**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 467−473
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1902

# Analysis of Community Sentiment on Twitter towards COVID-19 Vaccine Booster Using Ensemble Stacking Methods

**Syifa Khairunnisa Salsabila[1], Jondri[2,*], Widi Astuti[3]**

School of Computing, Informatics Study Program, Telkom Univeristy, Bandung, Indonesia
Email: [1]syifaks@student.telkomuniveristy.ac.id, [2,*]jondri@ telkomuniveristy.ac.id, [3]widiwdu@telkomuniversity.ac.id
Email Penulis Korespondensi: jondri@ telkomuniveristy.ac.id
Submitted:**21/07/2022**; Accepted:**18/08/2022**; Published**: 30/09/2022**

**Abstract**−The outbreak of the COVID-19 virus in Indonesia has not ended until the government has made various efforts to reduce this outbreak, such as the Large-Scale Social Restriction (PSBB) policy and the obligation of the entire community to vaccinate against COVID-19. The government has made a new policy for the community: booster vaccination for people who have already been vaccinated against COVID-19 1 and vaccinated against COVID-19 2. With this new policy, many people have given opinions on social media. One of them is Twitter social media. Positive and negative opinions given by Twitter users can be used as a source of information data. Because of these problems, researchers conducted a sentiment analysis of the booster vaccine using the Ensemble Stacking method. The dataset that has collected as many as 6,500 data from Twitter will be grouped into positive and negative class sentiments. The best results from this study using ensemble stacking and oversampling have an accuracy value of 80%.

**Keywords**: Vaccinate; Booster; Twitter; Ensemble Stacking; Sentiment Analysis

## 1. INTRODUCTION

The outbreak of the COVID-19 virus has made many changes in aspects of life in society, from the economic and educational to social aspects of experiencing the impact of this virus [1]. The government continues to seek ways to reduce the spread of the COVID-19 virus. The efforts included the Large-Scale Social Restriction (PSBB) policy [2] to the obligation of the entire community to carry out COVID-19 vaccinations to booster vaccinations. Since the first injection of the COVID-19 vaccine to President Joko Widodo at the Presidential Palace. The government continues encouraging the public to vaccinate the first and second doses of the Covid-19 vaccine. This national vaccination program was continued until the Ministry of Health (Kemenkes) through the Directorate General (Directorate General) of Disease Prevention and Control issues Circular (SE) Number HK.02.02/II/252/2022 concerning Advanced Dose (Booster) COVID-19 Vaccination [3]. Based on data from the Ministry of Health, 40.37 million booster vaccines have been given to the public as of May 3, 2022 [4].

The Circular provides many responses and opinions to the public on various social media. Although the government has socialized it, there are still many people who still do not understand the use of vaccines. In addition, many people are afraid or hesitant to be given the COVID-19 vaccine [5]. The many types of vaccines circulating in the market and the different side effects of each vaccine brand are also of the most common discussions on social media [6]. Based on data from We Are Social, the number of social media users in Indonesia will be more than 190 million. Twitter is one of the social media used by many people, which reaches 58% of the total social media users in Indonesia [7]. Due to a large number of Twitter social media users, many Indonesian people use tweets as a feature or tool to interact in disseminating information [8].

Every tweet made by the community contains sentiment. Sentiment analysis is a field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. Sentiment analysis is done to see opinions on an issue, whether they tend to be positive or negative opinions [9].

In a previous study, 2021 conducted by W. Yulita, E. D. Nugroho, and M. H. Algifari with the research topic of sentiment analysis on public opinion about the COVID-19 vaccine by implementing the Naive Bayes method. The results found that the positive sentiment was 60.3%, the neutral sentiment was 34.4%, and the negative sentiment was 5.4%. Furthermore, it produces an accuracy value of 93% from 3780 tweet data processed [10]. In a similar study in 2021 on sentiment analysis with a different method, H. Hayati and M. R. Alifi conducted a similar study. The author implements the Support Vector Machine algorithm that uses 360,000 tweet data with 120,000 data labeled positive, negative, and neutral each. This study resulted in a value of 84% of all aspects of measurement, namely accuracy, recall, f-measure, and precision [11].

In 2021, a similar study will be conducted in a study conducted by A. Baita, Y. Pristyanto, and N. Cahyono, only focusing on one type of vaccine, namely the Sinovac vaccine. The method used in this research is Support Vector Machine and K-Nearest Neighbor. With negative sentiment of 52%, much greater than the positive sentiment of 18%. Likewise, it is smaller than neutral sentiment, which has a value of 31%. This study resulted in an accuracy rate of 81% of the 221 processed data [12]. Different results are shown in the 2021 study conducted by F. Fitriana, E. Utami, and H. Al Fatta. The author applies the Support Vector Machine (SVM) and Naïve Bayes algorithms on the topic of the covid-19 vaccine. Produces the accuracy rate of 90.47% using Support Vector Machine, while the Naïve Bayes accuracy result is 88.64%. The author says that the Support Vector Machine algorithm is superior in accuracy performance, while Naïve Bayes is superior in timing performance [13].

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 467−473
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1902

In a study conducted by A. K. Santoso, A. Noviriandini, A. Kurniasih, B. D. Wicaksono, and A. Nuryanto, the topic of Twitter users' perceptions of COVID-19. The method used is Logistic Regression by varying the L2 and None hyperparameters. Furthermore, the results on the L2 hyperparameter obtained an accuracy value of 77% and an F1 score of 74%. Furthermore, in the hyperparameter variation None, the accuracy value is 74%, and the F1 Score is 70%. Therefore, the author says that the L2 hyperparameter value is the best variation on the Logistic Regression method [14].

In previous studies that have been carried out, the authors only used one or two classification algorithms. In this study, we will use a different method: Ensemble Stacking on the COVID-19 Vaccine Booster sentiment. The purpose of this study is to determine the level of accuracy in the Ensemble Stacking method with datasets resulting from collecting or crawling data on Twitter with the specified time and keywords and tweets taken using Indonesian-language tweets with sentiment grouping into two classes, namely positive sentiment and negative sentiment.

## 2. SEARCH METHODOLOGY

### 2.1 Research Stages

The stages in this research are crawling data through Twitter social media using the Twitter API, then preprocessing to clean the data, then giving sentiment labels manually, which is done by three people, then feature extraction using TF-IDF, then data sharing, then stage 3 classification basis, followed by meta classification, and evaluated using a confusion matrix. The research stages are used in Figure 1.

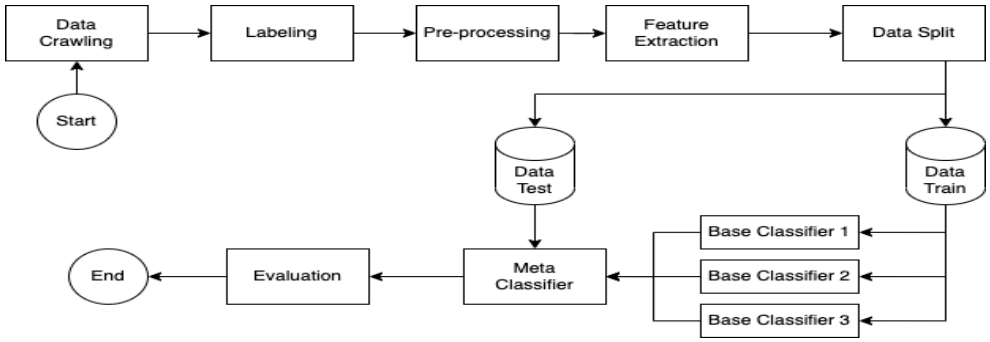


**Figure 1.** System Model

### 2.2 Data

Dataset collection from Twitter social media in real-time using the Twitter API (Application Programming Interface) by using keywords related to the title, namely 'booster vaccine'. The dataset used in this study is in Indonesian. The dataset collection started in March 2022 and continued until May 2022, which resulted in 6,500 tweet data from the data selection results. Then the dataset that has been collected is labeled with sentiment in the form of positive and negative sentiment, which is done manually by three people. Furthermore, after labeling, the dataset will be split into two parts: train data and test data, where 80% train data and 20% test data. This method is obtained from the results of TF-IDF which divides the data into two 80:20 parts.

### 2.3 Pre-processing

Pre-processing data is the stage of preparing text data so that it can be used in the next process. For example, the pre-processing used in this study is as follows.

a. Cleansing data is a process of reducing symbols or punctuation marks, emoticons, and URLs.
b. Case Folding is the process of converting capital letters to lowercase letters
c. Tokenization is the process of separating a sentence into several words.
d. Stopword Removal is a process of removing words that do not have significant meaning, and the word is not used.
e. Stemming is the process of removing affixes and returning them to base words.

### 2.4 Feature Extraction

Feature extraction is used to calculate the weight of each word, calculated from how often a word appears in the document that aims to describe the identity of an object that will help in this study [15]. In this study, feature extraction uses the Term Frequency - Inverse Document Frequency (TF-IDF) algorithm.

$$W_{dt} = tf_{dt} * IDF_t \quad (1)$$

Equation (1) shows $W_{dt}$ is the value of the d document in the t word, then $TF_{dt}$ is the number of words searched for in a document, and $IDF_{ft}$ is the Inverse Document Frequency ($\log\left(\frac{N}{df}\right)$). The value of $N$ is the number of documents, and $df$ is the number of documents containing the searched word.

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 467−473
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1902

### 2.5 Data Split

The data is divided into two parts: train data and test data, where 80% train data and 20% test data. This method is obtained from the results of TF-IDF which divides the data into two 80:20 parts.

### 2.6 Classification

This study applies the model using ensemble stacking with three base classifiers, namely Support Vector Machine (SVM), Naïve Bayes, Logistic Regression and the meta classifier using Logistic Regression.

#### 2.6.1 Ensemble Stacking

The ensemble stacking method uses several base classifiers in the learning process. There are two stages in learning to stack. In stage 1, each base classifier used is trained using the same dataset to produce its prediction results. Stage 2, the meta classifier takes the prediction results from the base classifier as input to determine which class is the most likely to test data [16].

#### 2.6.2 Support Vector Machine (SVM)

Support Vector Machine is a supervised classifier algorithm that uses a hyperplane, which aims to separate one class from another. The simple concept of SVM will be to find a support vector and then look for the boundary line with the maximum margin [17].

#### 2.6.3 Naïve Bayes

Naive Bayes is one of the basic machine learning methods to find the highest probability value. Naive Bayes has the main characteristic that it is assumed that the dataset features are not related to other features in the same dataset [18].

$$P(c|w) = \frac{P(c)\,P(w|c)}{P(w)} \tag{2}$$

There is equation (3) showing $P(c|w)$ is the probability of the hypothesis based on the conditions, then $P(c)$ is the probability of the hypothesis, then $P(w|c)$ is the probability based on the conditions of the hypothesis, and $P(w)$ is the probability $w$.

#### 2.6.4 Logistic Regression

Logistic regression is the baseline of the supervised machine learning algorithm for classification that connects one or several independent variables (independent variables) with the dependent variable in the form of categories [19].

### 2.7 Evaluation

At the evaluation stage, to see the suitability of the model applied in this study, it is then measured using a confusion matrix to compare the results of two different datasets in tabular form. A confusion matrix with a table containing recall, accuracy, precision, and f1-score.

True Positive (TP) : the result is true
True Negative (TN) : no correct result
False Positive (FP) : an unexpected result
False Negative (FN) : wrong result

a. Accuracy is the level of accuracy in classifying correctly.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

b. Precision is the level of truth between the requested information and the system's answer.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

c. Recall is the success rate of the system in retrieving information.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

d. F1-Score is a comparison of the average precision and recall that has been weighted.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{\mathrm{Presicion} + \mathrm{Recall}} \tag{6}$$

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 467−473
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1902

# 3. RESULT AND DISCUSSION

## 3.1 Data

The data crawling process comes from the media about Twitter in the form of tweets using the keywords "vaccine booster," "vaccine," and "booster," with a period from March 2022 to June 2022, which produces 8,752 tweet data. Data collection is done manually using the Twitter API, which requires an access token obtained after turning the author's Twitter account into a developer account.

The results of data collection that have been carried out then enter the tweet selection process so that the tweets that will be processed follow the sentiment topics discussed in this study. Irrelevant Tweets will be delete. The results of this tweet selection make the data that was original 8,752 tweets become 6,500 data with 3,560 positive and 2,940 negative data.

## 3.2 Result of Text Pre-processing

After the data has been successfully selected, it enters the pre-processing stage, which is the stage for cleaning the data by eliminating noise. Below are the results of each pre-processing stage.

### 3.2.1 Cleansing Data

Dirty data will be cleaned by removing special characters, URLs, numbers, emojis, and whitespace. The results of data cleansing can be seen in the table below.

**Table 1.** Cleansing Data

| Before Cleasning | After Cleansing |
|---|---|
| Kukira sudah sembuh, ternyata cuma ilang sesaat, sakit lagi semua persendian badan, gini amat habis vaksin booster ya. | Kukira sudah sembuh ternyata cuma ilang sesaat sakit lagi semua persendian badan gini amat habis vaksin booster ya |

### 3.2.2 Case Folding

Then the data is case folded to change the capital letters to lowercase letters. The results of the case folding data can be seen in the table below.

**Table 2.** Case Folding

| Before Case Folding | After Case Folding |
|---|---|
| Kukira sudah sembuh ternyata cuma ilang sesaat sakit lagi semua persendian badan gini amat habis vaksin booster ya | kukira sudah sembuh ternyata cuma ilang sesaat sakit lagi semua persendian badan gini amat habis vaksin booster ya |

### 3.2.3 Tokenizing

The data is then tokenization to separate the sentences into several words. The results of tokenization can be seen in the table below.

**Table 3.** Tokenization

| Before Tokenization | After Tokenization |
|---|---|
| kukira sudah sembuh ternyata cuma ilang sesaat sakit lagi semua persendian badan gini amat habis vaksin booster ya | ['kukira', 'sudah', 'sembuh', 'ternyata', 'cuma', 'ilang', 'sesaat', 'sakit', 'lagi', 'semua', 'persendian', 'badan', 'gini', 'amat', 'habis', 'vaksin', 'booster', 'ya'] |

### 3.2.4 Stopword Removal

The data is then carried out in Indonesian language stopword removal. The results of stopword removal can be seen in the table below.

**Table 4.** Stopword Removal

| Before Stopword Removal | After Stopword Removal |
|---|---|
| ['kukira', 'sudah', 'sembuh', 'ternyata', 'cuma', 'ilang', 'sesaat', 'sakit', 'lagi', 'semua', 'persendian', 'badan', 'gini', 'amat', 'habis', 'vaksin', 'booster', 'ya'] | ['kukira', 'sembuh', 'ilang', 'sakit', 'persendian', 'badan', 'gini', 'habis', 'vaksin', 'booster'] |

### 3.2.5 Stemming

The steaming process then carries out the data. The results of stemming can be seen in the table below.

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 467−473
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1902

**Table 5.** Stemming

| Before Stemming | After Stemming |
|---|---|
| ['kukira', 'sembuh', 'ilang', 'sakit', 'persendian', 'badan', 'gini', 'habis', 'vaksin', 'booster'] | ['kira', 'sembuh', 'ilang', 'sakit', 'sendi', 'badan', 'gini', 'habis', 'vaksin', 'booster'] |

### 3.3 TF-IDF

After the pre-processing stage, enter the word weighting stage by implementing the Term Frequency - Inverse Document Frequency (TF-IDF) algorithm, which works by calculating the weight of each word and displaying the results of the word frequency score with a high frequency of occurrence in the document. Then the data is split with a ratio of 80:20 with 5,200 train data and 1,300 test data. The following results from word weighting with TF-IDF were applied to the dataset.

| | aa | aaa | aaaa | aaaaa | aaaaaaaa | aaaaaaaaaarrrrggghhhhh | aaah | aaakkkkkk | aaamiinnnn | aakkkhhhhh | ... | zii | zivifax | zombie | zon | zonk | zoom | zulfan | zuzur | zzz | zzzzz |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 6495 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 6496 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 6497 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 6498 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 6499 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

**Figure 2.** Result of TF-IDF

In Figure 2 it can be seen that a word contained in the corpus is sorted alphabetically and then the probability of a word appearing in a document is calculated.

### 3.4 Classification

In this study, implementing ensemble stacking, the first stage will classify each base classifier, namely Naive Bayes, Support Vector Machine, and Logistic Regression, to train on the same train data to get predictive results from each base classifier. After getting the prediction results from the three base classifiers, enter the meta classifier stage with the addition of the Logistic Regression algorithm to optimise the final prediction results. This research will be carried out with three scenarios. In the first scenario, the meta classifier on the dataset does not use stemming, then the second scenario on the dataset will use oversampling, and the third scenario on the dataset will use undersampling.

### 3.5 Model Evaluation

At the model evaluation stage, the dataset applied in this study implements the ensemble stacking method, which has two stages. In the first stage, the base classifier uses the Support Vector Machine, Naive Bayes, and Logistic Regression algorithms, where each base classifier will be trained using trained data. Table 7 shows that the base classifier with the highest accuracy value is the Support Vector Machine, with an accuracy value of 78.5%. The following results from each base classifier trained using the data train.

**Table 6.** Performance of Base Classifier

| Method | Class | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| Support Vector Machine | Positive | 80% | 80% | 80% | **78.5%** |
| | Negative | 74% | 75% | 74% | |
| Naïve Bayes | Positive | 81% | 79% | 80% | 77.8% |
| | Negative | 74% | 76% | 75% | |
| Logistic Regression | Positive | 81% | 81% | 81% | 78.3% |
| | Negative | 75% | 76% | 75% | |

Furthermore, the second stage is a meta classifier that takes the prediction results from the first stage (base classifier) and then compares it with the test data added by a new model, namely Logistic Regression, to get more optimal prediction results. It can be seen in table 8 that at the meta classifier stage, the accuracy becomes better to 79%. The following are the results of the meta classifier.

**Table 7.** Performance of Meta Classifier

| Method | Class | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | Positive | 82% | 79% | 81% | **79%** |
| | Negative | 75% | 78% | 76% | |

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 467−473
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1902

Furthermore, in this study, in the first scenario stage, the meta classifier will test the test data without stemming using the same model as before, namely Logistic Regression, to see the difference. When comparing table 8 and table 9, that meta classifier using stemming has a higher accuracy value than beyond. Meta classifier with no stemming has an accuracy value of 78%. The following is the result of the meta classifier not using stemming.

**Table 8.** Performance of Meta Classifier without Stemming

| Method | Class | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression (without stemming) | Positive | 81% | 79% | 80% | 78% |
| | Negative | 74% | 76% | 75% | |

Then in the second scenario, oversampling will be carried out because the dataset in this study is not balanced. Oversampling has several methods; one method is the Random Over Sampling (ROS) method. The use of oversampling aims to add some data to the minority class so that the data in the minority class becomes more balanced against the majority class [20]. The minority class in this study is negative data of 2,940. After oversampling, the number of negative data becomes 3,560 and becomes balanced with positive data. After oversampling, the final result in this study has an accuracy rate of 80%.

Moreover, in the third scenario, undersampling is carried out, which aims to reduce the data in the majority class so that the data in the majority class becomes more balanced against the minority class. The method used in undersampling is Random Under Sampling (RUS). The majority class in this study is found in positive data, amounting to 3,560 data. After undersampling, the positive data becomes 2,960 and is balanced with negative data. After undersampling, the final result in this study has an accuracy rate of 77%. The results of the classification using oversampling and undersampling can be seen in the following table.

**Table 9.** Performance of Meta Classifier Using Oversampling and Undersampling

| Method | Class | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression (with oversampling) | Positive | 83% | 77% | 80% | **80%** |
| | Negative | 77% | 83% | 80% | |
| Logistic Regression (with undersampling) | Positive | 78% | 74% | 76% | 77% |
| | Negative | 76% | 80% | 78% | |

## 4. CONCLUSION

The results of this study indicate that the Support Vector Machine as a base classifier has the highest accuracy value of 78.5% compared to Naive Bayes and Logistic Regression. However, the performance results are better using the ensemble stacking method, with an accuracy value of 79%. Compared to ensemble stacking without stemming, the value obtained is 78%. That shows that the stemming stage dataset affects the final performance result. However, the use of oversampling on the dataset produces an accuracy value of 80%, while the use of undersampling on the dataset produces an accuracy value of 77%. The results of oversampling and undersampling show that with more datasets, it will have more optimal performance results.

## REFERENCES

[1] D. T. Purnama, V. Juliansyah, and Chainar, "PANDEMI COVID-19, PERUBAHAN SOSIAL DAN KONSEKUENSINYA PADA MASYARAKAT," Proyeksi: Jurnal Ilmu Sosial dan Humaniora, vol. 25, 2020.
[2] "Peta Sebaran | Covid19.go.id." https://covid19.go.id/peta-sebaran (accessed Apr. 01, 2022).
[3] "SE terkait Penambahan Regimen Vaksinasi COVID-19 Dosis Lanjutan (Booster) bagi sasaran yang mendapat vaksin primer Sinovac | Covid19.go.id." https://covid19.go.id/artikel/2022/05/26/se-terkait-penambahan-regimen-vaksinasi-covid-19-dosis-lanjutan-booster-bagi-sasaran-yang-mendapat-vaksin-primer-sinovac (accessed Apr. 03, 2022).
[4] C. Annur, "Capaian Vaksin Booster di Indonesia Baru 19,39% | Databoks," May 04, 2022. https://databoks.katadata.co.id/datapublish/2022/05/04/capaian-vaksin-booster-di-indonesia-baru-1939 (accessed Apr. 03, 2022).
[5] E. Pedersen, L. Loft, S. Jacobsen, B. Søborg, and J. Bigaard, "Strategic health communication on social media: Insights from a Danish social media campaign to address HPV vaccination hesitancy," Vaccine, vol. 38, Jul. 2020, doi: 10.1016/j.vaccine.2020.05.061.
[6] L. A. Octafia, "Vaksin Covid-19: Perdebatan, Persepsi dan Pilihan.," Jurnal Emik, vol. 4, no.2, 2021
[7] S. Kemp, "Digital 2022: Indonesia — DataReportal – Global Digital Insights," Feb. 15, 2022. https://datareportal.com/reports/digital-2022-indonesia (accessed Apr. 06, 2022).
[8] S. Dewanty Rehatta, E. Sediyono, and I. Sembiring, "JURNAL MEDIA INFORMATIKA BUDIDARMA Analisis Penyebaran Informasi Vaksin Covid-19 Pada Twitter Menggunakan Kolaborasi SNA dan Sentiment Analysis," 2022, doi: 10.30865/mib.v6i2.3955.
[9] B. Liu, Sentiment analysis and opinion mining. Springer, 2012.
[10] W. Yulita, E. D. Nugroho, and M. H. Algifari, "Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier," Jurnal Data Mining dan Sistem Informasi, vol. 2, no. 2, pp. 1–9, 2021.

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 467−473
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1902

[11] H. Hayati and M. R. Alifi, "ANALISIS SENTIMEN PADA TWEET TERKAIT VAKSIN COVID-19 MENGGUNAKAN METODE SUPPORT VECTOR MACHINE," Jurnal Teknologi Terapan, vol. 7, no. 2, Aug. 2021.
[12] A. Baita, Y. Pristyanto, and N. Cahyono, "ANALISIS SENTIMEN MENGENAI VAKSIN SINOVAC MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DAN K-NEAREST NEIGHBOR (KNN)," Information System Journal (INFOS), vol. 4, no. 2, 2021.
[13] F. Fitriana, E. Utami, and H. al Fatta, "Analisis Sentimen Opini Terhadap Vaksin Covid-19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes," Jurnal Komtika (Komputasi dan Informatika), vol. 5, no. 1, May 2021.
[14] A. K. Santoso, A. Noviriandini, A. Kurniasih, B. D. Wicaksono, and A. Nuryanto, "KLASIFIKASI PERSEPSI PENGGUNA TWITTER TERHADAP KASUS COVID-19 MENGGUNAKAN METODE LOGISTIC REGRESSION," Jurnal Informatika Kaputama, vol. 5, no. 2, Jul. 2021.
[15] D. Satria, "Perbandingan Metode Ekstraksi Ciri Histogram dan PCA untuk Mendeteksi Stoma pada Citra Penampang Daun Freycinetia Comparison of Histogram and PCA as Feature Extraction Methods in Detecting Stoma in Freycinetia Leaf Images", [Online]. Available: http://journal.ipb.ac.id/index.php.jika
[16] R. Ananda Fitriansyah, "Penerapan Ensemble Stacking Untuk Klasifikasi Multi Kelas," 2016. [Online]. Available: http://ars.ilkom.unsri.ac.id
[17] A. Wibowo Haryanto and E. Kholid Mawardi, "Influence of Word Normalization and Chi-squared Feature Selection on Support Vector Machine (SVM) Text Classification."
[18] R. Feldman and J. Sanger, The text mining handbook : advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.
[19] H. Wooi, A. Ramli, and A. Kumar, "Designing Early Warning System: Prediction Accuracy of Currency Crisis by Using k-Nearest Neighbour Method A Novel Nearest Neighbour Tree for Currency Crisis Forecast ing Nor Azuana Ramli An Analysis on T wo Different Dat a Set s by using Ensemble of k-Nearest Neighbor Classifiers Noraida Ramli Comparat ive St udy of Classificat ion Techniques for Breast Cancer Diagnosis."
[20] A. Y. Triyanto and R. Kusumaningrum, "Implementasi Teknik Sampling untuk Mengatasi Imbalanced Data pada Penentuan Status Gizi Balita dengan Menggunakan Learning Vector Quantization Implementation of Sampling Techniques for Solving Imbalanced Data Problem in Determination of Toddler Nutritional Status using Learning Vector Quantization," vol. 19, pp. 39–50, 2017.