Draft Jurnal Analisis Sentimen Transportasi Kereta Jakarta - Adriel Savero (20200803069).docx

by arjunafadli@gmail.com 1

Submission date: 25-Jul-2024 11:21AM (UTC+0700)

Submission ID: 2422124793

File name: Draft_Jurnal_Analisis_Sentimen_Transportasi_Kereta_Jakarta_-_Adriel_Savero_20200803069_.docx (6.26M)

Word count: 4469

Character count: 24461

ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Public Opinion Sentiment Analysis on Train Transport in Jakarta Using a Hybrid Model

Adriel Savero[1], Sawali Wahyu[2*]
[1]Dept. of Information System, Universitas Esa Unggul, Indonesia
[2]Dept. of Informatics Engineering, Universitas Esa Unggul, Indonesia
[1]saveroa4@gmail.com, [2]sawaliwahyu@esaunggul.ac.id

Abstract. Transportation is a key element in smoothing the wheels of the economy and connecting various regions, especially in big cities like Jakarta which has a high population density. This leads to dense and complex traffic conditions. Improving the quality and facilities of public transportation is important to overcome these problems. However, people are still reluctant to use public transportation for various reasons. Therefore, it is important to understand public sentiment towards public transportation in Jakarta. This research focuses on sentiment analysis of train-based transportation, namely KRL, MRT, and LRT. Sentiment analysis is conducted using a hybrid learning model with a voting model method, which combines SVM, logistic regression, and CNN algorithms. The data used is labeled with InSet sentiment dictionary and extracted features using TF-IDF method. The modeling results show that this hybrid model produces 89% accuracy for the KRL dataset, 88% for the MRT dataset, and 81% for the LRT dataset. However, this model still has difficulty in predicting neutral and positive classes. The results of this study show that hybrid learning with the voting model method can provide quite good results in public transportation sentiment analysis, but there is still room for improvement in the classification of neutral and positive sentiments. The findings provide important insights for the development of strategies to improve the quality of public transportation and encourage people to use the service more.
.

Keywords: CNN, Hybrid Model, KRL, Logistic Regression, LRT, MRT, Sentiment Analysis, SVM

INTRODUCTION

Transportation is a very important and strategic means of accelerating the economy, connecting regions, and influencing various aspects of life [1], [2]. The importance of the role of transportation in an area is reflected in the increased mobility of people and goods, which is influenced by population growth rates, especially in big cities like DKI Jakarta. DKI Jakarta is one of the provinces with the highest population density in Indonesia, which has an impact on traffic congestion. Improving the quality and facilities of public transportation aims to reduce congestion, but there are still many people who use private vehicles for reasons such as travel time, safety, and convenience [3]. Although public transportation such as KRL offers low cost, full conditions during peak hours make users have to share seats [4]. Besides KRL, the transportation system in Jakarta also includes MRT and LRT, each of which has its own advantages and disadvantages [5].

Collecting public opinion on KRL, MRT, and LRT in Jakarta can be done through social media, with Twitter as one of the effective platforms. Twitter user comments can be used as a data source using crawling techniques for public sentiment analysis [4].

Previous research entitled "Comparison of Algorithms for Sentiment Analysis on Commuterline Public Transportation Twitter" compared SVM, SVM-PSO, Naïve Bayes, and NB-Adaboost algorithms based on accuracy, precision, recall, and AUC, with an average value of SVM: 78.15%, SVM-PSO: 79.47%, Naive Bayes: 76.7%, and NB-Adaboost: 78.80% [6]. The research "Sentiment Analysis on Twitter Social Media towards Jakarta MRT Using Machine Learning" uses the Naïve Bayes algorithm to assess MRT services with the results of 48.8% positive sentiment, 22.4% negative, and 28.8% neutral with 76.21% accuracy [7]. The research "Analysis of Public Sentiment towards the Jakarta LRT Trial Using Improved K-Nearest Neighbor and Information Gain" using information gain feature selection and K-fold cross validation, found that variations in the k value did not significantly affect the f-measure and increasing the number of features or thresholds was not always directly proportional to running time [5].

DOI: 10.24014/coreit.vxxx.xxxx

This research aims to analyze public sentiment towards the use of KRL, MRT, and LRT in Jakarta using a hybrid model that combines machine learning and deep learning. This approach is expected to improve prediction accuracy compared to a single model [8]. This hybrid model uses a voting system for sentiment classification, where the final prediction of the model is if there are 2 or more models that have the same prediction then the prediction becomes the final prediction, but if there is no similarity in predictions from all models then the prediction results are selected based on the average probability (soft voting).

This research is entitled "Sentiment Analysis of Public Opinion on the Use of Rail Transportation in Jakarta Using Hybrid Model" and is intended to produce high levels of accuracy, precision, and recall so as to provide more precise information about public sentiment towards KRL, MRT, and LRT. This hybrid model will combine three algorithms namely Support Vector Machine (SVM), Logistic Regression and Convolutional Neural Network (CNN).

METHODS

Sentiment analysis is a technique for extracting information from opinions by automatically processing and understanding text data to identify and understand expressions in opinions, which aims to reveal and categorize user opinions into positive, negative, or neutral sentiments [9], [10]. By utilizing Natural Language Processing (NLP) technology and Machine Learning (ML) techniques, sentiment analysis extracts and analyzes large text data to understand the opinions of individuals or groups on a subject, both for personal and business purposes [11]. The sentiment analysis process involves the stages of data collection, preprocessing, data transformation, feature selection, and classification. In the preprocessing stage, activities such as cleaning the text from irrelevant characters (cleaning), homogenizing the text into lowercase letters (case folding), removing common words (filtering), and converting words into basic forms (stemming) are carried out [12].

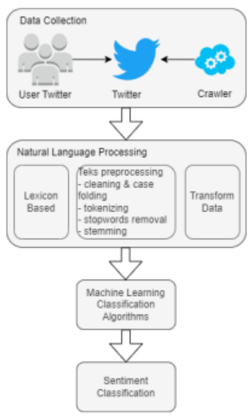

Figure 1 . Architecture sentiment analysis

This research focuses on sentiment analysis of public transportation in Jakarta through the process of collecting data from Twitter social media using crawler tools. The collected data is then processed through a series of natural language processing (NLP) stages, which include text cleaning, sentiment labeling, and data transformation using the TF-IDF method. After going through these stages, the data was analyzed using a hybrid model to produce a sentiment classification. This research provides important insights into the public perception of public transportation in Jakarta, which can be used as a basis for improving the quality and facilities of such transportation.

The classification process uses machine learning and deep learning with a hybrid model that combines the strengths of several machine learning and deep learning algorithms to improve the accuracy of sentiment analysis. This hybrid model uses a voting model approach that combines Support Vector Machine (SVM), Logistic Regression, and Convolutional Neural Network (CNN). SVM is an extension of the optimal separation hyperplane method by transforming the input space into a high-dimensional feature space through a pre-selected non-linear mapping, then constructing an optimal separation hyperplane in the feature space [13]. This method allows to relate independent variables to the probability of occurrence of dependent variables in the form of categories, such as 0 or 1, yes or no, large or small. Independent variables in this context can be categorical variables that describe certain categories or groups in regression analysis [14]. Convolutional Neural Networks (CNN) are used because they are proven to be able to detect information with a high level of accuracy [15].

Evaluation is done to find the best classification results through hyperparameter tuning of the classification algorithm used. This research focuses on classification performance measures reflected through the confusion matrix. Confusion matrix is a very useful tool for analyzing the extent to which a classification method is able to identify observed objects from various classes properly or accurately [16].

Data processing in this study goes through the process of collecting data, text preprocessing, data labeling, feature extraction, classification using hybrid models and calculating model performance.

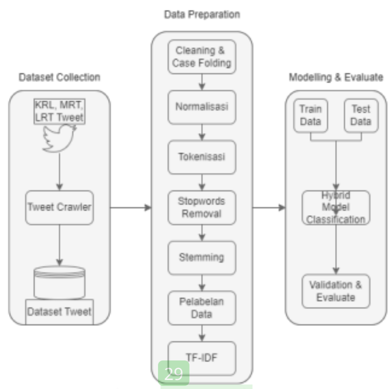

Figure 2. Research Stage

Figure 1 is the flow of this research.

1. **Collecting Data :**
   The data was obtained from Twitter using a crawling technique using the keywords "Jakarta KRL", "Jakarta MRT", and "Jakarta LRT" within the range of January 1 - December 31, 2023, except for "Jakarta LRT" from August 26 - December 31, 2023.
2. **Data Preperation :**
   In this stage there are several activities carried out, namely cleaning & case folding, normalization, tokenization, stopwords removal, stemming, sentiment labeling and word embedding.
3. **Modelling & Evaluate :**
   The modeling and evaluation stage is the stage of building a hybrid model and evaluating it to get the best results.

RESULT AND DISCUSSION

In this study, the tools used to process and build models using the python programming language. in data collection, data about KRL is obtained as much as 7795 lines of data, data about MRT as much as 9952 lines of data, data about LRT 2308 lines of data. from the various attributes obtained, only the "full_text" attribute will be used as input data in this study.

created_at	id_str	full_text	uote_cou	eply_coun	tweet_cou	orite_cou	lang	user_id_st	ersation_i	username	tweet_url
Sat Sep 02	1,7E+18	@miaetsu	0	1	0	0	in	7,9E+07	1,7E+18	mrtjakarta	https://tw
Wed Jun 1	1,7E+18	Keujanan	0	0	0	0	in	1,8E+08	1,7E+18	chaderyst	https://tw
Sun Mar 1	1,6E+18	@tanyarlf	0	0	0	1	in	1,1E+18	1,6E+18	tahukegej	https://tw
Mon Mar	1,6E+18	@adriansy	0	0	0	0	in	1,4E+18	1,6E+18	umbiygud	https://tw
Sat Mar 18	1.6E+18	@JakartaA	0	0	0	0	in	1.6E+18	1.6E+18	sutoyo442	https://tw

Figure 3. Data about KRL

created_at	id_str	full_text	uote_cour	eply_count	tweet_cou	vorite_cou	lang	user_id_st	ersation_i	username	tweet_url
Fri Dec 29	1,7E+18	Ada Perub	0	0	0	0	in	1,5E+18	1,7E+18	idtodaycoi	https://twi
Fri Dec 29	1,7E+18	@UGM_FE	0	0	0	0	in	1,4E+18	1,7E+18	rioakbar_p	https://twi
Fri Dec 29	1,7E+18	https://t.c	0	0	0	0	en	1,2E+18	1,7E+18	khailim26	https://twi
Fri Dec 29	1,7E+18	@Watchm	0	0	0	0	in	1,1E+09	1,7E+18	Rifaldiptr	https://twi
Fri Dec 29	1,7E+18	siapa disin	0	0	0	0	in	1,7E+18	1,7E+18	wonnyuw	https://twi

Figure 4. Data about MRT

created_at	id_str	full_text	uote_cour	eply_count	tweet_cou	vorite_cou	lang	user_id_st	ersation_i	username	tweet_url
Tue Aug 2!	1,7E+18	LRT Jabod	0	1	0	0	in	1,58E+18	1,7E+18	buihdilaut	https://tw
Tue Aug 2!	1,7E+18	Dewi Pers	0	0	0	0	in	96970566	1,7E+18	infoterkini	https://tw
Tue Aug 2!	1,7E+18	Viral Rok N	0	0	0	0	in	96970566	1,7E+18	infoterkini	https://tw
Tue Aug 2!	1,7E+18	Kebangun	0	1	1	4	in	7,51E+17	1,7E+18	firzaaak	https://tw
Tue Aug 2!	1,7E+18	Saat @eric	0	0	0	1	in	1,38E+18	1,7E+18	Siputty1	https://tw

Figure 5. Data about LRT

The data that has been collected and merged previously, will be processed using the python programming language. This is done to make it easier for the model to classify because the data is more structured.

Preprocessing Text

1. Cleansing and Case Folding

In this process, the data will be cleaned from various characters such as URLs, numbers, emoticons, mentions, hashtags, and whitespace, and uniformize letters to lowercase.

index	full_text	clean_text
0	@miaetsumi Hai, Kak. Untuk kartu KRL Commuter Line dapat digunakan untuk naik MRT Jakarta dengan minimal saldo Rp 14000, dan pastikan saldo mencukupi saat melakukan tap in. Terima kasih. -AL.	hai kak untuk kartu krl commuter line dapat digunakan untuk naik mrt jakarta dengan minimal saldo rp dan pastikan saldo mencukupi saat melakukan tap in terima kasih al
1	Keujanan di Bogor, menggigil selama di KRL, pas sampe Jakarta anget :))) ku senang kali ini	keujanan di bogor menggigil selama di krl pas sampe jakarta anget ku senang kali ini

Figure 6. Cleansing data

2. Tokenization

Tokenization is the stage of breaking text into smaller units in the form of words, phrases or symbols called tokens. This tokenization is intended to facilitate the process of normalization, stopwords removal and stemming.

index	clean_text	tokenized_text
0	hai kak untuk kartu krl commuter line dapat digunakan untuk naik mrt jakarta dengan minimal saldo rp dan pastikan saldo mencukupi saat melakukan tap in terima kasih al	['hai', 'kakak', 'untuk', 'kartu', 'krl', 'commuter', 'line', 'dapat', 'digunakan', 'untuk', 'naik', 'mrt', 'jakarta', 'dengan', 'minimal', 'saldo', 'rp', 'dan', 'pastikan', 'saldo', 'mencukupi', 'saat', 'melakukan', 'tap', 'di', 'terima', 'kasih', 'al']
1	keujanan di bogor menggigil selama di krl pas sampe jakarta anget ku senang kali ini	['keujanan', 'di', 'bogor', 'menggigil', 'selama', 'di', 'krl', 'pas', 'sampai', 'jakarta', 'anget', 'ku', 'senang', 'kali', 'ini']

Figure 7. Tokenization

3. Normalization

Normalization is the process of converting words that are not standard, wrong typing, abbreviations and slang into the actual form or standard words. This process is carried out using a collection of words that have been normalized and created by adeariniputri and uploaded to the github.com platform.

index	tokenized_text	normalisasi
0	['hai', 'kakak', 'untuk', 'kartu', 'krl', 'commuter', 'line', 'dapat', 'digunakan', 'untuk', 'naik', 'mrt', 'jakarta', 'dengan', 'minimal', 'saldo', 'rp', 'dan', 'pastikan', 'saldo', 'mencukupi', 'saat', 'melakukan', 'tap', 'di', 'terima', 'kasih', 'al']	hai kakak untuk kartu krl commuter line dapat digunakan untuk naik mrt jakarta dengan minimal saldo rp dan pastikan saldo mencukupi saat melakukan tap di terima kasih al
1	['keujanan', 'di', 'bogor', 'menggigil', 'selama', 'di', 'krl', 'pas', 'sampai', 'jakarta', 'anget', 'ku', 'senang', 'kali', 'ini']	keujanan di bogor menggigil selama di krl pas sampai jakarta anget ku senang kali ini

Figure 8. Normalization

4. Stopwords Removal

Stopwords removal is the process of removing words or tokens that have no influence on sentiment such as conjunctions, prepositions and articles. Stopwords removal is intended to minimize features that have no influence on sentiment so that the model will be more precise in predicting sentiment.

index	normalisasi	stopwords_removal
0	hai kakak untuk kartu krl commuter line dapat digunakan untuk naik mrt jakarta dengan minimal saldo rp dan pastikan saldo mencukupi saat melakukan tap di terima kasih al	['hai', 'kakak', 'kartu', 'krl', 'commuter', 'line', 'digunakan', 'naik', 'mrt', 'jakarta', 'minimal', 'saldo', 'rp', 'pastikan', 'saldo', 'mencukupi', 'melakukan', 'tap', 'terima', 'kasih', 'al']
1	keujanan di bogor menggigil selama di krl pas sampai jakarta anget ku senang kali ini	['keujanan', 'bogor', 'menggigil', 'selama', 'krl', 'pas', 'jakarta', 'anget', 'ku', 'senang', 'kali']

Figure 9. Stopwords removal

5. Stemming

Stemming is the process of removing affixes, suffixes and inserts so that words or tokens will become basic words. The main purpose of Stemming in natural language processing (NLP) and text analysis is to simplify and unify different forms of words that have the same basic meaning. The Stemmer used in this process is StemmerFactory from Sastrawi.

index	stopwords_removal	stemming
0	['hai', 'kakak', 'kartu', 'krl', 'commuter', 'line', 'digunakan', 'naik', 'mrt', 'jakarta', 'minimal', 'saldo', 'rp', 'pastikan', 'saldo', 'mencukupi', 'melakukan', 'tap', 'terima', 'kasih', 'al']	hai kakak kartu krl commuter line guna naik mrt jakarta minimal saldo rp pasti saldo cukup laku tap terima kasih al
1	['keujanan', 'bogor', 'menggigil', 'selama', 'krl', 'pas', 'jakarta', 'anget', 'ku', 'senang', 'kali']	keujanan bogor gigil lama krl pas jakarta anget ku senang kali

Figure 10. Stemming

After these stages, null data and duplicate data caused by the data cleansing process, Stopwords removal and Stemming will be cleaned by deleting them. Therefore, the amount of data will be reduced due to the presence of null data and there is also data that has duplication. The process results in the following amount of data:

```
Data KRL : 7426
Data MRT : 8613
Data LRT : 2182
```

Figure 11. Data after preprocess

6. Sentiment Labeling

Data labeling is done using a lexicon-based method. This method uses a sentiment dictionary from InSet built by Fajri Koro and Gemala Y. Rahmaningtyas where the word dictionary contains 6609 negative sentiments or equivalent to 65% of the positive sentiments of 3609 words [17]. At this stage, each token contained in the InSet sentiment dictionary will be summed based on the score on the word to get a polarity score. If the total polarity score is more than 0 is a positive sentiment, a polarity score of 0 is a neutral sentiment and a polarity score smaller than 0 is a negative sentiment [18].

index	stemming	polarity_score	sentiment
0	hai kakak kartu krl commuter line guna naik mrt jakarta minimal saldo rp pasti saldo cukup laku tap terima kasih al	5	Positif
1	keujanan bogor gigil lama krl pas jakarta anget ku senang kali	-9	Negatif
2	aku bkn orang jakarta tiap kesana seru gitu kalau naik transport umum entah trans jkt krl atau mrt bahkan kalau zombie lari suatu seru sendiri bagi wkwk mungkin beda kalau aku tetap lama jkt	-29	Negatif

Figure 12. Sentimen labeling

From the results of sentiment labeling using the lexicon-based method and using the sentiment dictionary from InSet, it produces tweets with the most negative sentiment from the three datasets, namely datasets about KRL, MRT and LRT. In the KRL opinion data there are 6277 or 84.53% negative sentiments, 797 or 10.73% positive sentiments and 352 or 4.74% neutral sentiments. In MRT opinion data there are 6413 or 74.46% negative sentiments, 1441 or 16.73% positive sentiments and 759 or 8.81% neutral sentiments. While in the LRT opinion data there are 1681 or 77.04% negative sentiments, 370 or 16.96% negative sentiments and 131 or 6% neutral sentiments.

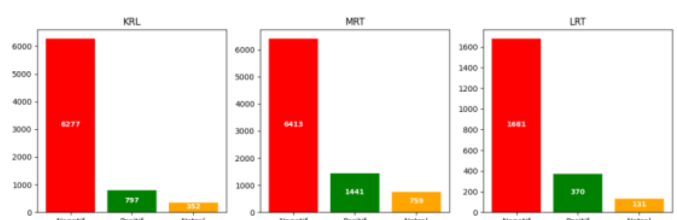

Figure 13. Sentiment distribution

7. feature extraction

This feature extraction stage uses the TF-IDF method, this method is used because the dataset is not too large. with this TF-IDF method the data is limited to a max feature of 2500 words that have important meanings, the minimum occurrence of words is in 8 lines of data, and words that appear in more than 70% of the entire dataset will be ignored.

Token	TF-IDF Value
accesscaccess	0.234475
aplikasi	0.175016
berangkat	0.169481
cilebutjakarta	0.330996
commuterline	0.183067

Figure 14. TF-IDF result

Modelling

Modeling is done by dividing the dataset into 70% training data and 30% test data. Hybrid modeling is done by combining prediction results from SVM models, logistic regression models and CNN models with a voting system.

1. Modeling of KRL data

By using the svm model on data about KRL, the following hyperparameters are used kernel = 'rbf' with C = 100 and gamma = 0.1, and probability = True. These hyperparameters are obtained from several configurations using GridSeaarchCV(). So from modeling using SVM, the accuracy value is 87.52%, precision 85.69%, recall 87.52%, and f1-score 84.77%.

```
Accuracy Score   : 87.52
Precision Score  : 85.69
Recall Score     : 87.52
f1-score         : 84.77
```

Figure 15. SVM model evaluation result of KRL data

While in the logistic regression model, a matching of the hyperparameter configuration that produces the best evaluation value using GridSearchCV() is also carried out and the best configuration is obtained, namely solver = 'saga', C = 10 and toll = 0.01. So that it produces an accuracy value of 88.6%, precision 86.2%, recall 88.6%, and f1-score 86.6%.

```
Accuracy Score   : 88.42
Precision Score  : 86.01
Recall Score     : 88.42
f1-score         : 86.59
```

Figure 16. Logistic regression model evaluation result of KRL data

In the CNN model, several configurations are performed for multiclass cases. The first layer of the model is Conv1D which has 64 filters and kernel size = 3 and ReLU activation function, followed by MaxPooling1D layer with pool size = 1 to reduce the dimensionality of the data while retaining important information. Then, the Flatten layer converts the output into a one-dimensional vector, followed by the dense layer with 100 units and a ReLU activation function. The last layer has the number of units corresponding to the number of classes in the target data and uses a softmax activation function. The model was compiled with the optimizer 'adam' and loss function 'categorical_crossentropy', then trained with 10 epochs and batch size 32.

Using this configuration, the CNN model can produce an accuracy value of 87.88%, precision 86.89%, recall 87.88% and f1-score of 87.28%.

```
Accuracy Score   : 88.24
Precision Score  : 87.05
Recall Score     : 88.24
f1-score         : 87.46
```

Figure 17. CNN model evaluation result of KRL data

Models from the SVM, Logistic Regression and CNN algorithms that have been built previously will be trained using the soft voting method, this is done to get the prediction results from the soft voting. Where these results will be the final choice if there are not 2 or more models that have similar predictions. From the evaluation results of the hybrid model, the model has an accuracy rate of 89%, precision of 87%, recall of 89% and f1-score of 87%.

```
Accuracy Score   : 88.69
Precision Score  : 86.77
Recall Score     : 88.69
f1-score         : 87.07
```

Figure 18. Hybrid model evaluation result of KRL data

Figure 19. shows that this hybrid model shows better performance compared to the performance of models that work individually but the performance improvement of this hybrid model is not too significant because it can be seen that the accuracy, precision, recall and f1-score values are not too much different.

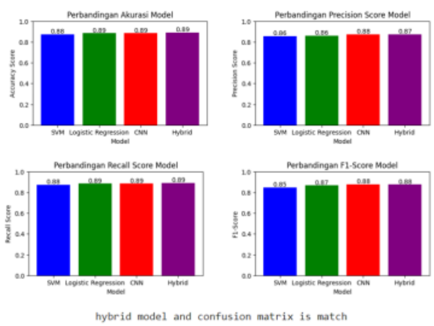

Figure 19. Comparison of KRL Data Evaluation Results

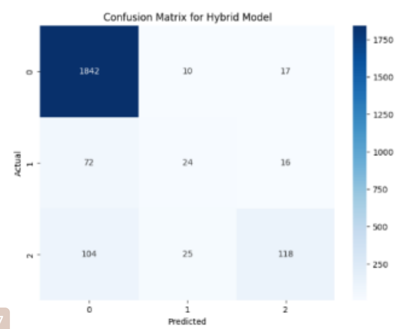

Figure 20. Confusion matrix of hybrid model for KRL data

In figure 20. the confusion matrix shows that the model has difficulty in predicting the neutral class (1) and in the positive class (2) it still has a little difficulty in predicting it, while in the negative class (0) the model can predict it properly. From the confusion matrix in figure 20. the precision, recall and f1-score values of each sentiment class can be calculated as follows.

Table 1. Evaluate values of each sentiment class of KRL Data

	Negative (0)	Neutral (1)	Positive (2)
Precision	$\frac{1842}{1842+72+104} = 0{,}91$	$\frac{24}{10+24+25} = 0{,}41$	$\frac{118}{17+16+118} = 0{,}78$
Recall	$\frac{1842}{1842+10+17} = 0{,}99$	$\frac{24}{72+24+16} = 0{,}21$	$\frac{118}{104+25+118} = 0{,}48$
F1-score	$2 \times \frac{0{,}91 \times 0{,}99}{0{,}91+0{,}99} = 0{,}95$	$2 \times \frac{0{,}41 \times 0{,}21}{0{,}41+0{,}21} = 0{,}28$	$2 \times \frac{0{,}78 \times 0{,}48}{0{,}78+0{,}48} = 0{,}59$

2. Modelling of MRT data

By using the svm model on data about MRT, the following hyperparameters are used kernel = 'rbf' dengan C = 10 dan gamma = 1 serta probability = True. These hyperparameters are obtained from several configurations using GridSeaarchCV(). So from modeling using SVM, the accuracy value is 81,07%, precision 76,73%, recall 81,07%, and f1-score 77,25%.

```
Accuracy Score   : 81.07
Precision Score  : 76.73
Recall Score     : 81.07
f1-score         : 77.25
```

Figure 21. SVM model evaluation result of MRT data

While in the logistic regression model, a matching of the hyperparameter configuration that produces the best evaluation value using GridSearchCV() is also carried out and the best configuration is obtained, namely solver = 'liblinear', C = 10 dan tol = 0.001. So that it produces an accuracy value of 87%, precision 87%, recall 87%, and f1-score 87%.

```
Accuracy Score   : 87.46
Precision Score  : 86.56
Recall Score     : 87.46
f1-score         : 86.59
```

Figure 22. Logistic regression model evaluation result of MRT data

In the CNN model, several configurations are performed for multiclass cases. The CNN model is configured with Conv1D with 32 filters, kernel size 3, and ReLU activation function, followed by a MaxPooling1D layer with pool size 2 to reduce the dimensionality of the data. Next, the Flatten layer converts the output into a one-dimensional vector, followed by a dense layer with 100 units, and an output layer with the number of units corresponding to the number of classes in the target data and softmax activation function. The model was compiled with the optimizer 'adam' and loss function 'categorical_crossentropy', then trained using the rescaped training data with 10 epochs and batch size 32. Using this configuration, the CNN model can produce accuracy, precision, recall and f1-score values of 86%.

```
Accuracy Score   : 85.84
Precision Score  : 85.27
Recall Score     : 85.84
f1-score         : 85.51
```

Figure 23. CNN model evaluation result of MRT data

Models from the SVM, Logistic Regression and CNN algorithms that have been built previously will be trained using the soft voting method, this is done to get the prediction results from the soft voting. Where these results will be the final choice if there are not 2 or more models that have similar predictions. From the evaluation results of the hybrid model, the model has an accuracy rate of 87,5%, precision 86,67%, recall 87,5% dan f1-score sebesar 86,57%.

```
Accuracy Score   : 87.50
Precision Score  : 86.67
Recall Score     : 87.50
f1-score         : 86.57
```

Figure 24. Hybrid model evaluation result of MRT data

Figure 25. shows that using the hybrid model on the MRT dataset shows improved performance in terms of accuracy, precision, recall, and fi-score. Of all the evaluation metrics, the hybrid model scored the highest compared to the models working in isolation.

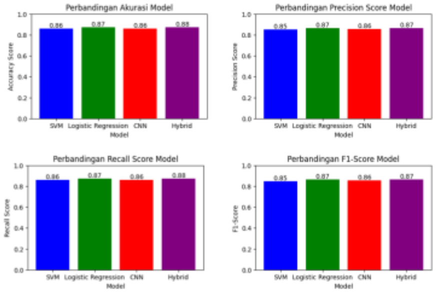

Figure 25. Comparison of MRT Data Evaluation Results

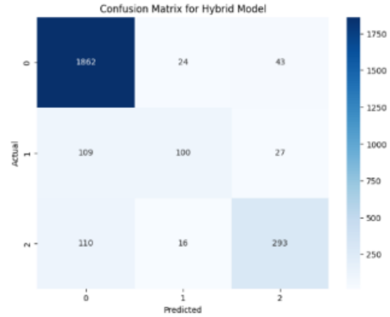

Figure 26. Confusion matrix of hybrid model for MRT data

In figure 26. the confusion matrix shows that the model can predict the negative class (0) well, and the prediction of the positive class (2) is quite good while the prediction of the neutral class (1) is still a struggle. From the confusion matrix in figure 26. the precision, recall and f1-score values of each sentiment class can be calculated as follows.

Table 2. Evaluate values of each sentiment class of KRL Data

	Negative (0)	Neutral (1)	Positif (2)
Precision	$\frac{1862}{1862+109+110} = 0{,}89$	$\frac{100}{24+100+16} = 0{,}71$	$\frac{293}{43+27+293} = 0{,}81$
Recall	$\frac{1862}{1862+24+43} = 0{,}97$	$\frac{100}{109+100+27} = 0{,}42$	$\frac{293}{110+100+293} = 0{,}7$
F1-score	$2 \times \frac{0{,}89 \times 0{,}97}{0{,}89+0{,}97} = 0{,}93$	$2 \times \frac{0{,}71 \times 0{,}42}{0{,}71+0{,}42} = 0{,}53$	$2 \times \frac{0{,}81 \times 0{,}7}{0{,}81+0{,}7} = 0{,}75$

3. Modelling of LRT data

Modeling on LRT data is also carried out the same process as modeling KRL and MRT data. For SVM models using GridSearchCV() shows the best hyperparameter configuration using kernel = 'rbf' with C = 10 and gamma = 1 and probability = True. With this configuration, the SVM model can produce an accuracy evaluation value of 81.07%, precision 76.73%, recall 81.07%, and f1-score 77.25%.

```
Accuracy Score    : 81.07
Precision Score   : 76.73
Recall Score      : 81.07
f1-score          : 77.25
```

Figure 27. SVM model evaluation result of LRT data

As for the logistic regression model using GridSearchCV(), the best hyperparameter configuration is C = 10, solver = 'saga', toll = 0.01. This configuration produces an accuracy value of 80.92%, precision 76.99%, recall 80.92%, and f1-score 78.34%.

```
Accuracy Score    : 80.92
Precision Score   : 76.99
Recall Score      : 80.92
f1-score          : 78.34
```

Figure 28. Logistic regression model evaluation result of LRT data

And for the deep learning model using CNN algorithm, the first layer is configured with Conv1D with 64 filters, kernel size 3, and ReLU activation function, followed by MaxPooling1D layer with pool size 1. Next, the Flatten layer converts the output into a one-dimensional vector, followed by a dense layer with 100 units, and an output layer with the number of units corresponding to the number of classes in the target data and softmax activation function. The model is compiled with the optimizer 'adam' and loss function 'categorical_crossentropy', then trained using rescaled training data with 10 epochs and batch size 64. With various configurations of CNN models and compilers, the accuracy value was 78.93%, precision 76.86%, recall 78.93% and f1-score 77.79%.

```
Accuracy Score   : 78.93
Precision Score  : 76.86
Recall Score     : 78.93
f1-score         : 77.79
```

Figure 29. CNN model evaluation result of LRT data

The prediction results of the three models are combined into one and voted to select the final prediction based on the highest frequency or soft voting method. With this hybrid model on LRT data, the accuracy value is 80%, precision is 76%, recall is 80% and f1-score is 78%.

```
Accuracy Score   : 81.37
Precision Score  : 76.86
Recall Score     : 81.37
f1-score         : 78.24
```

Figure 30. Hybrid model evaluation result of LRT data

In figure 31. the hybrid model on LRT data does not show significant differences with models built individually. In fact, the accuracy value of the hybrid model is the same as the SVM, logistic regression and CNN models. Which means this hybrid model has relatively the same performance as models that work individually.

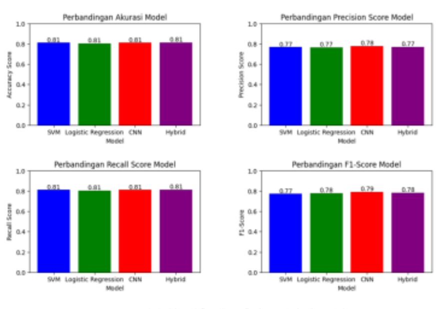

Figure 31. Comparison of KRL Data Evaluation Results

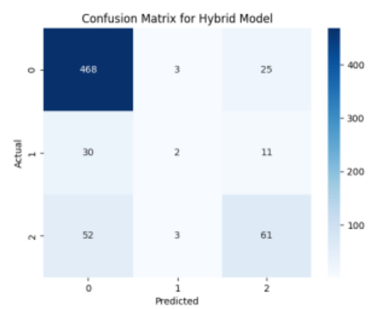

Figure 32. Confusion matrix of hybrid model for LRT data

In figure 32. the confusion matrix shows that the model can predict the negative class (0) well, and the prediction of the positive class (2) is not very good, while the prediction of the neutral class (1) the model has a terrible time. From the confusion matrix in figure 32. the precision, recall and f1-score values of each sentiment class can be calculated as follows.

Table 3. Hybrid model evaluation result of LRT data

	Negatif (0)	Netral (1)	Positif (2)
Precision	$\frac{468}{468+30+52} = 0{,}85$	$\frac{2}{3+2+3} = 0{,}25$	$\frac{61}{25+11+61} = 0{,}63$
Recall	$\frac{468}{468+3+25} = 0{,}94$	$\frac{2}{30+2+11} = 0{,}05$	$\frac{61}{52+3+61} = 0{,}53$
F1-score	$2 \times \frac{0{,}85 \times 0{,}94}{0{,}85+0{,}94} = 0{,}89$	$2 \times \frac{0{,}25 \times 0{,}05}{0{,}25+0{,}05} = 0{,}08$	$2 \times \frac{0{,}63 \times 0{,}53}{0{,}63+0{,}53} = 0{,}57$

CONCLUSION

The hybrid method used is a vote model where the model will produce a final prediction based on the majority prediction of each individual model, but if there is no similarity in the predictions of the three models then the final prediction is based on the average of the probabilities of the three models. From the results of data modeling with the hybrid method, the modeling results on the KRL dataset produced an accuracy value of 89%, the MRT dataset with an accuracy value of 88% and the LRT dataset with an accuracy value of 81%. In all modeling, both SVM, logistic regression, CNN and hybrid models have difficulty predicting neutral and positive classes on the three datasets, while positive classes can be well predicted by each model. This can be caused by imbalanced sentiment data on each dataset.

Modeling using a hybrid model on KRL and MRT data showed improved but not significant performance compared to models working individually. Whereas, modeling on LRT data, this hybrid model does not show an increase in performance but the evaluation value shows relatively the same as other models.

REFERENCES

[1] A. Fatoni and D. Hardianti, "Pengaruh Fasilitas Dan Kualitas Pelayanan Terhadap Keputusan Menggunakan Jasa Transportasi Mrt Atau Mass Rapid Transit," *Mediastima*, vol. 26, no. 1, pp. 117–134, 2020, doi: 10.55122/mediastima.v26i1.99.

[2] M. Kadarisman, "Kebijakan Transportasi Kereta Cepat Jakarta Bandung Dalam Mewujudkan Angkutan Ramah Lingkungan," *J. Manaj. Transp. Logistik*, vol. 4, no. 3, p. 251, 2018, doi: 10.54324/j.mtl.v4i3.167.

[3] D. I. Sari, Y. F. Wati, and Widiastuti, "Analisis Sentimen Dan Klasifikasi Tweets Berbahasa Indonesia Terhadap Transportasi Umum Mrt Jakarta Menggunakan Naïve Bayes Classifier," *J. Ilm. Inform. Komput.*, vol. 25, no. 1, pp. 64–75, 2020, doi: 10.35760/ik.2020.v25i1.2427.

[4] A. Novantirani, M. K. Sabariah, and V. Effendy, "Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine," *e-Proceeeding Eng.*, vol. 2, no. 1, pp. 1–7, 2015.

[5] M. Okza Pradhana and S. Adinugroho, "Analisis Sentimen Masyarakat terhadap Uji Coba LRT Jakarta Menggunakan Improved K-Nearest Neighbor dan Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 6, pp. 1888–1896, 2020, [Online]. Available: http://j-ptiik.ub.ac.id

[6] R. Novaneliza, F. Handayani, R. J. Suhandar, H. Surono, N. S. Azzahra, and D. Nadilla, "Perbandingan Algoritma Untuk Analisis Sentimen Pada Twitter Transportasi Umum Commuterline," *J. Sains Komput. Inform. (J-SAKTI*, vol. 7, no. 1, pp. 13–21, 2023.

[7] D. Agustina and F. Rahmah, "Analisis Sentimen pada Sosial Media Twitter terhadap MRT Jakarta Menggunakan Machine Learning," *Insearch (Information Syst. Res. J.*, vol. 2, pp. 1–6, 2022.

[8] M. A. Iftikar and Y. Sibaroni, "Analisis Sentimen Twitter: Penanganan Pandemi Covid-19 Menggunakan Metode Hybrid Naïve Bayes, Decision Tree, dan Support Vector Machine," *e-Proceeding Eng.*, vol. 9, no. 3, pp. 1809–1826, 2022.

[9] Z. Kastrati, F. Dalipi, A. S. Imran, K. P. Nuci, and M. A. Wani, "Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study," *Appl. Sci.*, vol. 11, no. 9, 2021, doi: 10.3390/app11093986.

[10] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *J. SIMETRIS*, vol. 10, no. 2, pp. 681–686, 2019.

[11] N. A. Salsabila, "ANALISIS SENTIMEN PADA MEDIA SOSIAL TWITTER TERHADAP TOKOH GUS DUR MENGGUNAKAN METODE NAÏVE BAYES DAN SUPPORT VECTOR MACHINE (SVM)," UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH JAKARTA, 2022.

[12] Nugroho Arif Sudibyo, Ardymulya Iswardani, Kartika Sari, and Siti Suprihatiningsih, "Penerapan Data Mining Pada Jumlah Penduduk Miskin Di Indonesia," *J. Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat.*, vol. 1, no. 3, pp. 199–207, 2020, doi: 10.46306/lb.v1i3.42.

[13] M. O. Stitson, J. A. E. Weston, V. Vovk, and V. Vapnik, "Theory of Support Vector Machines," 1996.

[14] F. D. Pramakrisna, F. D. Adhinata, and N. A. F. Tanjung, "Aplikasi Klasifikasi SMS Berbasis Web Menggunakan Algoritma Logistic Regression," *Teknika*, vol. 11, no. 2, pp. 90–97, 2022, doi: 10.34148/teknika.v11i2.466.

[15] S. N. Listyarini and D. A. Anggoro, "Analisis Sentimen Pilkada di Tengah Pandemi Covid-19

Menggunakan Convolution Neural Network (CNN)," *J. Pendidik. dan Teknol. Indones.*, vol. 1, no. 7, pp. 261–268, 2021, doi: 10.52436/1.jpti.60.

[16] I. F. Yuliati and P. R. Sihombing, "Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 2, pp. 417–426, 2021, doi: 10.30812/matrik.v20i2.1174.

[17] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," *Proc. 2017 Int. Conf. Asian Lang. Process. IALP 2017*, vol. 2018-Janua, no. December, pp. 391–394, 2017, doi: 10.1109/IALP.2017.8300625.

[18] B. Kholifah, I. Thoib, N. Sururi, and N. D. Kurnia, "Analisis Sentimen Warganet Terhadap Isu Layanan Transportasi Online Berbasis InSet Lexicon Menggunakan Logistic Regression," *Kumpul. J. Ilmu Komput.*, vol. 11, no. 1, pp. 14–25, 2024.

Draft Jurnal Analisis Sentimen Transportasi Kereta Jakarta - Adriel Savero (20200803069).docx

ORIGINALITY REPORT

14%	9%	11%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	polgan.ac.id Internet Source	2%
2	Yik Hern Ong, Nurul Ashikin Abdul-Kadir, Nasrul Humaimi Mahmood, Muhammad Amin Abd Wahab et al. "Atrial Fibrillation Dynamic Features Recognition using 1-D Convolutional Neural Networks", 2023 IEEE International Biomedical Instrumentation and Technology Conference (IBITeC), 2023 Publication	1%
3	mdpi-res.com Internet Source	1%
4	Devendra Sahu, Ritesh Kumar Dewangan, Surendra Pal Singh Matharu. "Fault Diagnosis of Rolling Element Bearing with Operationally Developed Defects Using Various Convolutional Neural Networks", Journal of Failure Analysis and Prevention, 2024 Publication	1%

5	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2019", Springer Science and Business Media LLC, 2019 Publication	1 %
6	Khasyi Al Mukarram, M. Anang Mukhlas, Amalia Zahra. "Enhancing speech emotion recognition with deep learning using multi-feature stacking and data augmentation", Bulletin of Electrical Engineering and Informatics, 2024 Publication	1 %
7	Rajkumar Soundrapandiyan, Adhiyaman Manickam, Moulay Akhloufi, Yarlagadda Vishnu Srinivasa Murthy et al. "An Efficient COVID-19 Mortality Risk Prediction Model Using Deep Synthetic Minority Oversampling Technique and Convolution Neural Networks", BioMedInformatics, 2023 Publication	1 %
8	tunasbangsa.ac.id Internet Source	1 %
9	www.iieta.org Internet Source	1 %
10	article.nadiapub.com Internet Source	1 %

Lecture Notes in Computer Science, 2005.

11	Publication	<1%
12	Rizki Aulia Putra, Rice Novita, Tengku Khairil Ahsyar, Zarnelly. "Implementation of Classification Algorithm for Sentiment Analysis: Measuring App User Satisfaction", Teknika, 2024 Publication	<1%
13	bmrat.org Internet Source	<1%
14	journals.lww.com Internet Source	<1%
15	ojs.stmik-banjarbaru.ac.id Internet Source	<1%
16	journal.budiluhur.ac.id Internet Source	<1%
17	Amala Jayanthi M, Elizabeth Shanthi I. "Quest_SA: Preprocessing Method for Closed-Ended Questionnaires Using Sentiment Analysis through Polarity", Mobile Information Systems, 2022 Publication	<1%
18	Tasya Aisyah Amini, Kiki Setiawan. "Application of the Naive Bayes Algorithm in Twitter Sentiment Analysis of 2024 Vice Presidential Candidate Gibran Rakabuming	<1%

Raka using Rapidminer", International Journal Software Engineering and Computer Science (IJSECS), 2024
Publication

19	www.researchgate.net Internet Source	<1 %
20	Arbania Fitriani, Dominikus David Biondi Situmorang. "The differential item functioning (DIF) testing for the WOCC (Ways of Coping Checklist) instrument based on gender", Heliyon, 2023 Publication	<1 %
21	docplayer.net Internet Source	<1 %
22	econtlaw.nlu.edu.ua Internet Source	<1 %
23	repository.ub.ac.id Internet Source	<1 %
24	Sriani, Aidil Halim Lubis, Lia Putri Ashari Lubis. "Sentiment analysis on twitter about the death penalty using the support vector machine method", TEKNOSAINS : Jurnal Sains, Teknologi dan Informatika, 2024 Publication	<1 %
25	Yesi Novaria Kunang, Widya Putri Mentari. "Analysis of the Impact of Vectorization Methods on Machine Learning-Based	<1 %

Sentiment Analysis of Tweets Regarding Readiness for Offline Learning", JUITA : Jurnal Informatika, 2023
Publication

26	ejournal.uin-suska.ac.id Internet Source	<1 %
27	eurchembull.com Internet Source	<1 %
28	journal.universitasbumigora.ac.id Internet Source	<1 %
29	repository.mercubuana.ac.id Internet Source	<1 %
30	www.mdpi.com Internet Source	<1 %
31	www.utupub.fi Internet Source	<1 %
32	Ainul Yaqin, Ahlihi Masruro, Yoga Nawassarief. "Analysis Sentiment of President Vladimir Putin on Twitter Social Media Using Multinomial Algorithm and Bernoulli", 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2022 Publication	<1 %
33	Ahmad Hanif Asyhar, Yuniar Farida, Nurissaidah Ulinnuha, Dian Candra Rini	<1 %

Novitasari, Ahmad Zaenal Arifin. "Sentiment Analysis of Regional Head Candidate's Electability from the National Mass Media Perspective Using the Text Mining Algorithm", Advances in Science, Technology and Engineering Systems Journal, 2019

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		