ILKOM Scientific Journal Vol. 16, No. 2, August 2024, pp. 184-189
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article **Open Access (CC–BY-SA)**

Enhancing Accuracy by Using Boosting and Stacking Techniques on the Random Forest Algorithm on Data from Social Media X

Teri Ade Putra [a,1,*]; **Vicky Ariandi** [a,2] ; **Sarjon Defit** [a,3]
[a] Universitas Putra Indonesia YPTK Padang, Jl. Raya Lubuk Begalung Padang – West Sumatra, 25221, Indonesia
[1]teriadeputra@upiyptk.ac.id, [2]vicky_ariandi@upiyptk.ac.id, [3]sarjon_defit@upiyptk.ac.id
** Corresponding author*

Article history: Received April 22, 2024; Revised May 30, 2024; Accepted July 06, 2024; Available online August 20, 2024

Abstract

Online loans (commonly referred to as *Pinjol*) have become a widespread phenomenon in Indonesia, both in legal and illegal forms. It is undeniable that this is in line with the rapid development and innovation of technology. *Pinjol* cannot be separated from public comments, both positive and negative, on social media X. The study examined the communication patterns of Indonesian people using a sentiment analysis approach. The research utilized the Random Forest algorithm to perform sentient analysis. This algorithm combined the output of several decision trees to achieve a more accurate result. In addition to using a random forest algorithm, this study also made improvements by using stacking and boosting. The results of this study indicated that the highest accuracy of 86% was obtained by the SMOTE+RF+Adaboost (Boosting) model. In contrast, the lowest accuracy of 60% was obtained in the RF+Adaboost model with a stacking technique.

Keywords: Boosting; Random Forest; SMOTE; Social Media X; Stacking.

Introduction

Ensemble methods can be defined as techniques that builds multiple models and combines them to produce better results [1]. The Ensemble method in machine learning typically produces more accurate solutions than single models [2]. There are several techniques used such as stacking, bagging, boosting and voting [3]. These techniques have their own advantages and disadvantages.

The stacking technique has a strong ensemble learning strategy in machine learning that combines the predictions of various basic models such as SVM, Random Forest, Naïve Bayes and so on, to get the final prediction with better performance [4]. This technique is also known as stacked ensemble or stacked generalization. In addition, the Boosting technique is ensemble learning that combines a weak set of algorithms into a strong algorithm to minimize training errors. In boosting, a sample of data is randomly selected, equipped with a model, and then trained sequentially, i.e., each model tries to compensate for the weaknesses of its predecessor model [4].

This study compared the text data taken from social media X about online loans or commonly called *pinjol*. Currently, *pinjol* is a phenomenon that is being hotly discussed by many people in Indonesia [5] . This is because there are many news reports that state that loans are often detrimental to the community [6]. However, some people also say that the online loans can be a viable alternative solution to borrow money other than banks [7].

Previous studies used sentiment analysis to analyze tweets by classifying them as positive, negative, and neutral. For example, research with sentiment analysis was conducted by [8] using ensemble stacking with linear kernel SVM meta classifier and meta classifier logistic regression. The accuracy result was 88%. Then [9] conducted sentiment analysis research using stacking (LR+SVM+RF), it obtained an accuracy of 86%. Furthermore [10] , using the multinomial Naïve Bayes algorithm with Adaboost and information gain obtained an accuracy of 87.87%. The last study [11], it employed the boosting technique with XGBoost, LightGBM, Adaboost, and Gradient Boosting algorithms. The obtained accuracy was 96.75%.

Despite the insights from the previous research, there has not been a comparison between techniques between the different ensemble methods. Therefore, this study used a comparison of ensemble techniques, namely stacking and boosting by using Random Forest algorithm as the base model. The two techniques were compared to see their accuracy. Random Forest was employed because it is a popular machine learning algorithm included in supervised learning techniques. It can be used for classification and regression issues in machine learning.

https://doi.org/10.33096/ilkom.v1xix.xxx.x-x

Before the modeling process was carried out, this study conducted preprocessing to ensure the data was clean and free from noise [12]. The preprocessing used was data cleaning, case folding, text normalization, filtering, stemming, and data transformation. After the data was cleaned, a 70:30 data sharing process was carried out from 3391 tweets. Then modeling was carried out with random forests, along with boosting and stacking using the Adaboost algorithm. This study also used Synthetic Minority Oversampling Technique (SMOTE) to balance classes so that the processed data was better than without SMOTE.

Method

Figure 1 is the research flow used to make it easier to conduct the research. This flow starts from searching for tweets from social media to modeling using a random forest algorithm.

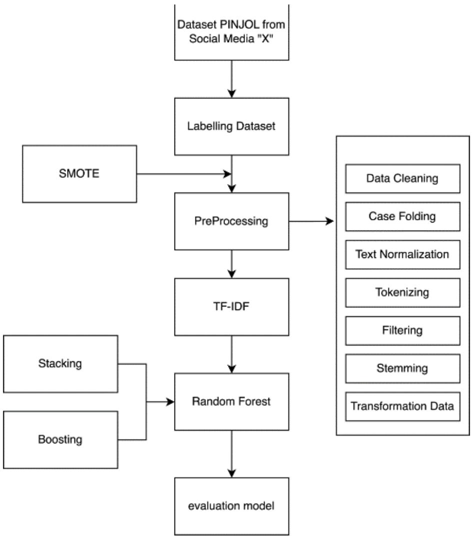


Figure 1. Research Flow

A. Datasets and labelling

The dataset in this study was an online loan data (known as *pinjol*) taken from social media X starting from July 2022 to July 2023. During this period, a total of 3,391 tweets were collected. Then the dataset was labeled as positive, negative, and neutral.

B. Class balancing

Classroom imbalance is a common issue in machine learning, where the proportion of data in each class is unbalanced. These issues often arise in a variety of contexts, including medical diagnostics, spam screening, fraud detection, emotional classification, and a variety of other fields [13]. When there is a class imbalance in the training data, machine learning models often tend to overclassify the majority class due to the increased prior probability. Consequently, machine learning algorithms become vulnerable to misclassifying minority classes [14]. For this reason, the need for class balancing is necessary, one of the methods used is the SMOTE method [15]. The Synthetic Minority Oversampling Technique SMOTE is a statistical technique used to increase the number of cases in a dataset to create a balance. This component works by generating a new instance of an existing minority case, which is provided as input [16].

C. Preprocessing

Preprocessing in this study involved 7 steps: data cleaning, case folding, text normalization, tokenizing, filtering, stemming, and data transformation. data cleaning was used in this study to eliminate unnecessary columns such as links, authors, followers. the data processed only required mentions and sentiment labels. then case folding was a stage in text pre-processing used to standardize characters on the data. The case folding process was the process of converting all letters to lowercase letters [17]. Then text normalization is a concept used to convey ideas by changing the format of the text to fit a specific purpose [18].

Tokenization was a pre-processing step in the field of information retrieval and Natural Language Processing (NLP), where text was divided into smaller units called "tokens". The goal of tokenization was to break down the text into more manageable units, making it easier to analyze and process [19]. Then filtering was the process of cleaning up unnecessary symbols in the document, such as punctuation, numbers, and emoticons [20]. Stemming was the process of removing affixes both at the beginning and at the end of the word. The goal was to get the root word [21]. Finally, preprocessing in the study used transformation data to convert categories into numbers [22].

D. TF-IDF

Term Frequency (TF) is a metric used to find out how often a word appears in a document. This can be calculated by dividing the number of occurrences of those words in the document by the total number of words in the document [23]. For example, if the word "machine learning" appears 10 times in a document containing 100 words, then the TF of the word is 10/100 = 0.1. On the other hand, Inverse Document Frequency (IDF) is a measure used to assess how common a word is across a document corpus. It is calculated by dividing the total number of documents in the corpus by the number of documents containing the word and taking the logarithms from the results.

The general formula of IDF is log(N/n), where N is the total number of documents in the corpus, and n is the number of documents containing the word. The fewer documents that contain the word, the higher the IDF value [24]. Furthermore, the TF*IDF score is the result of the multiplication between the TF and IDF scores for a word in a document. The general formula is TF * IDF. For example, if the word "machine learning" has a TF score of 0.1 and an IDF score of 2, then the TF*IDF score for that word is 0.1 * 2 = 0.2. [25].

E. Modeling

In this study, 6 models were used, namely Based Random Forest, SMOTE+ Random Forest, Random Forest+Stacking (Adaboost), SMOTE+Random Forest+Stacking (Adaboost), Random Forest+Boosting (Adaboost), SMOTE+Random Forest+Boosting (Adaboost). After modeling, the next step was to evaluate the model by examining accuracy, precision, recall, and F1-Score.

Results and Discussion

The following are the results and discussion of the theme analysis of online loans using a sentiment analysis approach using the random forest algorithm. The first step was to preprocess the data first using data cleaning, case folding, text normalization, tokenizing, filtering, stemming, and data transformation. After the data was cleaned, the next process was carried out as described below:

A. Data Balancing

The imbalanced data can be seen in the **Figure 2**. On the neutral label, there was a considerable difference in the number of neutral labels compared to the positive and negative labels.

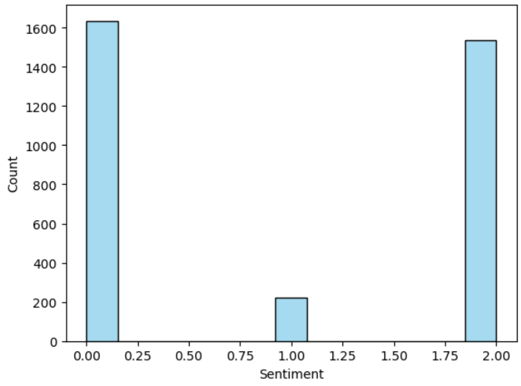


Figure 2. Unbalanced label graphics

Then the label was balanced prior to further processing. The data balancing process used the SWATCH method. **Figure 3** shows the balanced labels.

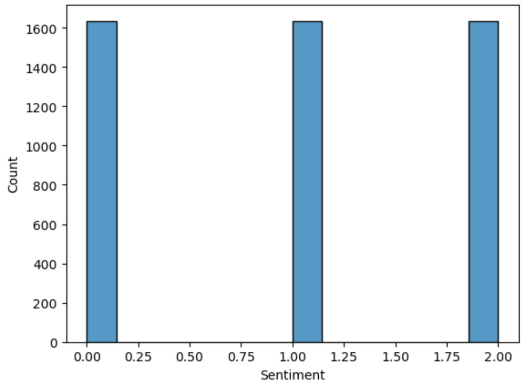
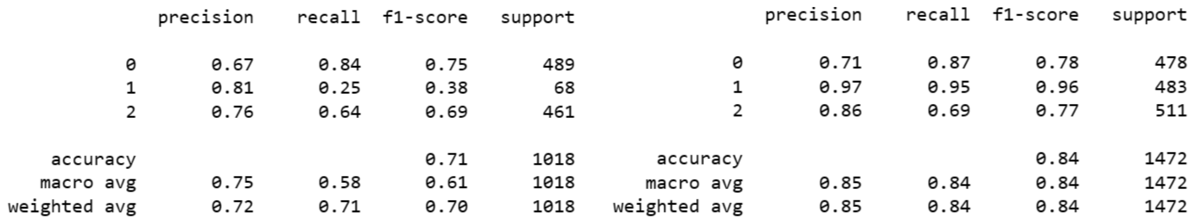


Figure 3. Balanced label graphics

After the data balancing process was carried out, the next step was to preprocess and weight words with TF-IDF.

B. Modeling with the Random Forest algorithm

After the data cleaning and word weighting process, the next step was to model with a random forest algorithm. The first trial was to conduct tests using the random forest algorithm. Figure 1 is the result of Random Forest (RF) using SMOTE and without SMOTE.



```
              precision    recall  f1-score   support

           0       0.67      0.84      0.75       489
           1       0.81      0.25      0.38        68
           2       0.76      0.64      0.69       461

    accuracy                           0.71      1018
   macro avg       0.75      0.58      0.61      1018
weighted avg       0.72      0.71      0.70      1018
```

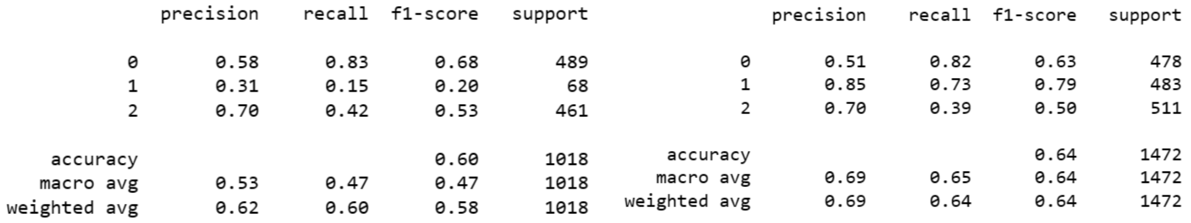
a. RF

```
              precision    recall  f1-score   support

           0       0.71      0.87      0.78       478
           1       0.97      0.95      0.96       483
           2       0.86      0.69      0.77       511

    accuracy                           0.84      1472
   macro avg       0.85      0.84      0.84      1472
weighted avg       0.85      0.84      0.84      1472
```

b. SMOTE+RF

Figure 4. Random Forest accuracy results

The results obtained from RF modeling was shown in Figure 4a. The obtained accuracy was only 71%, but when the data was balanced using SMOTE, the results increased significantly in Figure 4b, which was 13% or 84%. Then the RF test used the stacking technique, figure 5 is the result of RF accuracy using the stacking technique.



```
              precision    recall  f1-score   support

           0       0.58      0.83      0.68       489
           1       0.31      0.15      0.20        68
           2       0.70      0.42      0.53       461

    accuracy                           0.60      1018
   macro avg       0.53      0.47      0.47      1018
weighted avg       0.62      0.60      0.58      1018
```

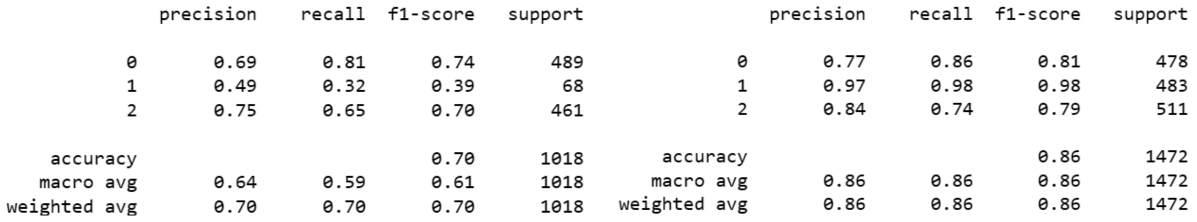
a. RF+Adaboost (Stacking)

```
              precision    recall  f1-score   support

           0       0.51      0.82      0.63       478
           1       0.85      0.73      0.79       483
           2       0.70      0.39      0.50       511

    accuracy                           0.64      1472
   macro avg       0.69      0.65      0.64      1472
weighted avg       0.69      0.64      0.64      1472
```

b. SMOTE+RF+Adaboost (Stacking)

Figure 5. Random Forest accuracy results using the Stacking technique

From Figure 5, it can be seen that the use of the ensemble method with stacking techniques has a significant decreased, as seen in Figure 5a, which is down 11% from the RF base without SMOTE, which is 60%. After adding the SMOTE as seen in Figure 5b, accuracy increased by 4% to 64%. Then the next test used RF with the Boosting technique, Figure 6 is the result of RF accuracy using the Boosting Technique.



```
              precision    recall  f1-score   support

           0       0.69      0.81      0.74       489
           1       0.49      0.32      0.39        68
           2       0.75      0.65      0.70       461

    accuracy                           0.70      1018
   macro avg       0.64      0.59      0.61      1018
weighted avg       0.70      0.70      0.70      1018
```

a. RF+Adaboost (Boosting)

```
              precision    recall  f1-score   support

           0       0.77      0.86      0.81       478
           1       0.97      0.98      0.98       483
           2       0.84      0.74      0.79       511

    accuracy                           0.86      1472
   macro avg       0.86      0.86      0.86      1472
weighted avg       0.86      0.86      0.86      1472
```

b. SMOTE+RF+Adaboost (Boosting)

Figure 6. Random Forest accuracy results using Boosting Technique

In Figure 6a, it can be seen that the use of Boosting performed better than stacking but not better than based RF, with 70% accuracy. Then in Figure 6b it can be seen that the significant increase when the boost was combined to SMOTE. It increased to 86%.

C. Comparison

Based on the tests that have been carried out, it was found that the use of SMOTE has a significant effect on the random forest algorithm, both on the base model, Adaboost, and Boosting. Table 1 is a comparison of the tests carried out in this study.

Table 1. Model Comparison

It	Accuracy Comparison		
	Type	***Without SMOTE***	***SMOTE***
1	RF	71%	84%
2	RF+Adaboost (Stacking)	60%	64%
3	RF+Adaboost (Boosting)	70%	86%

Table 1 showed that that the highest accuracy improvement was obtained in RF with the Boosting technique with 86%. In stacking, there was a significant accuracy drop to 60%, although it increased with SMOTE, but it only reach 64%. This indicated that the Stacking technique in Random Forest did not work well using data from social media X.

Conclusion

This study has several conclusions. Firstly, the best accuracy in this study was 86% obtained by the SMOTE+RF+Adaboost (Boosting) model. Then the stacking technique on social media X data did not perform well, because it only achieved 64%. Furthermore, the use of the SMOTE method to balance classes or labels greatly affected the entire model used.

This study also has a limitation, such as using only the default parameters of the RF algorithm, the parameters used were still random, so it is necessary to conduct several tests until the best accuracy was obtained. The need to use the hyperparameter tuning method to make it easier to get the best parameters. Several types of hyperparameter tuning that can be used are random search, optuna, grid search and others.

Acknowledgement

Thank you to Universitas Putra Indonesia "YPTK" Padang for facilitating and funding this research.

References

[1] L. C. Maretva and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," *Jurnal Masyarakat Informatika*, vol. 13, no. 1, pp. 33–44, 2022, doi: 10.14710/jmasif.13.1.42912.

[2] Y. Pristyanto, "Penerapan Metode Ensemble Untuk Meningkatkan Kinerja Algoritme Klasifikasi Pada Imbalanced Dataset," *Jurnal TEKNOINFO*, vol. 13, no. 1, pp. 11–16, 2019, doi: 10.33365/jti.v13i1.184.

[3] Y. M. Awaludin and F. Budiman, "Optimasi Analisis Kesuburan Tanah Dengan Pendekatan Soft Voting Ensemble," *Jurnal SIMETRIS*, vol. 14, no. 2, pp. 261–275, 2023, doi: 10.24176/simet.v14i2.11285.

[4] A. K. Putri and H. Suparwito, "Uji Algoritma Stacking Ensemble Classifier pada Kemampuan Adaptasi Mahasiswa Baru dalam Pembelajaran Online," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 3, no. 1, pp. 1–12, 2023.

[5] S. Wijayanti and Hartiningrum, "Dampak Aplikasi Pinjaman Online Terhadap Kebutuhan Dan Gaya Hidup Konsumtif Buruh Pabrik," *Mizania: Jurnal Ekonomi dan Akuntansi*, vol. 2, no. 2, pp. 230–235, 2022, doi: 10.47776/mizania.v2i2.592.

[6] J. Z. Y. Arvante, "Dampak Permasalahan Pinjaman Online dan Perlindungan Hukum Bagi Konsumen Pinjaman Online," *Ikatan Penulis Mahasiswa Hukum Indonesia Law Journal*, vol. 2, no. 1, pp. 73–87, Feb. 2022, doi: 10.15294/ipmhi.v2i1.53736.

[7] A. Priyonggojati, "Perlindungan Hukum Terhadap Penerima Pinjaman Dalam Penyelenggaraan Financial Technology Berbasis Peer To Peer Lending," *Jurnal Usm Law Review*, vol. 2, no. 2, p. 162, Nov. 2019, doi: 10.26623/julr.v2i2.2268.

[8] Y. Setiawan, J. Jondri, and W. Astuti, "Twitter Sentiment Analysis on Online Transportation in Indonesia Using Ensemble Stacking," *Jurnal Media Informatika Budidarma*, vol. 6, no. 3, pp. 1452–1458, Jul. 2022, doi: 10.30865/mib.v6i3.4359.

[9] R. Jayapermana, A. Aradea, and N. I. Kurniati, "Implementation of Stacking Ensemble Classifier for Multi-class Classification of COVID-19 Vaccines Topics on Twitter," *Scientific Journal of Informatics*, vol. 9, no. 1, pp. 8–15, May 2022, doi: 10.15294/sji.v9i1.31648.

[10] H. N. Cahyani and R. Arifudin, "Improving the Accuracy of Multinomial Naïve-Bayes Algorithm with Adaptive Boosting Using Information Gain for Classification of Movie Reviews Sentiment Analysis," *Journal of Advances in Information Systems and Technology*, vol. 4, no. 1, 2022, doi: 10.15294/jaist.v4i1.60267.

[11] S. M. Ganie, P. K. D. Pramanik, M. Bashir Malik, S. Mallik, and H. Qin, "An ensemble learning approach for diabetes prediction using boosting techniques," *Front Genet*, vol. 14, pp. 1–15, 2023, doi: 10.3389/fgene.2023.1252159.

[12] M. K. Anam, M. I. Mahendra, W. Agustin, Rahmaddeni, and Nurjayadi, "Framework for Analyzing Netizen Opinions on BPJS Using Sentiment Analysis and Social Network Analysis (SNA)," *Intensif*, vol. 6, no. 1, pp. 2549–6824, 2022, doi: 10.29407/intensif.v6i1.15870.

[13] N. Chamidah, M. M. Santoni, and N. Matondang, “Pengaruh Oversampling pada Klasifikasi Hipertensi dengan Algoritma Naïve Bayes, Decision Tree, dan Artificial Neural Network (ANN),” *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 635–641, 2020, doi: 10.29207/resti.v4i4.2015.

[14] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, “Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 677–690, Jul. 2022, doi: 10.30812/matrik.v21i3.1726.

[15] M. K. Anam *et al.*, “Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm,” *ILKOM Jurnal Ilmiah*, vol. 15, no. 2, pp. 290–302, 2023, doi: 10.33096/ilkom.v15i2.1590.290-302.

[16] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, “Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 153–160, Oct. 2023, doi: 10.57152/malcom.v3i2.897.

[17] M. K. Anam, Rahmaddeni, M. B. Firdaus, H. Asnal, and Hamdani, “Sentiment Analysis to analyze Vaccine Enthusiasm in Indonesia on Twitter Social Media,” *JAIA – Journal Of Artificial Intelligence And Applications*, vol. 1, no. 2, pp. 23–27, 2021.

[18] A. N. Ulfah and M. K. Anam, “Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM),” *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 7, no. 1, pp. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[19] R. S. Putra, W. Agustin, M. K. Anam, L. Lusiana, and S. Yaakub, “The Application of Naïve Bayes Classifier Based Feature Selection on Analysis of Online Learning Sentiment in Online Media,” *Jurnal Transformatika*, vol. 20, no. 1, p. 44, Jul. 2022, doi: 10.26623/transformatika.v20i1.5144.

[20] H. A. Putranto, O. Setyawati, and Wijono, “Pengaruh Phrase Detection dengan POS-Tagger terhadap Akurasi Klasifikasi Sentimen menggunakan SVM,” *JNTETI*, vol. 5, no. 4, pp. 252–259, 2016.

[21] A. N. Ulfah, M. K. Anam, N. Y. S. Munti, S. Yaakub, and M. B. Firdaus, “Sentiment Analysis of the Convict Assimilation Program on Handling Covid-19,” *JUITA: Jurnal Informatika*, vol. 10, no. 2, pp. 209–216, 2022, doi: 10.30595/juita.v10i2.12308.

[22] R. Latifah, E. S. Wulandari, and P. E. Kreshna, “Model Decision Tree untuk Prediksi Jadwal Kerja menggunakan Scikit-Learn,” in *Seminar Nasional Sains dan Teknologi*, 2019, pp. 1–6.

[23] C. H. Yutika, A. Adiwijaya, and S. Al Faraby, “Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes,” *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, p. 422, Apr. 2021, doi: 10.30865/mib.v5i2.2845.

[24] V. Talasila, M. V Mohan, and N. M. R, “Enhancing Text-to-Image Synthesis with an Improved Semi-Supervised Image Generation Model Incorporating N-Gram, Enhanced TF-IDF, and BOW Techniques,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 7s, pp. 381–397, 2023.

[25] R. Kaur and V. Bhardwaj, “Gurmukhi Text Emotion Classification System using TF-IDF and N-gram Feature Set Reduced using APSO,” *International Journal on Emerging Technologies*, vol. 10, no. 3, pp. 352–362, 2019.