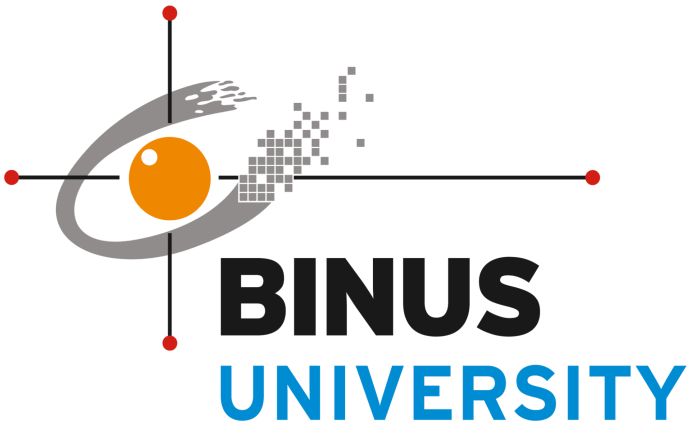
PROJECT REPORT
NATURAL LANGUAGE PROCESSING

Indonesian Hate Tweets Sentiment Analysis



2602157195 - Maesa Ken Neobi Arief
2602153234 - Edrick Givian Hulbert
2602159471 - Stephen Jo
2602177606 - Matthew William Siregar
2602199941 - Ganesha Warendra Sindhunata
2602151405 - Vincent Phillipus Li

BINUS University
2024

TABLE OF CONTENTS

ABSTRACT

Social media can be a place where hate speech spreads to an individual or group. This can trigger social conflicts that can damage national unity. This study analyzes tweets from Indonesia that may contain hate speech, utilizing the "Indonesian Abusive and Hate Speech Twitter Text" dataset obtained from Kaggle. The dataset is preprocessed and then balanced and split for formatting. Each data label is visualized to determine the method to be used in modeling. After modeling, fitting is performed to the SVM classification algorithm and complement Naive Bayes Classifier. Classification was conducted using the confusion matrix method and a classification report. The evaluation results show good performance, although there are some problems with unknown factors. Cultural and trend factors may affect the model's performance, leading to inconsistent results. Adjustments to the model are needed to make the results more accurate and consistent.

CHAPTER I
INTRODUCTION

In the digital age, social media has been one of the main platforms for interacting, information sharing, and discussing. However, with the platform being a method of freedom of expression, social media has also been a place that sources hate speech and abusive speech towards a certain individual or group. In Indonesia, this phenomenon has been increasingly devastating as the effects of them cause a significant amount of conflict.

In that context, it is why this study is prepared. To find a fitting algorithm and model to create a sentiment analysis of the tweets inside twitter. This study is aimed to compare the effectiveness of two models, which are Support Vector Machines (SVM) and Naive Bayes. Both models are commonly used in sentiment analysis and classification in general. The embedding used in this study is TF-IDF and Word2Vec. By this method, the aim is to give a contribution to prevent hate speech in the digital world.

CHAPTER II
RELATED WORKS

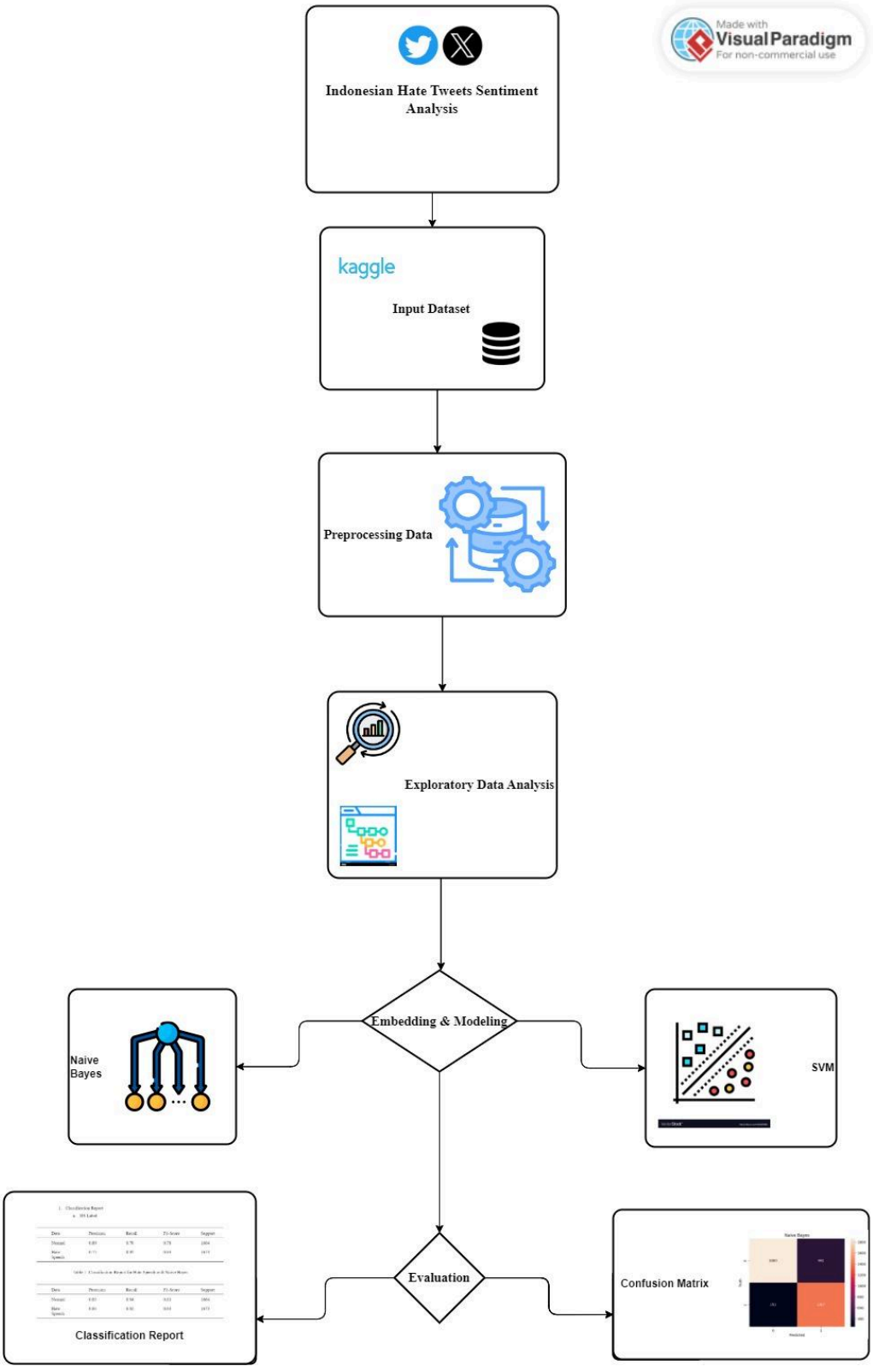
In 2022, Rahmat Syahputra, Gomal Juni Yaris, and Deci Irmayani conducted a study on sentiment analysis of Twitters users towards the PeduliLindungi application, which is used by the Indonesian government to monitor the spread of Covid-19. This study aimed to compare two classification algorithms in the supervised learning category, Support Vector Machines (SVM) and Naive Bayes, in analyzing Twitter user reviews on the PeduliLindungi application, including collecting data from user tweets using a crawling technique, text preprocessing, word weighting using TF-IDF method, training model using SVM and Naives Bayes, k-fold cross-validation test, and providing conclusions. The results showed that the accuracy of SVM with the k-fold test method was 86%, and the split 80-20 technique resulted in an accuracy of 89%. Meanwhile, the Naive Bayes algorithm produced an accuracy of 85% with k-fold, and 80% accuracy with a split of 80-20. From these results, it can be concluded that both algorithms have similar performance levels, but differ in processing time, with Naive Bayes algorithm is better, requiring only 0.0094 seconds.

In 2022, Laurenzius Julio Anreaja, Norma Nobuala Harefa, Julius Galih Prima Negara, Venantius Nathan Hermanu Pribyantara, and Agung Budi Prasetyo conducted a study on sentiment analysis of user reviews for the Opensea application on the Indonesian Play Store. The research aimed to understand public perception of the Opensea application using SVM and Naives Bayes. Both methods were assigned to compare public responses, with review data labeled as positive, negative, and neutral. The findings revealed that the Naive Bayes achieved a class precision of 87.31%, class recall of 71.02%, and an accuracy of 89.81%. In Contrast, the SVM algorithm demonstrated superior performance with a class precision of 94.23%, class recall of 71.96%, and an accuracy of 90.78%. In conclusion, the SVM outperformed the Naive Bayes algorithm in the sentiment analysis of Opensea user reviews.

In 2023, Ahmad Zahri Ruhban Adam and Erwin Budi Setiawan conducted a research aimed at analyzing sentiments expressed on Twitter in Bahasa Indonesia. Recognizing Twitter as a relevant platform for public opinion in Indonesia, the researchers focused on tweets labeled as

positive, negative, and neutral, represented by 1, -1, and 0, respectively. The study assigned the Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) methods to classify tweet sentiments. The research processes included feature selection, feature expansion, preprocessing, and balancing data with SMOTE. The results showed that the CNN-GRU achieved the highest accuracy of 97.58%. The study concluded that sentiment analysis using the CNN and GRU methods on Twitter data can yield high performance, and that feature expansion testing significantly enhances the accuracy of deep learning models.

CHAPTER III
PROPOSE METHOD



1. **Input Dataset**

The first thing to do is enter the required dataset. The dataset is data.csv, abusive.csv which contains abusive words, and new_kamusalay.csv which contains shortened and abbreviated words or phrases so that the model can still read the data. Because the language in our dataset is Indonesian, we use a different encoding from English encoding. We use CP437 and latin-1 encoding to read Indonesian documents.

The dataset used in this research is Indonesian Abusive and Hate Speech Twitter Text, taken from Kaggle. This dataset is designed to support multi-label detection of hate speech and abusive language in Indonesian on the Twitter platform. The main dataset consists of 13 columns with tweets in Indonesian with the following information labels:

1. **HS:** Hate speech label.
2. **Abusive:** Abusive words label.
3. **HS_Individual:** Hate speech targeted at individuals.
4. **HS_Group**: Hate speech targeted at groups.
5. **HS_Religion:** Hate speech targeted at religion.
6. **HS_Race:** Hate speech targeted at race.
7. **HS_Physical:** Hate speech related to physical body.
8. **HS_Gender:** Hate speech related to gender.
9. **HS_Others:** Hate speech related to other things.
10. **HS_Weak:** Weak hate speech.
11. **HS_Moderate:** Moderate hate speech.
12. **HS_Strong:** Strong hate speech.

For each label, it has a value of 1 and 0. 1 means that the tweet includes that label and 0 means that the tweet does not include that label. And due to Twitter's policy, the dataset provided does not have a Tweet ID and all usernames and URLs in the dataset have been changed to USER and URL.

2. **Preprocessing Data**

After inputting the dataset, the preprocessing is performed on the text from the dataset to prepare the data before fitting it to the model for further processing. First of all, word tokenization is carried out to break sentences into words and then remove Indonesian stopwords if they are detected. It is done so that the data can be cleaner before fitting it to the model.After that, each label is balanced and split to tidy up the data format. Each label is separated by 0 and 1 and then combined into its own .csv file.

3. **Exploratory Data Analysis**

For the EDA, first, we look at the components and features contained in the dataset. Then each data label is visualized to determine the appropriate method for the modeling part. WordCloud is used to see the words that appeared most frequently in the tweet data.

4. **Embedding & Modeling**

In the modeling section, the data is split into 2, namely X and Y. Data X contains tweets data, and Y contains category labels contained in the dataset. Then, the train_test_split function is used to divide 70% of the training data and 30% of the testing data for X and Y. Then, the text is embedded using the TF-IDF vectorizer embedding system and uses n grams for the range of text embedding. Then in the vectorizer too, the stopwords are removed in the dataset to clean the data and improve model performance.

After that, we fitted the classification algorithm. We use the SVM (Support Vector Machine) classification algorithm and also the Complement Naive Bayes Classifier.

5. **Evaluation**

For the evaluation part, evaluations are carried out for all hate speech type and category labels to determine the model's performance on this dataset. Two methods are used, namely classification report and confusion matrix. The reason is because the classification report immediately displays all evaluation components such as accuracy, precision, recall, and also f1-score. Then the confusion matrix is used to provide a visualization of the performance of determining HS labels in the model we developed to compare the performance of the SVM and Naive Bayes classification algorithms.

CHAPTER IV
EXPLORATORY DATA ANALYSIS (EDA)

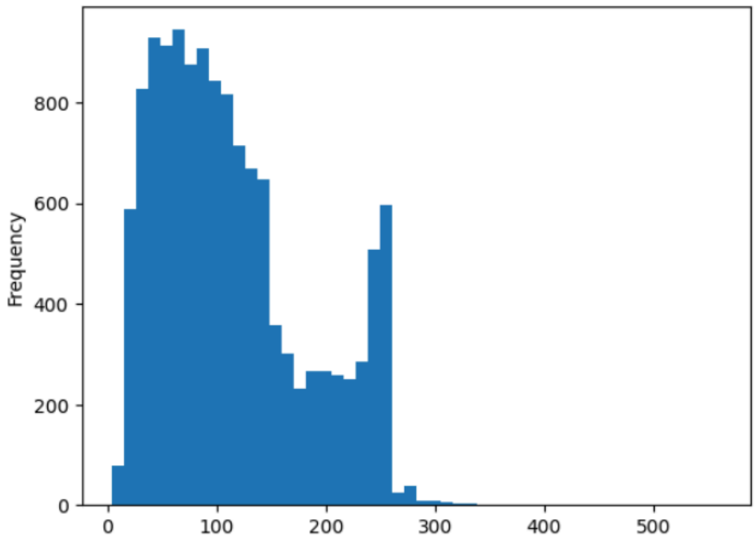


Fig 1. Histogram of Amount of Characters

Fig 1. Visualizes the number of characters through the following histogram. It can be seen that the most data distribution is between 50 - 100 characters in a tweet.

Fig 2. WordCloud of Tweets

Fig 2. is a WordCloud that visualizes the words that have the most frequencies in the tweets.

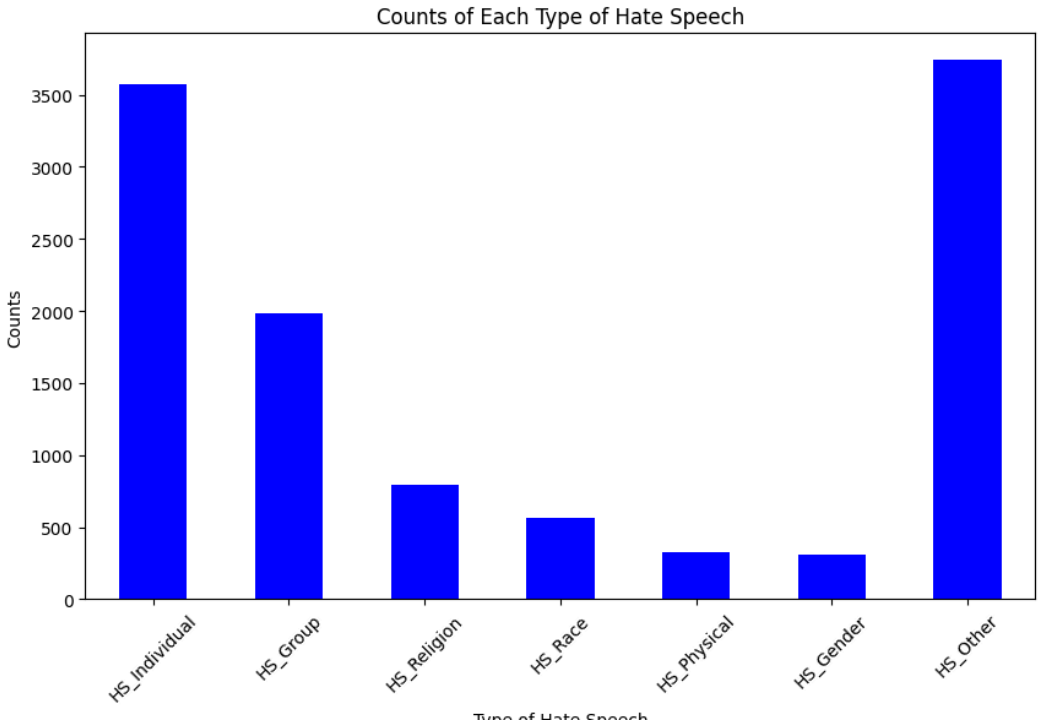


Fig 3. Counts of Types of Hate Speech

Fig 3. shows the distribution of target types of hate speech in tweets in the dataset is being used. There are individual, group, religious, racial, physical, gender, etc. targets. The type that has the most frequency is the HS_Individual label which shows that most of the hate speech tweets are subjected to individuals the most.

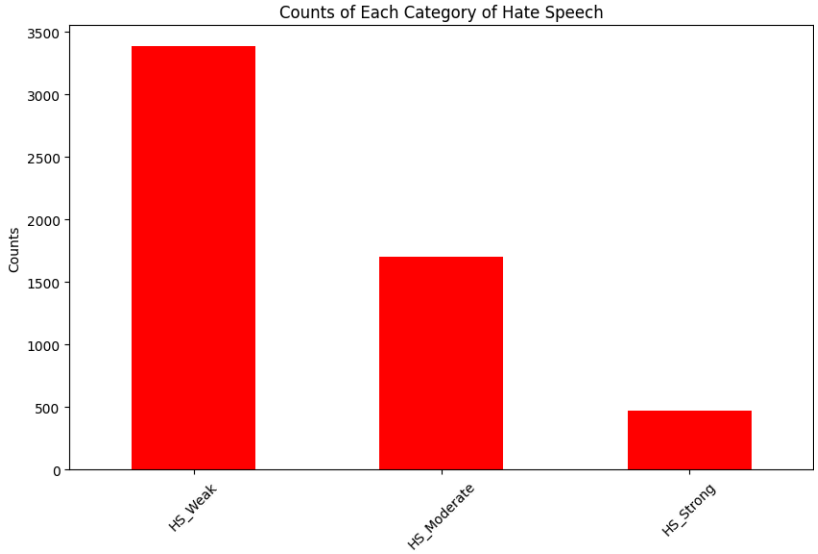


Fig 4. Bar Chart of Hate Speech Categories

Fig 4. visualizes the distribution of three hate speech categories in the dataset, from weak, moderate, and strong. From this bar chart, the weak category has the most frequency in the dataset.

CHAPTER V
PERFORMANCE ANALYSIS

To evaluate the performance of the model we developed, 2 metrics are used to evaluate the model:

1. Classification Report
    a. HS Label

Data	Precision	Recall	F1-Score	Support
Normal	0.89	0.70	0.78	1664
Hate Speech	0.75	0.92	0.83	1673

Table 1. Classification Report for Hate Speech with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.83	0.84	0.83	1664
Hate Speech	0.84	0.82	0.83	1673

Table 2. Classification Report for Hate Speech with Support Vector Machine

Table 1 shows the classification report for Hate Speech with Naive Bayes, which shows precision, recall, f1-score and support. Table 2 shows the classification report for Hate Speech with Support Vector Machine with the same metrics.

b. Abusive Label

Data	Precision	Recall	F1-Score	Support
Normal	0.91	0.79	0.84	1477
Abusive	0.82	0.92	0.87	1549

Table 3. Classification Report for Abusive with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.88	0.90	0.89	1477
Abusive	0.91	0.89	0.90	1549

Table 4. Classification Report for Abusive with Support Vector Machine

Table 3 shows the classification report for Abusive with Naive Bayes, which shows precision, recall, f1-score and support. Table 4 shows the classification report for Abusive with Support Vector Machine with the same metrics.

c. Label HS_Individual

Data	Precision	Recall	F1-Score	Support
Normal	0.86	0.67	0.75	1042
HS_Individual	0.74	0.89	0.81	1103

Table 5. Classification Report for HS_Individual with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.79	0.80	0.79	1042
HS_Individual	0.81	0.80	0.80	1103

Table 6. Classification Report for HS_Individual with Support Vector Machine

Table 5 shows the classification report for HS_Individual with Naive Bayes, which shows precision, recall, f1-score and support. Table 6 shows the classification report for HS_Individual with Support Vector Machine with the same metrics.

d. Label HS_Group

Data	Precision	Recall	F1-Score	Support
Normal	0.83	0.67	0.74	578
HS_Group	0.74	0.87	0.80	614

Table 7. Classification Report for HS_Group with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.74	0.82	0.78	578
HS_Group	0.81	0.73	0.77	614

Table 8. Classification Report for HS_Group with Support Vector Machine

Table 7 shows the classification report for HS_Group with Naive Bayes, which shows precision, recall, f1-score and support. Table 8 shows the classification report for HS_Group with Support Vector Machine with the same metrics.

e. Label HS_Religion

Data	Precision	Recall	F1-Score	Support
Normal	0.90	0.73	0.80	240
HS_Religion	0.77	0.92	0.84	236

Table 9. Classification Report for HS_Religion with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.85	0.92	0.88	240
HS_Religion	0.91	0.83	0.87	236

Table 10. Classification Report for HS_Religion with Support Vector Machine

Table 9 shows the classification report for HS_Religion with Naive Bayes, which shows precision, recall, f1-score and support. Table 10 shows the classification report for HS_Religion with Support Vector Machine with the same metrics.

f. Label HS_Race

Data	Precision	Recall	F1-Score	Support
Normal	0.91	0.75	0.82	158
HS_Race	0.81	0.93	0.87	182

Table 11. Classification Report for HS_Race with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.86	0.93	0.90	240
HS_Race	0.94	0.87	0.90	236

Table 12. Classification Report for HS_Race with Support Vector Machine

Table 11 shows the classification report for HS_Race with Naive Bayes, which shows precision, recall, f1-score and support. Table 12 shows the classification report for HS_Race with Support Vector Machine with the same metrics.

g. Label HS_Physical

Data	Precision	Recall	F1-Score	Support
Normal	0.85	0.60	0.71	88
HS_Physical	0.73	0.92	0.82	106

Table 13. Classification Report for HS_Physical with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.86	0.93	0.90	88
HS_Physical	0.94	0.87	0.90	106

Table 14. Classification Report for HS_Physical with Support Vector Machine

Table 13 shows the classification report for HS_Physical with Naive Bayes, which shows precision, recall, f1-score and support. Table 14 shows the classification report for HS_Physical with Support Vector Machine with the same metrics.

h. Label HS_Other

Data	Precision	Recall	F1-Score	Support
Normal	0.85	0.60	0.71	88
HS_Other	0.73	0.92	0.82	106

Table 15. Classification Report for HS_Other with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.86	0.93	0.90	240
HS_Other	0.94	0.87	0.90	236

Table 16. Classification Report for HS_Other with Support Vector Machine

Table 15 shows the classification report for HS_Other with Naive Bayes, which shows precision, recall, f1-score and support. Table 16 shows the classification report for HS_Other with Support Vector Machine with the same metrics.

i. Label HS_Weak

Data	Precision	Recall	F1-Score	Support
Normal	0.83	0.63	0.72	995
HS_Other	0.71	0.87	0.78	1035

Table 17. Classification Report for HS_Other with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.79	0.79	0.79	995
HS_Other	0.80	0.80	0.80	1035

Table 18. Classification Report for HS_Other with Support Vector Machine

Table 17 shows the classification report for HS_Weak with Naive Bayes, which shows precision, recall, f1-score and support. Table 18 shows the classification report for HS_Weak with Support Vector Machine with the same metrics.

j. Label HS_Moderate

Data	Precision	Recall	F1-Score	Support
Normal	0.82	0.63	0.71	502
HS_Other	0.71	0.86	0.78	521

Table 19. Classification Report for HS_Moderate with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.71	0.79	0.75	502
HS_Other	0.77	0.69	0.73	521

Table 20. Classification Report for HS_Moderate with Support Vector Machine

Table 19 shows the classification report for HS_Moderate with Naive Bayes, which shows precision, recall, f1-score and support. Table 20 shows the classification report for HS_Moderate with Support Vector Machine with the same metrics.

k. Label HS_Strong

Data	Precision	Recall	F1-Score	Support
Normal	0.93	0.76	0.84	144
HS_Other	0.79	0.94	0.86	140

Table 21. Classification Report for HS_Strong with Naive Bayes

Data	Precision	Recall	F1-Score	Support
Normal	0.87	0.96	0.91	144
HS_Other	0.95	0.85	0.90	140

Table 22. Classification Report for HS_Strong with Support Vector Machine

Table 21 shows the classification report for HS_Strong with Naive Bayes, which shows precision, recall, f1-score and support. Table 22 shows the classification report for HS_Strong with Support Vector Machine with the same metrics.

2. Confusion Matrix

a. Naive Bayes

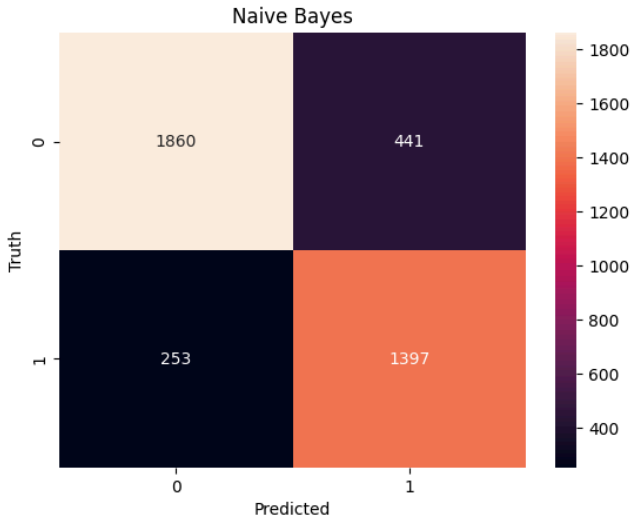


Fig 5. Confusion Matrix for Naive Bayes

Fig 5. visualizes the confusion matrix for the Naive Bayes, it can be seen that the number of true positives and true negatives has far exceeded the false positives and false negatives in the model, which indicates that the model's accuracy is sufficient for testing the dataset.

b. Support Vector Machine

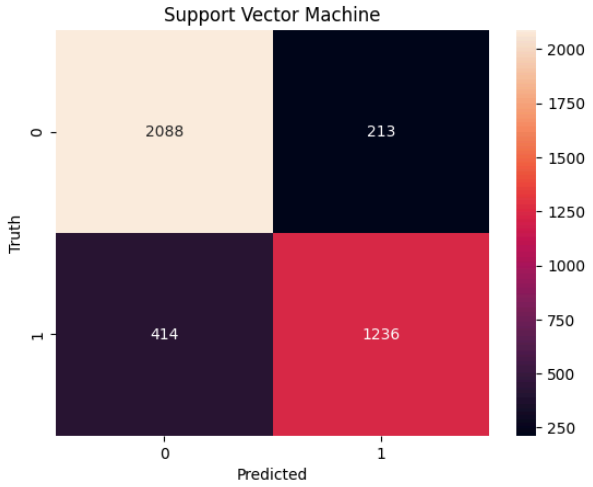


Fig 6. Confusion Matrix for Support Vector Machine

Fig 6. visualizes the confusion matrix for the Support Vector Machine, it can be seen that the number of true positives and true negatives has far exceeded the false positives and false negatives in the model, which indicates that the model's accuracy is sufficient for testing the dataset.

After carrying out an evaluation using evaluation metrics on Sklearn, the model developed from both algorithms was starting to achieve fairly good accuracy. The Naive Bayes algorithm seems to perform better in true negatives than true positives according to the confusion matrix. For the Support Vector Machine algorithm, it performs better in the true positives than the true negatives according to the confusion matrix. From the classification report, both models showed similar results as each label had different frequencies.

CHAPTER VI
CONCLUSION

The conclusion of the sentiment analysis of hate tweets in Indonesia shows that the model developed has achieved a satisfactory level of accuracy overall. The evaluation results using the confusion matrix also show good performance, although there are several factors that have not been identified that influence the model's ability to classify certain words. Both models showed that the classification is accurate, however it needs to be tested in a broader context.

However, it is important to recognize that in sentiment analysis, there are complexities in the language that are often difficult for models to fully understand. It is possible that contextual factors, cultural nuances, or even changing trends in language use may influence model performance. Therefore, further research is needed to understand these factors and improve model performance.

To improve model performance, a more holistic approach could be considered, including the use of more sophisticated NLP techniques, more representative data collection, and more appropriate adaptation of the model to the context of Indonesian language use. By doing this, it is hoped that the model can provide more consistent and reliable results in identifying hate sentiment in online content in Indonesia.

REFERENCES

Putra, Ilham. “Indonesian Abusive and Hate Speech Twitter Text.” Kaggle, 2020, https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text?select=citation.bib. Accessed 01 April 2024.

A. Fauzi, M. F. Akbar, dan Y. F. A. Asmawan, "Sentimen Analisis Berinternet Pada Media Sosial dengan Menggunakan Algoritma Bayes," J. Inform., vol. 6, no. 1, hal. 77-83, Apr. 2019.

R. Syahputra, G. J. . Yanris, and D. . Irmayani, “SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter”, *SinkrOn*, vol. 6, no. 2, pp. 671-678, May 2022. [Related Works]

Anreaja, L., Harefa, N., Negara, J., Pribyantara, V., & Prasetyo, A. (2022). Naive Bayes and Support Vector Machine Algorithm for Sentiment Analysis Opensea Mobile Application Users in Indonesia. *JISA(Jurnal Informatika dan Sains), 5*(1), 62-68. doi:https://doi.org/10.31326/jisa.v5i1.1267. [Related Works]

Adam, A., & Setiawan, E. (2023). Social Media Sentiment Analysis Using Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU). *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika, 9*(1), 119-131. doi:http://dx.doi.org/10.26555/jiteki.v9i1.25813. [Related Works]

APPENDIX

Source Code: Kelompok1_NLP.ipynb