Comparison of Naive Bayes, K-Nearest Neighbor, and Support Vector Machine Classification Methods in Semi-Supervised Learning for Sentiment Analysis of Kereta Cepat Jakarta Bandung (KCJB)

M Farhan[1], **R D L R Manik**[1], **H R Jannah**[1], **L H Suadaa**[1,*]
[1] Politeknik Statistika STIS, Jl. Otto Iskandardinata No. 64C, East Jakarta, DKI Jakarta

*Corresponding author's e-mail: lya@stis.ac.id

Abstract. Transportation technology has developed very rapidly in the 21st century; one of them is high-speed trains. Currently, the Indonesian government is implementing the construction of the Kereta Cepat Jakarta-Bandung (KCJB) project in collaboration with China. The construction of this fast train project has attracted various comments and opinions from the public on Twitter and social media. This research aims to compare the classification methods of Naïve Bayes, K-Nearest Neighbor (K-NN), and Support Vector Machine (SVM) in classifying sentiment in tweets about high-speed trains obtained by scraping Twitter. The comparison process was carried out using semi-supervised learning, and the results showed that the semi-supervised SVM model had the best performance with an average accuracy of 86%, followed by the semi-supervised Naïve Bayes model and semi-supervised K-NN with an average accuracy of 81% and 58% respectively. Overall, the prediction results from the three models conclude that there are more tweets with negative sentiment than tweets with positive and neutral sentiment.

1. Introduction

Technology is developing all the time, including transportation technology. Public transportation as a passenger transport service is available for the general public. The higher level of mobility raises the need for public transportation that can run more efficiently [1]. The Indonesian government is trying to improve transportation services and support development in the Jakarta-Bandung area by building a fast train. The project is a collaboration between Indonesia and China. A consortium called PT. Kereta Api Indonesia China (KCIC) was formed as the executor of the Kereta Cepat Jakarta-Bandung development and construction project [2].

The construction of the Jakarta-Bandung high-speed train sparked a lot of attention and varied opinions from various levels of society. Former Minister of BUMN, Rini Mariani Soemarno, said that the benefits of building the Jakarta-Bandung high-speed train include improving the economy and the tourism sector and opening up new job opportunities [3]. However, during its construction, there needed to be more attention to smooth access in and out of toll roads, allowing the accumulation of materials that disrupted drainage functions, building LRT pillars without permits, and occupational safety and

health (K3) issues [4]. Few also focus on economic figures regarding how much the country profits and losses [5].

Nowadays, people can express their opinions through various media. One is the social media Twitter, which is massively used because it is considered adequate for expressing their opinions and thoughts, including voicing their perspectives on the construction of the Jakarta-Bandung Fast Train. Indonesian people tweeted many comments regarding this project. Some comments are in acceptance, others are against, and some are neither. Things like this are usually called sentiments [6].

Sentiment analysis can be performed to determine public opinion on an event [7]. Sentiment analysis, also known as opinion mining, is the study of how a person's opinions, sentiments or feelings, behavior, and emotions about an entity are expressed through writing. The information obtained can be in the form of positive, negative, or neutral sentiments. Through a data mining approach, a classification model can be built to label a tweet, whether it is positive, negative, or neutral, automatically so that conclusions and information gathering can be made quickly [8].

Various methods can be used to carry out sentiment analysis, such as the Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) classification technique [9]. This method is a supervised classification method that requires the availability of labels from all data to carry out the training and testing process. In the case of limited labeled data, a semi-supervised approach can be taken to perform sentiment analysis [10][11]. Semi-supervised methods can also improve the performance of sentiment analysis models and achieve higher levels of accuracy [10][11]. Based on that, this research aims to compare the Naïve Bayes, K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM) classification methods in conducting sentiment analysis regarding the Kereta Cepat Jakarta Bandung (KCJB) with a semi-supervised learning approach.

2. Literature Review

2.1. Theoretical Basis

2.1.1. Preprocessing. The data preprocessing stage aims to obtain structured data from unstructured raw data. Several stages are passed in this process, including eliminating duplicate data, case folding, tokenization, normalization, removal of stopwords, and stemming.

- *Case folding.* Case folding is a step in data processing that aims to change all capital letters in documents to lowercase [12].
- *Tokenization.* Tokenization is the process of breaking sentences into words, punctuation marks, and other meaningful expressions according to the rules of the language [12]. The results of this process are called tokens [6].
- *Normalization*. Normalization is a preprocessing stage to change unstructured datasets into structured ones and simplify data processing [12].
- *Stopword.* A technique called Stopword Removal eliminates insignificant words [12]. The reason for removing stop words is that their use is too every day, so users can focus on other words that are much more important [13].
- *Stemming.* Stemming removes prefixes, suffixes, and confixes to change words with affixes into their basic form [12].

2.1.2. Sentiment Analysis. Sentiment analysis is a subfield within computational research to investigate a text's opinions, feelings, and emotions. Sentiment analysis is used to understand statements created by internet users and explain how a product is received by them [14]. The main goal is to process, extract, summarize, and analyze text information to obtain the author's emotions and point of view from the text [12].

2.1.3. Semi-Supervised Learning. Semi-supervised learning uses unlabeled data complemented by labeled data to form a classification model. Semi-supervised learning is commonly used for sentiment analysis with lots of unlabeled data. This model produces quite good results when the dataset is large [15]. According to Jamshid Bagherzadeh's research, there are several semi-supervised learning techniques: self-training, generative, co-training, and semi-supervised boosting. In self-training methods, there are pseudo-labeling methods that use a small amount of labeled data to label more extensive set of unlabeled data [16]. Pseudo-labeling techniques are simple and robust and do not need to focus too much on tuning parameters [17].

2.1.4. Feature Selection. Term Frequency-Inverse Document Frequency (TF-IDF) is a weighting matrix that measures the relationship between words by combining two parts, namely Term Frequency (TF) to count the number of occurrences of each word in a document and Inverse Document Frequency (IDF) to calculate the number of related documents which contains certain words [18]. The TF-IDF feature is one of the extraction techniques by assigning a value to each word in the data to determine how important the word that represents the sentence is. The calculation value is determined by the frequency of the word's appearance in the document [19].

2.1.5. Classification Method

- *Naïve Bayes.* Naive Bayes Classifier is one of the algorithms popularly used for data mining purposes because of its fast processing time, ease of implementation with its relatively simple structure, and high level of effectiveness [20]. Naïve Bayes is a classification method that requires calculating an opportunity for each attribute to build a Naïve Bayes model. This method requires instructions to determine the class for sample data [21].
- *K-NN.* K-nearest neighbor (K-NN) is a non-parametric classification method or has no assumptions regarding the underlying data distribution [22]. Non-parametric algorithms such as K-NN use a flexible number of parameters, and the number of parameters often increases as the data increases. Non-parametric algorithms are computationally slower but make fewer assumptions about the data [23]. This algorithm works by classifying data into several classes from several data as the nearest neighbors of that point [21].
- *SVM.* Support Vector Machine (SVM) is an algorithm using optimization theory and hypothetical space in a series of pattern recognition fields and using the kernel to map the input space [24]. SVM is often implemented for various problems and purposes such as pattern identification, bioinformatics, and text classification by decomposing hyperplanes as input features consisting of two classes, then re-optimized into more than two classes [25]. The unlimited function in searching for hyperplanes in the Support Vector Machine method is an advantage, where processing will always be possible regardless of the data used [6].

2.1.6. Resampling. Due to the uneven data distribution among categories, resampling is performed to address class imbalance. In this study, resampling was carried out using the Synthetic Minority Over-Sampling Techniques (SMOTE) method. The SMOTE method combines under-sampling and over-sampling, which are both the majority. SMOTE is performed by adding the difference in weight between the samples and their nearest neighbors from synthetic samples randomly generated in the over-sampling section. In contrast, the under-sampling section is only a general procedure. The existing synthetic sample aims to establish a decision boundary and reduce the effect of overfitting at the time of classification [26].

2.1.7. Evaluation Method. Cross-validation is a method or technique to evaluate an analysis's results. In general, this cross-validation technique is used to predict the level of accuracy of a predictive model, where the data is divided by considering the same amount of data into k parts. K-Fold Cross Validation

is the primary form of cross-validation. This method is used to estimate the error of the predictive model and evaluate the model assisted by data testing [27].

2.2. Related Research

Various studies on sentiment analysis have been carried out, especially using semi-supervised learning methods. Macrohon et al. [10] present a semi-supervised approach to sentiment analysis of tweets during the 2022 Philippine Presidential Election. The researchers collected tweets related to the two primary candidates and used Natural Language Processing techniques to classify the tweets into positive, neutral, and negative polarities. They achieved an accuracy of 84.83% using the Self-Training model with Multinomial Naïve Bayes as the base classifier, utilizing 30% unlabeled data. This study used a semi-supervised approach because it allows for utilizing both labeled and unlabeled data, which can be beneficial in scenarios where labeled data is limited or expensive to obtain. By incorporating unlabeled data, the researchers were able to improve the performance of their sentiment analysis model and achieve a higher accuracy rate.

Other research that uses a semi-supervised learning approach is research from Vivian et al. [11]. This research discusses using semi-supervised learning in sentiment classification and compares its performance to supervised learning. The authors trained a deep neural network using a combination of labeled and unlabeled data and found that the unlabeled data improved the model's performance. Semi-supervised learning differs from supervised learning in deep neural networks regarding the amount of labeled data used for training [11]. In supervised learning, the data is fully labeled, and the model is trained using only the labeled data.

On the other hand, in semi-supervised learning, the data is partially labeled, with a small part of labeled data and a large part of unlabeled data. The model is first trained with the labeled data, then the trained model is used to make predictions on the unlabeled data. The predicted outputs from the unlabeled data are then used along with the labeled data to train the model further. This combination of labeled and unlabeled data helps reduce the effort and time needed for data labeling [11]. The study found that incorporating unlabeled data in the training process of a semi-supervised deep neural network resulted in improved performance compared to a baseline model trained only with labeled data.

Besides that, research conducted by Haekal et al. [9] discusses sentiment analysis of Twitter data related to the 2024 presidential candidates in Indonesia. The study compares the performance of three classification algorithms, namely Support Vector Machine (SVM), Naïve Bayes, and K-Nearest Neighbors (K-NN), in analyzing the sentiment of tweets. The results show that the SVM algorithm provides superior performance with a total accuracy rate of 88% and demonstrated high consistency in precision and recall for all sentiment categories. Furthermore, SVM aims to find an optimal hyperplane that maximizes the margin between different classes in the data, allowing for better classification accuracy and generalization to new data. This characteristic of SVM makes it well-suited for sentiment analysis tasks where it aims to accurately classify tweets into positive, negative, or neutral sentiments [9].

3. Method

The complete steps of this research can be seen in Figure 1.

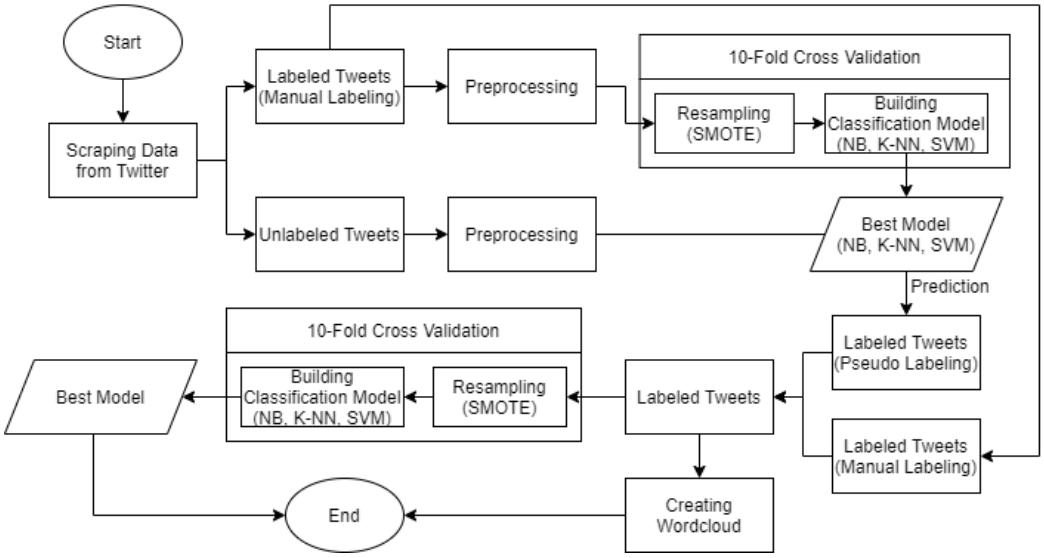

Figure 1. Flowchart of Research Steps

In this study, data was collected using web scraping techniques on Twitter pages using the tweet-harvest package and the keyword "kereta cepat." Web scraping is a technique for retrieving information in text form by automatically extracting information from websites [28]. Data was collected from November 6 to December 31, 2022. This timeline is due to the dynamic test of the Kereta Cepat Jakarta-Bandung, which took place on November 16, 2022 [29]. This event was expected to generate much public discussion and comments about fast trains on the social media platform Twitter.

After the tweet data has been collected, the next step is to label some of the tweets manually. Each tweet is labeled with a category of positive, negative, or neutral. Furthermore, both tweets that have been manually labeled and tweets that have not been labeled will be preprocessed. After the preprocessing step is complete, labeled tweets will be used to build Naïve Bayes, K-NN, and SVM classification models by first resampling the SMOTE method on the training data for each iteration of the 10-fold cross-validation process. The model with the best accuracy from each classification method will predict unlabeled tweets. This process is called pseudo labelling, where the predicted label is considered the actual label of the tweet.

Tweets resulting from manual labeling and pseudo-labeling will then be combined into labeled tweets. These labeled tweets will be applied to the same modeling process as in the previous stage, namely by building a semi-supervised Naïve Bayes, semi-supervised K-NN, and semi-supervised SVM model by first carrying out resampling using the SMOTE method on the training data for each iteration in the 10-fold cross-validation process, so that the best model is obtained. Besides that, labeled tweets will also be used to visualize the distribution of words for each class category using the word cloud.

4. Research and Result

The scraping process obtained around 10,598 tweets, which were then checked for duplication by deleting duplicate tweets from the same username and timestamp. This resulted in 10,271 tweets without duplication. Following the semi-supervised learning concept, from a total of 10,271 tweets, 5,104 tweets from November 6 to 30, 2022, were labeled manually. This labeling involves two people for each tweet. Tweets that will later be used in modeling are tweets with labels in the same category as the results of labeling the two people.

The categories used in labeling are positive, negative, and neutral. Tweets categorized as positive can be interpreted as tweets supporting the existence of the Kereta Cepat Jakarta-Bandung. Tweets

categorized as unfavorable can be interpreted as tweets that do not support or reject the existence of the fast train. Meanwhile, tweets that are categorized as neutral are tweets that are not related to the topic of the Jakarta-Bandung high-speed train or tweets that neither support nor reject the construction of the high-speed train in Indonesia.

Table 1. Example of Tweet Category Labeling

Tweet	Category
Kami sebagai rakyat bangga sama pak Jokowi,yg penting barangnya ada perihal mau balik modal 38 THN lagi tidak ada masalah daripada duitnya dikorupsi dan pekerjaannya mangkrak tidak selesai. Kalo pemikirannya cuma untung rugi sampai kapan kita punya kereta cepat?	Positive
Telah terjadi suatu keputusan bodoh yg mengakibatkan kerugian, sudah mengerti akan ada kerugian tapi proyek masih tetap dikerjakan, proyek kereta cepat akan balik modal setelah 38 th itupun jika proyek itu selesai, kenapa kebodohan sering terjadi pd pemeritahan jkw?	Negative
Progres terbaru Stasiun Halim KCIC yang menjadi terminus kereta cepat tersebut di DKI Jakarta. Stasiun Halim akan memiliki 6 jalur untuk headway 68 perjalanan per hari.	Neutral

The results of manual labeling from two people led to an unequal labeling category in some tweets. These tweets were subsequently deleted because they were considered to cause ambiguity and could reduce the model's performance in classifying. So, as many as 3775 tweets were produced, which were assumed to have better quality for use in modeling. Following are details of the number of tweets for each category after manual labeling by two people.

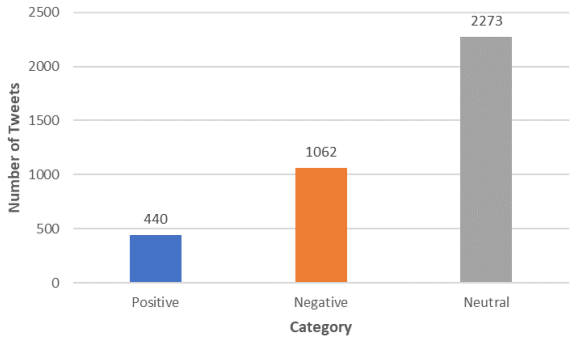

Figure 2. Manual Labeling Results by 2 People

After manual labeling, the next step is preprocessing. Preprocessing is done both on labeled tweets and unlabeled tweets. Preprocessing is a crucial step in sentiment analysis because it can help improve the quality of analysis results and minimize the impact of various noise or interference in the data. The following is an example of preprocessing results in this study.

Table 2. Preprocessing Steps and Results

Steps	Results
Before preprocessing	@ariarjiwinata02 @berlianidris Kereta dibanding sama Travel aman mana... ?? Kereta cepat adalah bentuk pilihan transportasi cepat dengan harga premium dan pantas.. Kalo anda naik Travel akan anda alami spot jantung.... https://t.co/BUWqFkuHbS'
Remove usernames, URLs, punctuation, emoji, hashtags, symbols, and numbers.	Kereta dibanding sama Travel aman mana Kereta cepat adalah bentuk pilihan transportasi cepat dengan harga premium dan pantas Kalo anda naik Travel akan anda alami spot jantung
Case folding	kereta dibanding sama travel aman mana kereta cepat adalah bentuk pilihan transportasi cepat dengan harga premium dan pantas kalo anda naik travel akan anda alami spot jantung
Tokenization	kereta dibanding sama travel aman mana kereta cepat adalah bentuk pilihan transportasi cepat dengan harga premium dan pantas kalo anda naik travel akan anda alami spot jantung
Stemming	kereta banding sama travel aman mana kereta cepat adalah bentuk pilih transportasi cepat dengan harga premium dan pantas kalo anda naik travel akan anda alami spot jantung
Remove stopword	kereta banding travel aman kereta cepat bentuk pilih transportasi cepat harga premium kalo travel alami spot jantung

Resampling is performed using the SMOTE method to address data imbalance before classification modeling. In each iteration of the cross-validation process, the training data is resampled. The resampled training data will be used to build and validate Naïve Bayes, K-NN, and SVM models through cross-validation. In this research, the process is referred to as Model 1. The number of folds used in Model 1 is 10-fold, and the accuracy results from each iteration for each model are as follows.

Table 3. 10-Fold Cross Validation Accuracy Results from Model 1

Iteration	Accuracy		
	Naïve Bayes	K-NN	SVM
1	0.7275	0.6455	0.7381
2	0.7196	0.6376	0.7725
3	0.7328	0.6534	0.8042
4	0.6693	0.6217	0.7249
5	0.7328	0.6138	0.7672
6	0.7188	0.6419	0.7931
7	0.7109	0.6340	0.7719
8	0.7347	0.6870	0.7905
9	0.6525	0.5968	0.7347
10	0.6790	0.6419	0.7798
Mean	0.71	0.64	0.77

Based on Table 3, it can be seen that the SVM model has the highest average accuracy, namely 0.77 or 77%. The Naïve Baye model followed them in second place with an accuracy of 71%, and the K-NN

model in third place with an accuracy of only 64%. Furthermore, the model with the highest accuracy among the ten iterations for each classification method will predict unlabeled tweets totaling around 5167 from 1 to December 31, 2022. The prediction results from each method provide details on the number of tweets for each category as follows.

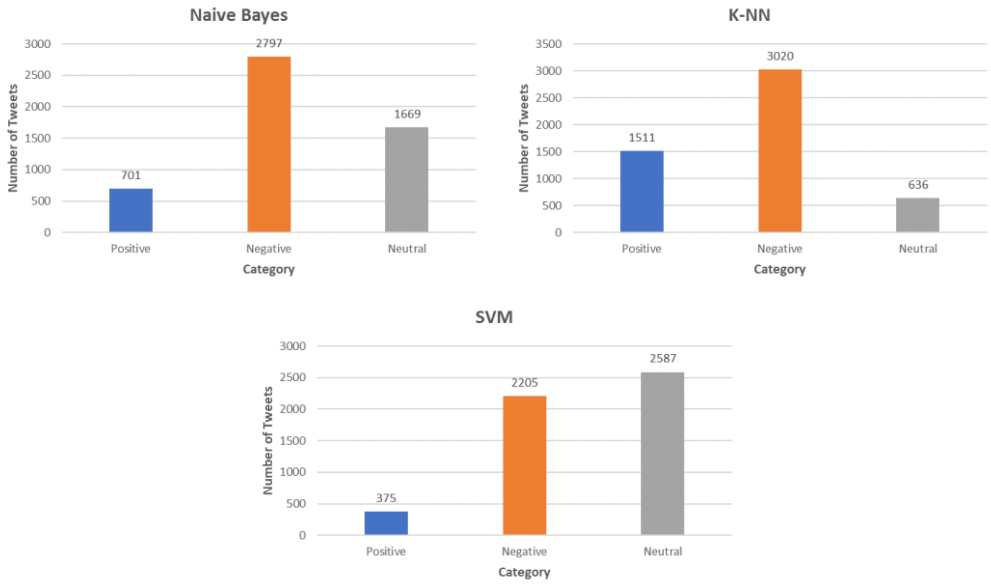

Figure 3. Prediction Results of Naive Bayes, K-NN, and SVM Classification Method

Based on Figure 2, overall, more tweets are categorized as negative than positive. This indicates that more people who use Twitter do not agree with the existence of the Kereta Cepat Jakarta-Bandung compared to people who agree. The prediction results from Model 1 are referred to as pseudo-labeled data. The labels resulting from these predictions are considered as the actual labels of the data.

After all, tweets have labels resulting from manual and pseudo-labeled labeling. The next step is to model all tweets with the same methods and techniques as in Model 1, namely by first balancing the number of tweets for each class category on the training data in each iteration on cross-validation using the SMOTE resampling method. Furthermore, from the training data that has been resampled, semi-supervised Naïve Bayes, semi-supervised K-NN, and semi-supervised SVM models will be built, which will then be validated in each iteration of the cross-validation process. This modeling is referred to as Model 2 in this study. The number of folds used in Model 2 is the same as in Model 1, which is 10-fold. The results of the accuracy of each iteration for each model in Model 2 are as follows.

Table 4. 10-Fold Cross Validation Accuracy Results from Model 2

Iteration	Accuracy		
	Naïve Bayes	K-NN	SVM
1	0.8022	0.5799	0.8704
2	0.8279	0.5899	0.8637
3	0.8087	0.5828	0.8803
4	0.8121	0.6085	0.8658
5	0.7964	0.5671	0.8535
6	0.8076	0.5906	0.8445
7	0.8065	0.5794	0.8468
8	0.8154	0.5872	0.8579
9	0.7942	0.5861	0.8423
10	0.7886	0.5716	0.8445
Mean	0.81	0.58	0.86

Based on Table 4, it can be seen that the semi-supervised SVM model has an accuracy of 0.86 or 86%, which makes it the model with the highest accuracy compared to the semi-supervised Naïve Bayes model with an accuracy of 81%, and the semi-supervised K-NN model which only has an accuracy of 58%.

In addition, this research also visualizes the distribution of words in tweet data about Kereta Cepat Jakarta-Bandung for each class category. Determining positive, negative, and neutral sentiment categories is based on the results of semi-supervised SVM modeling and predictions because it has the highest accuracy value compared to other models. The following is the distribution of words for each sentiment category displayed in the word cloud graph.

Figure 4. Word Cloud Category of Positive Sentiment

Based on Figure 4, the positive sentiment category contains words that are frequently used, including “indonesia”, “presiden” (president), “jokowi”, “presidenjokowihebat” (President Jokowi was great), “bumn”, “moga” (hopefully), “proyek” (project), “negara” (country), and “selesaikan” (finish). This shows that people who use Twitter in the positive tweet category are optimistic about completing the fast train project. Besides that, the people are also grateful to President Jokowi for bringing the fast train to Indonesia.

Figure 5. Word Cloud Category of Negative Sentiment

Meanwhile, based on Figure 5, it can be seen that the negative sentiment category consists of words like "proyek" (project), "argo parahyangan" (Argo Parahyangan train), "ikn" (IKN), "padalarang," "utang" (debt), "rugi" (loss), etc. This seems to indicate that Twitter users in the negative tweet category are concerned about some of the shortcomings of the fast train project, such as fears that it will eventually be able to delete the Argo Parahyangan train route and the high-speed train stopping only at Padalarang rather than in the city center of Bandung. Furthermore, people assume that the high-speed train project is the same as the IKN project, which is also financed with debt and is considered detrimental to the country and its future society.

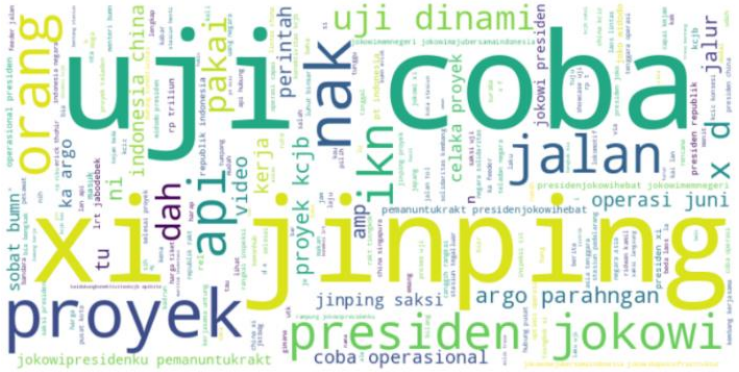

Figure 6. Word Cloud Category of Neutral Sentiment

Based on Figure 6, the tweets with a neutral category often contain words including “uji coba” (trials), “xi jinping”, “proyek” (project), and “presiden jokowi” (President Jokowi). This is relevant considering that the scraping period for collecting tweet data was taken during the period which included the Kereta Cepat Jakarta-Bandung dynamic test witnessed directly by President Jokowi and President Xi Jinping via video teleconference.

5. Conclusion

From the results of Model 1 and Model 2, it can be concluded that the SVM method performs better in predicting and classifying tweets according to their class categories than the Naïve Bayes and K-NN methods. This is indicated by the accuracy value of the SVM model in Model 1 and the semi-supervised SVM in Model 2, which has the highest accuracy compared to the other models. In addition, from the results of this study, it can also be seen that the semi-supervised learning process can increase the accuracy of the Naïve Bayes and SVM models but reduce the accuracy of the K-NN model. On the other hand, from the prediction results of Model 1, overall, more tweets are categorized as negative compared

to tweets that are categorized as positive and neutral. This also concludes that more Twitter users disagree with the existence of the Kereta Cepat Jakarta-Bandung compared to people who agree with the fast train.

It cannot be denied that this research also has several limitations, including only comparing three classification methods, namely Naïve Bayes, K-NN, and SVM. In addition, manual labeling by only two people resulted in more than 1,000 tweets being discarded and not used for modeling. Therefore, further research can use several other classification methods, such as Decision Tree, Random Forest, Gradient Boosting, Neural Network, etc., so that they can provide more comprehensive comparability. On the other hand, manual labeling can be done by more than two people, thereby reducing wasted tweets because each tweet has a label with the most label categories given by the person who labeled the tweet. Besides, collaborating with experts is also recommended in labeling because it can reduce researcher subjectivity.

References

[1] Karim H A, Lis Lesmini S H, Sunarta D A, SH M, Suparman A, SI S, ... and Bus M 2023 Manajemen transportasi Cendikia Mulia Mandiri

[2] Yamin M and Windymadaksa S 2017 Pembangunan kereta cepat Jakarta-Bandung sebagai mercusuar hubungan Indonesia-Tiongkok *Jurnal Politik Profetik* **5** 200-218

[3] Handyono 2016 Manfaat dari proyek kereta cepat Jakarta – Bandung. Diakses pada 29 Agustus 2023

[4] Gusman H 2020 Fakta dan Masalah Kereta Cepat Jakarta – Bandung https://tirto.id/fakta-dan-masalah-kereta-cepat-jakarta-bandung-eG7s

[5] Rosojati H, Darmastuti S and Atiandina D 2023 Menatap Sustainable Development pada Kereta Cepat. *Jurnal Sosial Ekonomi dan Humaniora* **9** 19-29

[6] Hadna N M S, Santosa P I and Winarno W W 2016 Studi literatur tentang perbandingan metode untuk proses analisis sentimen di Twitter. *Semin. Nas. Teknol. Inf. dan Komun* 57-64

[7] Salaatsa T K and Sibaroni Y 2022 Sentiment Analysis on the Construction of the Jakarta-Bandung High-Speed Train on Twitter Social Media Using Recurrent Neural Networks Method *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi* **13** 102-110

[8] Pravina A M, Cholissodin I and Adikara P P 2019 Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM) *Teknol. Inf. dan Ilmu …* http://j-ptiik.ub.ac.id/index.php/jptiik/article/view/4793

[9] Haekal H Z, Rolly M A and Woro I R 2023 Perbandingan Model Svm, Knn Dan Naïve Bayes Untuk Analisis Sentiment Pada Data Twitter: Studi Kasus Calon Presiden 2024. *JIMPS: Jurnal Ilmiah Mahasiswa Pendidikan Sejarah* **8** 2083-2093

[10] Macrohon J J E, Villavicencio C N, Inbaraj X A and Jeng J H 2022 A semi-supervised approach to sentiment analysis of tweets during the 2022 Philippine presidential election. *Information* **13** 484

[11] Lee V L S, Gan K H, Tan T P and Abdullah R 2019 Semi-supervised learning for sentiment classification using small number of labeled data. *Procedia Computer Science* **161** 577-584

[12] Kurniawan B, Aldino A A and Isnain A R 2022 Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (Pse) Menggunakan Algoritma Bidirectional Encoder Representations From Transformers (Bert) *J. Teknol. dan Sist. Inf*, **3** 98-106

[13] Ganesan K 2015 A Brief Note on Stop Words for Text Mining and Retrieval. Available at: http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html.

[14] Cvijikj I P and Michahelles F 2011 Understanding Social Media Marketing: A Case Study on Topics, Categories and Sentiment on a Facebook Brand Page

[15] Ligthart A, Catal C and Tekinerdogan B 2021 Systematic reviews in sentiment analysis: a tertiary study *Springer Netherlands* **54**

[16] Hesaputra A P, Saputra R D and Wibowo Y H 2022 Identifikasi Konten Dewasa pada Cuitan Twitter Menggunakan Metode BiLSTM Sebagai Upaya Mengatasi Penyebaran Pornografi Untuk Indonesia Maju. *Khazanah: Jurnal Mahasiswa* **14**
[17] He Y and Zhou D 2011 Self-training from labeled features for sentiment analysis *Inf. Process. Manag.* **47** 606–616, 2011
[18] Trstenjak B, Mikac S and Donko D 2014 KNN with TF-IDF based framework for text categorization *Procedia Eng.* **69** 1356–1364
[19] Lestandy M, Abdurrahim A and Syafa'ah L 2021 Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* **5** 802-808
[20] Taheri S and Mammadov M 2013 Learning The Naive Bayes Classifier With Optimization Models. *In International Journal of Applied Mathematics and Computer Science* 787– 795.
[21] Ardana N K, Ruhiyat R, Amany N, Irawan T K, Raymond R, Karunia R and Fauzia S 2023 Perbandingan Metode KNN, Naive Bayes, dan Regresi Logistik Binomial dalam Pengklasifikasian Status Ekonomi Negara *Jambura Journal of Mathematics* **5** 404-418
[22] Rohmansyah F A, Bintoro B and Santoso I 2023 Analisis Sentimen Terhadap Penerapan Sistem Ganjil Genap Menggunakan Metode K-Nearest Neighbor (KNN) *IKRA-ITH Informatika: Jurnal Komputer dan Informatika* **7** 165-169
[23] Ahluna F, Tutuarima C J and Santoso I 2023 Metode K-Nearest Neighbor Untuk Analisis Sentimen Tentang Penghapusan Ujian Nasional. *IKRA-ITH Informatika: Jurnal Komputer dan Informatika*, **7** 170-175
[24] Fitriansyah A R and Sibaroni Y 2023 Analisis Sentimen Terhadap Pembangunan Kereta Cepat Jakarta-Bandung Pada Media Sosial Twitter Menggunakan Metode SVM dan GloVe Word Embedding *eProceedings of Engineering* **10**
[25] Qisthiano M R, Ruswita I and Prayesy P A 2023 Implementasi Metode SVM dalam Analisis Sentimen Mengenai Vaksin dengan Menggunakan Python 3 *Teknologi: Jurnal Ilmiah Sistem Informasi* **13** 1-7
[26] Pamuji F Y and Putri S D A 2023 Komparasi Metode SMOTE dan ADASYN Untuk Penanganan Data Tidak Seimbang MultiClass *Jurnal Informatika Polinema* **9** 331-338
[27] Firmansyach W A, Hayati U and Wijaya Y A 2023 Analisa Terjadinya Overfitting Dan Underfitting Pada Algoritma Naive Bayes Dan Decision Tree Dengan Teknik Cross Validation *JATI (Jurnal Mahasiswa Teknik Informatika)* **7** 262-269
[28] Hernandez-Suarez A, Sanchez-Perez G, Toscano-Medina K, Martinez-Hernandez V, Sanchez V and Perez-Meana H 2018 A web scraping methodology for bypassing twitter API restrictions *arXiv preprint arXiv:1803.09875*
[29] Kompas 2023 Uji Dinamis Kereta Cepat 16 November 2022: Tujuan, Lokasi, Jarak & Kecepatan cited 30 August 2023 Available from: https://www.kompas.tv/nasional/348247/uji-dinamis-kereta-cepat-16-november-2022-tujuan-lokasi-jarak-kecepatan