwahidah, G. Dwilestari, N. Dienwati Nuris, And R. Narasati, "Analisis Sentimen Data Ulasan Pengguna Aplikasi Google Kelas Pada Google Play Store Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mahasiswa. Teknik. Informatika.)*, Vol. 7, No. 6, Pp. 3673–3678, 2024, Doi: 10.36040/Jati.V7i6.8245.

[15] A. J. Mohamme, M. M. Hassan, And D. H. Kadir, "Improving Classification Performance For A Novel Imbalanced Medical Dataset Using SMOTE Method," *International Journal of Advanced Trends in Computer Science and Engineering.*, Vol. 9, No. 3, Pp. 3161–3172, 2020, Doi: 10.30534/Ijatcse/2020/104932020.