Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Computer Science and Computational Intelligence (ICCSCI 2024)

Exploration of handling imbalanced data in hate speech detection on social media using machine learning models

Devita Wijaya Putri, Faldian Chandra Bulyono, Ghinaa Zain Nabiilah, Jurike V. Moniaga

Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Abstract

Various social media platforms provide open and very easy access for individuals to share content and exchange opinions through the comments column. However, with advances in technology, it turns out that there are still many people who cannot use it wisely, the spread of hate speech on social media has become an alarming phenomenon. Comments or tweets that are offensive can cause division, discrimination, and can have a negative impact on other people's lives. Individuals who are targets of hate speech are vulnerable to experiencing psychological disorders that can disrupt their mental health. Therefore, machine learning models are needed to analyze whether a comment from someone contains hate speech or not and to prevent hate speech from appearing on social media. The aim of this research is to compare the performance of the SVM and KNN algorithms in detecting hate speech. However, the drawback in developing machine learning models is the limited and unbalanced amount of data. So in this research an exploration will be carried out regarding handling dataset imbalances using a re-sampling technique, namely random over sampling. Experimental results show that the SVM method has the best performance after using the re-sampling technique with an accuracy of 96.1% from the previous 95.6%. This proves that using re-sampling techniques can improve model performance significantly.

Peer-review under responsibility of the scientific committee of the 9th International Conference on Computer Science and Computational Intelligence 2024
Keywords: Hate Speech, Support Vector Machine, K-Nearest Neighbor, Machine Learning, Imbalance, Re-sampling, Random Over Sampling.

* Corresponding author. Tel.: +6280-4169-6969
E-mail address: ghinaa.nabiilah@binus.ac.id

1. Introduction

Technological advances are currently growing rapidly, various kinds of technology have been created in this modern era. Social media is one of the most frequently used information technologies, social media is a digital platform that allows users to interact, share content and communicate with each other remotely. In 2023, internet users in Indonesia reached 213 million users out of Indonesia's total population of around 276 million people. Where the majority of internet use is aimed at activities on social media. This is shown by the number of social media users reaching 167 million users in 2023 or around 78.4% of the total internet users in Indonesia [1]. However, with technological advances, it turns out that there are still many people who have not been able to utilize these technological advances wisely, one of which is the spread of hate speech. The spread of hate speech on social media has become an alarming phenomenon. Various social media platforms provide very easy access for individuals to spread hoax and make offensive comment that can cause division, discrimination, harsh words, and also give negative impacts to other people. Individuals who are targets of hate speech are also vulnerable to experiencing psychological disorders, such as stress, anxiety and depression, which can damage their mental health [2].

In dealing with the large amount of hate speech on social media, the government's efforts to prevent and overcome this problem are proven by the creation of legislation in the form of the ITE Law. In the ITE Law article 28 paragraph 2 it is stated that netizens are prohibited from spreading information to cause feelings of hatred [3]. However, detecting whether someone's comments or tweets contain hate speech or not is difficult, especially when detecting it manually. When identifying hate speech manually, it will require a lot of effort and take a lot of time, because there are definitely more than thousands of comments or tweets from social media users.

Based on this problem, a system is needed that can automatically detect comments that contain hate speech. One of the easiest ways to reduce hate speech is to block every word that is considered part of hate speech. However, not all comments or tweets that use words that are considered hate speech can be classified as hate speech, sometimes some tweets may only cause hatred for some people, this happens because people's perceptions can be different. For example, the use of animal words such as dog, pig, monkey, which in the context could be about pets but could be misinterpreted as an expression of hatred. Therefore, machine learning is very necessary to help track hate speech on social media platforms, and to prevent hate speech from appearing on social media [4].

The reason for using machine learning is because machine learning can be used to process large amounts of data and then make predictions and classifications on new data that has never been seen before. By using machine learning, it can predict whether someone's comments contain hate speech or not and classify them based on the label. There are various machine learning (ML) algorithms that can be used to detect hate speech such as Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Decision Trees, K-Means, Random Forest, and so on [5]. These algorithms are useful for creating relationship models between given input data sets and will provide the expected output. By testing using various types of machine learning algorithms, it will be useful to find out which algorithm is more accurate in detecting hate speech.

Research related to hate speech that occurs through social media has been carried out before and is indeed an interesting thing to discuss. Previous research discusses the use of traditional machine learning algorithms, which are used to classify tweets whether they fall into the category of hate speech or non-hate speech, this research was conducted by Akib Mohi Ud Din Khanday, et al., [6]. This research discusses Logistic Regression, Multinomial Naïve Bayes, Support Vector Machine, and Decision Tree, with Logistic Regression using feature engineering, namely TF/IDF and Bag of Words. This research also mentions the effectiveness of the Ensemble Learning classifier. Stochastic Gradient Boosting showing the highest accuracy among other models, with 98.04% accuracy. Apart from that, this research also describes the process of classifying tweets into hate speech and non-hate speech categories, such as using tokenization, stop word removal, stemming, and feature extraction techniques.

Another research conducted by Nabila Putri Damayanti, et al., [7] also discusses the classification of hate and non-hate speech on Twitter using a combination of Logistic Regression (LR) and Support Vector Machine (SVM) algorithms. This research highlights the importance of addressing the negative impacts of technology, particularly hate, and emphasizes the need to increase accuracy in classifying such problems. This study compares its approach to previous research, and notes the potential for increased accuracy. The combination of LR and SVM is proven to produce a classification model with high accuracy of 96%. This could provide a promising solution to the challenge of identifying hate speech on social media platforms. This research also provides suggestions for further research, it is hoped that future research will use a more varied dataset and further explore text processing techniques in order to

improve the representation of text features. It is hoped that future research will use a more varied dataset and further explore text processing techniques in order to improve the representation of text features.

There is also research discussed by Vijay, et al., [8] regarding the comparison of machine learning algorithms, namely Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) in classifying text as hate speech or non-hate speech. This highlights that SVM is a machine learning algorithm known for its high performance in classifying text, SVM uses hyperplanes to classify data points clearly. KNN is described as a statistical classification algorithm that determines the class of a text, this depends on the K value. In this research it was found that the SVM algorithm produces higher accuracy than the KNN algorithm. SVM with linear kernel function provides an accuracy level of 65%, SVM with radial basis function provides an accuracy level of 63.8%, while KNN with K=1 provides 60.8% accuracy, with K=2 provides 62.65% accuracy and with K=3 provides 62.75% accuracy.

This research also uses machine learning algorithms, namely Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) to detect hate speech, and also uses feature extraction, namely TF-IDF. This research will discuss the comparison between the two methods, to determine which method has a higher level of accuracy in detecting hate speech. Datasets are very necessary in this research, so that the system has input and can provide output in the form of classifying text into two labels, namely hate speech or non-hate speech. The dataset used in this research comes from Twitter, which consists of 31,962 user data tweets. However, there are things that can make the model less effective, usually because the dataset used is not balanced. Therefore, in this research we will also discuss techniques for balancing the dataset, so that the model can work more accurately and effectively. The data balancing technique used in this research is the random over sampling technique [9].

2. Methodology

This section will explain the design of the tweet classification model into two labels, namely the hate speech label and the non-hate speech label, as well as explaining the stages of data collection and the methodology used. Fig. 1. shows the complete research methodology process, first starting from data collection, then preprocessing, splitting the data, carry out data re-sampling technique, using feature extraction, building a machine learning model, and evaluation [4].

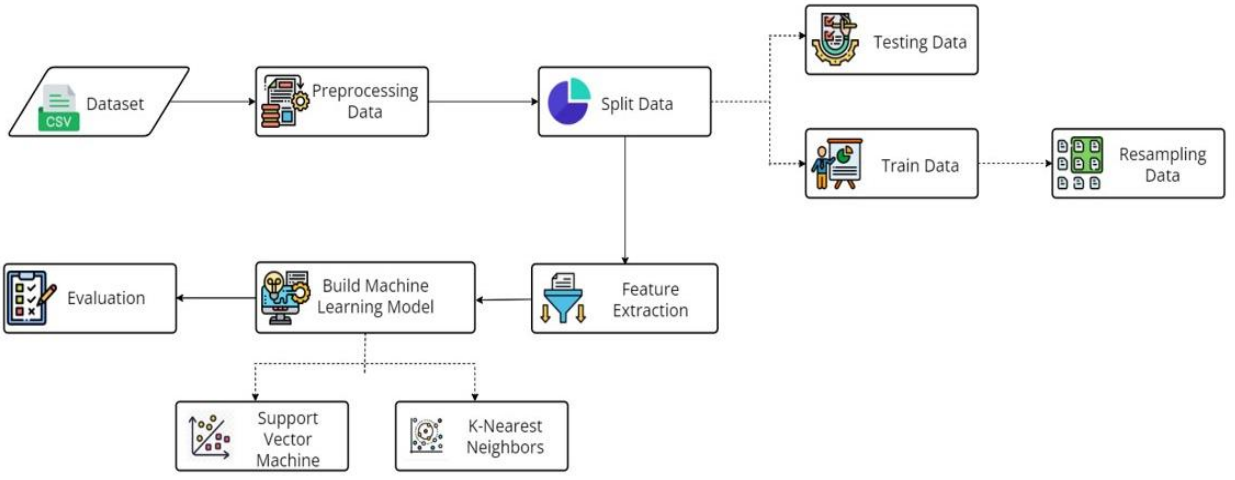

Fig. 1. Proposed Method Stage

2.1. Data collection

The dataset used in this research comes from Twitter, the dataset was uploaded by Subhajeet Das on Kaggle. This dataset consists of 31,962 tweets, and there are two labels, namely hate speech and non-hate speech [10]. The hate speech label contains tweets or comments that are racist, rude, harass someone, contain hatred, which can even cause division. Meanwhile, the non-hate speech label is a tweet or comment that is normal, does not corner anyone, and does not contain harsh words. To represent the dataset used, it can be seen in Table 1 [11].

Table 1. Representation of Hate Speech Data

Tweets	Hate Speech (*1*)	Non-Hate Speech (*0*)
@user yes lets do this,suppoing a openly,#prowar #anti #islamic,#homophobic,#rapist,who advocates more of same,#hypocrite	1	
use the power of your mind to #heal your body!!- #alwaystoheal #healthy #peace!		0

In this dataset there are 29,720 tweets that are classified as hate speech, and 2,242 tweets are classified as non-hate speech. It can be concluded that the dataset used in this research has an unequal or unbalanced distribution. Details of this distribution are shown in Fig. 2.

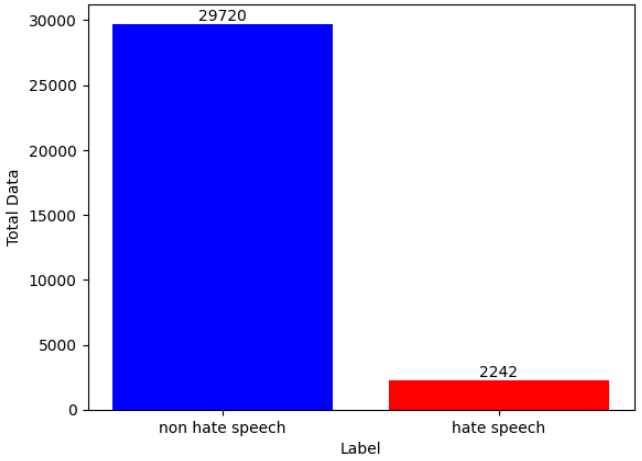

Fig. 2. Distribution of the Dataset for Each Label

2.2. Preprocessing

Preprocessing techniques are needed in building machine learning models, so that the classification process is more efficient and can provide the best results [12]. Preprocessing is a stage for cleaning data whose format is inconsistent, for example words or characters that are not needed in the classification process can be removed using preprocessing techniques [11]. In this research, there are five preprocessing techniques used, namely case folding, remove unnecessary characters, remove non-alphanumeric, stop word removal, and stemming. The steps can be seen in Fig. 3.

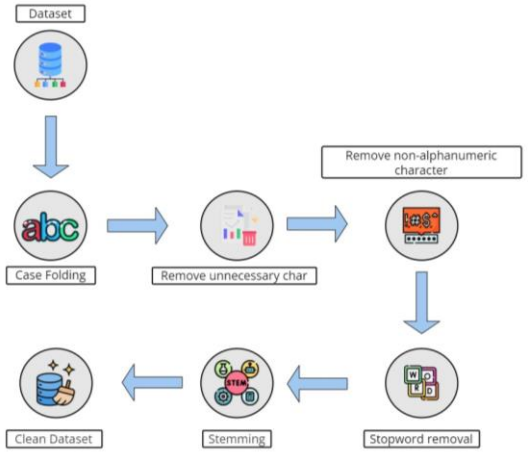

Fig. 3. Preprocessing Stage

- Case folding is useful for changing all letters to lowercase.
- Remove unnecessary characters can remove characters that are not useful in the classification process, for example URLs, retweets (rt), and user mentions.
- Remove non-alphanumeric is useful for removing characters other than numbers and letters.
- Stop word removal is useful for removing words that are less meaningful.
- Stemming is the process of removing words with affixes, so that each word only contains the basic word.

2.3. Split data

Machine learning is basically the process of how computers learn from data, without data the computer will not be able to learn anything. In this research, the data is divided into two parts, 80% is training data and 20% is test data. Training data is part of the dataset that is used to train a classification model so that the model can find data patterns, which will make it easier for the model to predict new data. Meanwhile, test data is part of the dataset used to test the accuracy of the model in predicting data [4].

2.4. Re-sampling technique

As is known, the dataset used in this research has data imbalance. When a dataset has data imbalance, and the dataset is split into train data and test data, this can cause the representation of minority data to be very low. Meanwhile, most algorithms can work well when the input data is balanced. So to avoid negative impacts that might occur due to imbalance in the dataset, data re-sampling techniques are needed [13]. The re-sampling technique used in this research is random over sampling, this technique is useful for adding data samples from existing minority data, so that the amount of minority data is equal to the majority data.

In this research, the random over sampling technique was applied in the training data section or during the model training process, this is useful for improving model performance because the model can learn patterns from minority data very well [14]. Fig. 4. (a) is the amount of training data before the re-sampling technique is carried out, and Fig. 4. (b) is the amount of training data after the re-sampling technique is carried out.

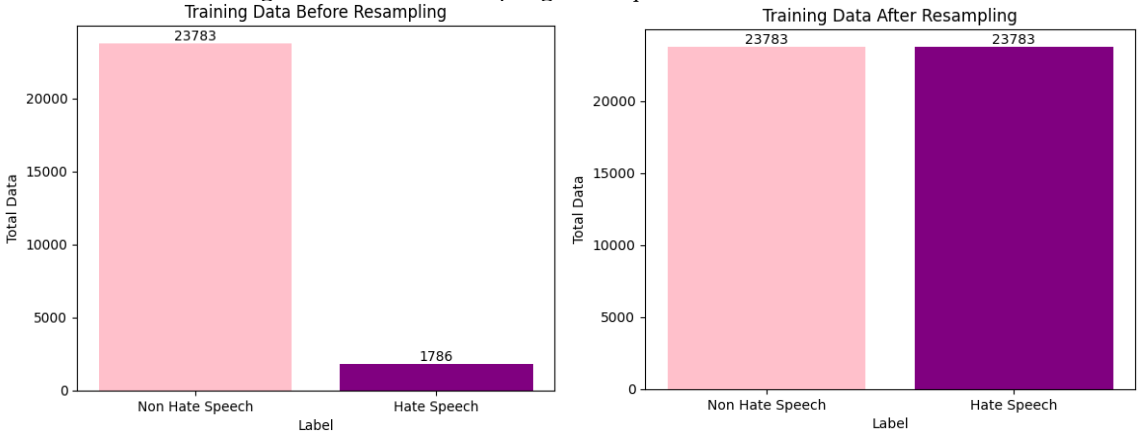

Fig. 4. (a) Training Data Before Re-sampling; (b) Training Data After Re-sampling.

2.5. Feature extraction

Feature extraction is a process for identifying and extracting important features from raw data, because raw data may contain many features that are not very relevant in the classification process or are redundant. Feature extraction can convert data that was previously text into a numeric format. So this feature extraction can help train machine learning models to process data more accurately [15]. In this research, the feature extraction used is Term Frequency - Inverse Document Frequency (TF-IDF), because this feature can help to produce a unique score for each word in the

document. Term Frequency (TF) refers to how often a word appears in a document. Meanwhile, Inverse Document Frequency (IDF) shows how rare a word is in the entire document [16].

2.6. Machine learning model

The aim of this research is to build a model that can classify tweets into two labels, namely the hate speech label and the non-hate speech label. To be able to classify tweets, a machine learning algorithm is needed. Machine learning can learn from a training data set and evaluate model performance using testing data [17]. In this research, the machine learning algorithms used are Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The two algorithms will be compared for their level of accuracy in detecting hate speech [6].

2.6.1. Support vector machine

The use of the Support Vector Machine method is usually used for classification or regression using linear or non-linear methods. The Support Vector Machine method is very suitable for diagnosing medical matters, detecting objects in images, and categorizing text [18]. This method can find the best hyperplane in N-dimensional space. Hyperplane is useful as a separator between one class and another. Therefore, the use of this method is suitable for detecting hate speech, because it can divide two different classes, namely tweets that contain hate speech and tweets that do not contain hate speech [19]. An illustration of the Support Vector Machine method can be seen in Fig. 5.

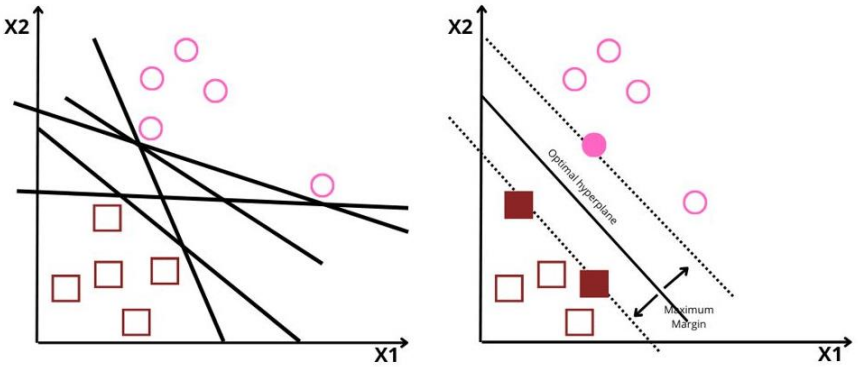

Fig. 5. Support Vector Machine

2.6.2. K - nearest neighbor

The K-Nearest Neighbor method is also a supervised learning method, where the results of the latest cases will be classified based on the majority of the k closest categories. The k value in the K-Nearest Neighbor method is the number of neighbors that will be examined to determine the classification of new data points. The choice of k value will really depend on the input data, because if the data has a lot of outliers or noise, then this method will work better with a higher k value. So this method can be useful for classifying new data that has not been labeled, then the data will be put into the closest majority class [20]. For an illustration of the K-Nearest Neighbor method, can be seen in Fig. 6.

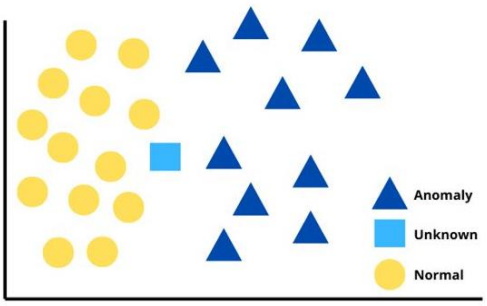

Fig. 6. K-Nearest Neighbor

2.7. Evaluation

This research uses a confusion matrix table to evaluate the performance of the classification algorithms. In the confusion matrix there are four terms that represent the results of the classification process. The four terms are True Negative, False Positive, False Negative, and True Positive. True Negative is a non-hate speech tweet that is correctly predicted as non-hate speech, True Positive is a hate speech tweet that is correctly predicted as hate speech, while False Negative is a tweet that is actually hate speech but incorrectly predicted as non-hate speech, and False Positive is a tweet that is actually non-hate speech but incorrectly predicted as hate speech [7]. The form of the confusion matrix table can be seen in Fig. 7.

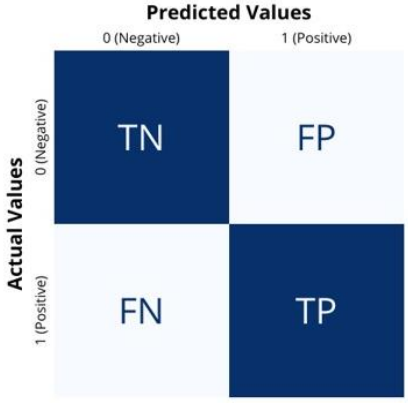

Fig. 7. Confusion Matrix

The confusion matrix can be used to calculate various performance metrics, or to measure the performance of algorithms in detecting hate speech. Some performance metrics that are commonly used are accuracy, accuracy describes how accurate the model is in carrying out classification. Then there is precision, which is the ratio of true positive predictions compared to the overall positive predicted results. Next there is recall, recall is the ratio of true positive predictions compared to the total true positive data. Then the last one is F1-score which is a calculation of the average between precision and recall values.

3. Result and Discussion

This experiment uses the python programming language along with the python library, and is carried out on the Google Colab platform, the reason is because Google Colab can use the python language and is easy to use. [9].

Table 2 below is the experimental results of the Support Vector Machine and K-Nearest Neighbor algorithms before the dataset balancing technique was carried out. In this research, the K-Nearest Neighbor algorithm uses the value k = 5.

Table 2. Experiment Results Before Balancing the Dataset

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	95.6%	93.9%	40.8%	56.9%
K-Nearest Neighbor	94.1%	97.5%	17.3%	29.4%

From the experimental results it can be seen that the Support Vector Machine algorithm produces a higher level of accuracy than the K-Nearest Neighbor algorithm, the accuracy is 95.6%. Then, to predict the success of the model in detecting hate speech in test data, a comparison is used between the predicted labels and the actual labels, or what is usually called a confusion matrix. This confusion matrix can also be used to obtain accuracy, precision, recall and F1-Score values. Fig. 8. (a) is the confusion matrix for the Support Vector Machine algorithm before the re-sampling

technique is carried out, and Fig. 8. (b) is the confusion matrix for the K-Nearest Neighbor algorithm before the re-sampling technique is carried out.

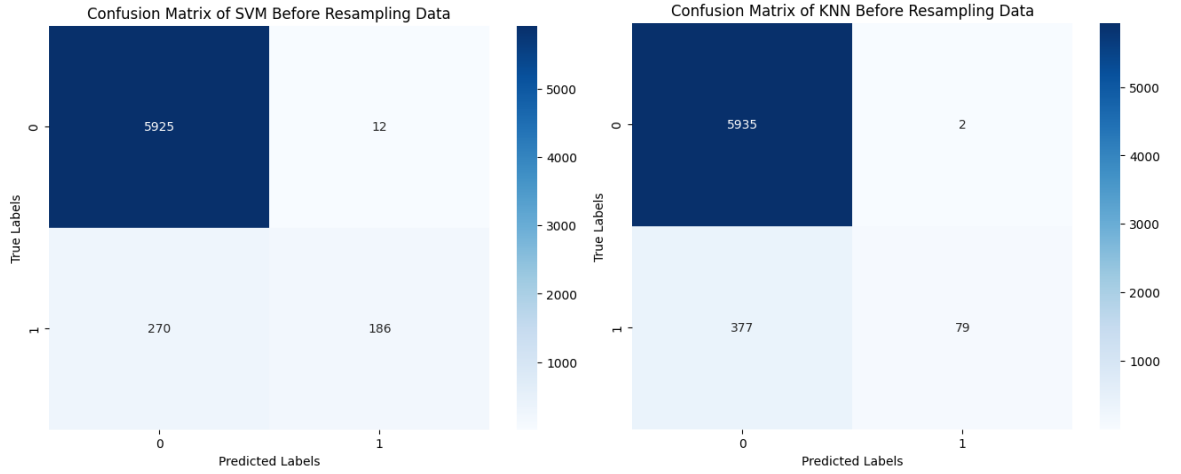

Fig. 8. (a) Confusion Matrix of SVM Before Re-sampling; (b) Confusion Matrix of KNN Before Re-sampling.

Next, Table 3 shows the experimental results of the Support Vector Machine and K-Nearest Neighbor algorithms after the dataset balancing technique was carried out. Then it is still the same for the K-Nearest Neighbor algorithm using the value k = 5.

Table 3. Experiment Results After Balancing the Dataset

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	96.1%	97.5%	96.1%	96.6%
K-Nearest Neighbor	95.2%	94.9%	95.2%	94.4%

It can be seen that by carrying out data re-sampling techniques, it can improve the performance of the two algorithms. This is shown by an increase in accuracy, precision, recall and F1-Score of each algorithm. The Support Vector Machine algorithm has an accuracy of 96.1% from the previous 95.6%, while the K-Nearest Neighbor algorithm has an accuracy of 95.2% from the previous 94.1%. After carrying out the data re-sampling technique, the Support Vector Machine algorithm still looks more effective than the K-Nearest Neighbor algorithm. Fig. 9. (a) and Fig. 9. (b) is the confusion matrix of the two algorithms after the dataset balancing technique is carried out.

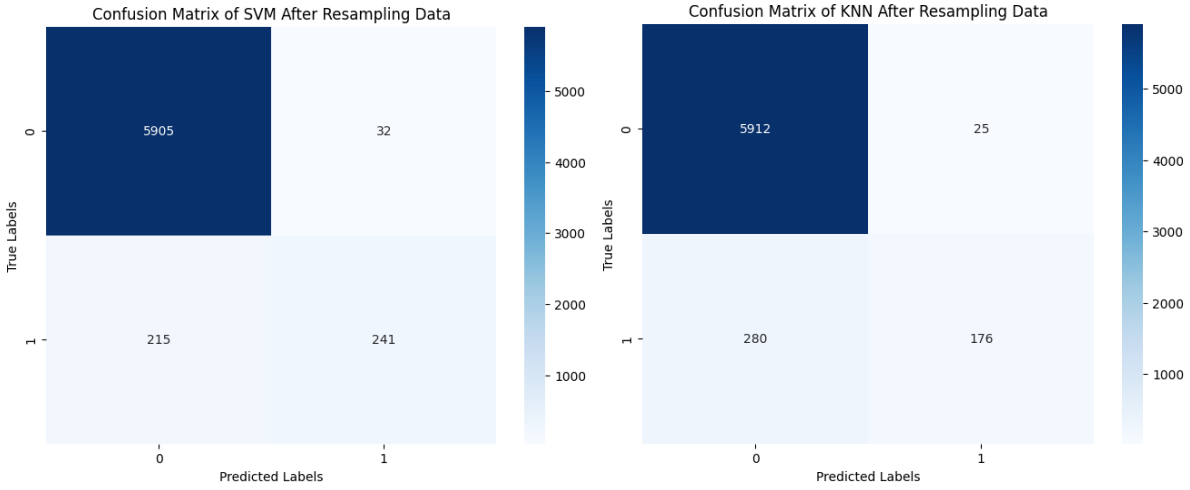

Fig. 9. (a) Confusion Matrix of SVM After Resampling; (b) Confusion Matrix of KNN After Resampling.

4. Conclusion

Based on the experiments that have been carried out, it can be said that the two algorithms have succeeded in classifying tweets that contain hate speech and tweets that do not contain hate speech. This is proven by the high level of accuracy of the two algorithms. Then, to handle cases of data imbalance in this research, a re-sampling technique was used. Although not all cases of data imbalance can be handled with re-sampling technique, in this case it turns out that the use of re-sampling technique can handle the problem of data imbalance and provide much better model performance results. Especially in the recall and F1-Score sections, the scores increased quite significantly. This is because a resampling technique was used in the training data, the re-sampling technique used is random over sampling. When the data is balanced during the training process, then the model can learn patterns from minority data very well. So when test data is given, the model will find it easier to classify the tweet whether it is labeled hate speech or non-hate speech.

In this research, the Support Vector Machine algorithm produces better accuracy than the K-Nearest Neighbor algorithm, this is because the Support Vector Machine algorithm has the advantage of finding hyperplanes that can directly separate two different types of data. Meanwhile, the K-Nearest Neighbor algorithm only relies on the k value or the majority value of its closest neighbors. For further research, it is recommended to experiment with a combination of the SVM and KNN algorithms and then compare them with the ensemble learning technique to find out which one has higher accuracy.

References

[1] wearesocial.com. DIGITAL 2024. *wearesocial*, https://wearesocial.com/id/blog/2023/01/digital-2023/ (2024, accessed 31 March 2024).

[2] Putri TTA, Sriadhi S, Sari RD, et al. A comparison of classification algorithms for hate speech detection. In: *IOP Conference Series: Materials Science and Engineering*. Institute of Physics Publishing, 2020. Epub ahead of print 18 May 2020. DOI: 10.1088/1757-899X/830/3/032006.

[3] Novaldi. Menkominfo: Persekusi di Dunia Maya Melanggar UU ITE. *kominfo.go.id*, https://www.kominfo.go.id/content/detail/9806/menkominfo-persekusi-di-dunia-maya-melanggar-uu-ite/0/sorotan_media (2017, accessed 1 April 2024).

[4] Abro S, Shaikh S, Ali Z, et al. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications* 2020; 11: 484–491.

[5] Hana KM, Adiwijaya, Al Faraby S, et al. Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines. In: *2020 International Conference on Data Science and Its Applications, ICoDSA 2020*. Institute of Electrical and Electronics Engineers Inc., 2020. Epub ahead of print 1 August 2020. DOI: 10.1109/ICoDSA50139.2020.9212992.

[6] Khanday AMUD, Rabani ST, Khan QR, et al. Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. *International Journal of Information Management Data Insights*; 2. Epub ahead of print 1 November 2022. DOI: 10.1016/j.jjimei.2022.100120.

[7] Damayanti NP, Prameswari DE, Puspita W, et al. Classification of Hate Comments on Twitter Using a Combination of Logistic Regression and Support Vector Machine Algorithm. *Journal of Information System Exploration and Research*; 2. Epub ahead of print 29 January 2024. DOI: 10.52465/joiser.v2i1.229.

[8] Verma P. Hate Speech Detection Using KNN and SVM. *International Journal of Scientific Research and Engineering Development*; 4, www.ijsred.com.

[9] Heri Cahyana N, Saifullah S, Fauziah Y, et al. *Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency*, www.ijacsa.thesai.org.

[10] SUBHAJEET DAS. Twitter Hate-Speech Detection (Different Model). *https://www.kaggle.com/code/subhajeetdas/twitter-hate-speech-detection-different-model/notebook.*

[11] Nabiilah GZ, Prasetyo SY, Izdihar ZN, et al. BERT base model for toxic comment analysis on Indonesian social media. In: *Procedia Computer Science*. Elsevier B.V., 2022, pp. 714–721.

[12] Ibrohim MO, Budi I. *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter*,

https://www.komnasham.go.id/index.php/ (2019).
[13] Ibrohim MO, Budi I. Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon*; 9. Epub ahead of print 1 August 2023. DOI: 10.1016/j.heliyon.2023.e18647.
[14] Zhu L, Zhou X, Zhang C. Rapid identification of high-quality marine shale gas reservoirs based on the oversampling method and random forest algorithm. *Artificial Intelligence in Geosciences* 2021; 2: 76–81.
[15] domino.ai. Feature Extraction. *Domino*, https://domino.ai/data-science-dictionary/feature-extraction (2024, accessed 5 April 2024).
[16] Khan A, Yousaf J, Muhammad T, et al. *HATE SPEECH DETECTION USING MACHINE LEARNING AND N-GRAM TECHNIQUES*.
[17] Seble H, Terefe T, Mekashaw F, et al. *HATE SPEECH DETECTION USING MACHINE LEARNING: A SURVEY*, www.academyjsekad.edu.ng.
[18] Asogwa DC, Chukwuneke CI, Ngene CC, et al. Hate Speech Classification Using SVM and Naive BAYES. Epub ahead of print 21 March 2022. DOI: 10.9790/0050-09012734.
[19] Rodríguez-Pérez R, Bajorath J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *J Comput Aided Mol Des* 2022; 36: 355–362.
[20] Lubis AR, Lubis M, Al-Khowarizmi. Optimization of distance formula in k-nearest neighbor method. *Bulletin of Electrical Engineering and Informatics* 2020; 9: 326–338.