SCIENCE AND INFORMATION TECHNOLOGY
SINTECH
JOURNAL

SINTECH JOURNAL | ISSN 2598-7305 | E-ISSN 2598-9642
Vol. 8 No 1 – April 2025 | https://s.id/sintechjournal
DOI: https://doi.org/10.31598
Publishing: Prahasta Publisher

Understanding Public Sentiment on Jakarta Public Transportation Using LSTM

Nadya Regina Djodjobo[1], Hasanul Fahmi[2]

[1,2] Master of Informatics, Faculty of Computer Science President University
Jl. Ki Hajar Dewantara, Cikarang, Indonesia
e-mail: nrnadyaregina@gmail.com[1], hasanul.fahmi@president.ac.id[2]

Received : January, 2025	Accepted : March, 2025	Published : April, 2025

Abstract

Traffic congestion is a major issue in Jakarta due to the number of vehicles exceeding road capacity. To address this, the Jakarta government has introduced various types of public transportation to drive commuters to shift to mass public transport from using private vehicles. However, the question remains: is the public transportation system optimal? Understanding public sentiment is crucial for optimizing the available transport options. This research aims to understand the public sentiment from social media using the LSTM method. By examining how people perceive the current state of mass public transportation, we can evaluate its effectiveness. From the approximately 1000 data from social media and google review that were processed in this research, the results show that using BiLSTM could reach an accuracy of 94% for datasets that only have Positive and Negative Sentiment and 74% for a dataset that consists of Neutral sentiment too. Ultimately, this study hopes to provide and deliver a summary of the public transportation system utilized by the NLP and LSTM algorithm, and also serve as an evaluation tool to improve services for encouraging a shift from private vehicles to public transport which can reduce traffic congestion in Jakarta.

Keywords: *sentiment analysis, mass public transportation, traffic congestion, NLP, LSTM*

Abstrak

Kemacetan lalu lintas adalah masalah utama di Jakarta yang terjadi akibat jumlah kendaraan yang telah melampaui kapasitas jalan. Untuk mengatasi hal tersebut, pemerintah Jakarta telah menghadirkan beragam jenis transportasi umum untuk mengajak masyarakat beralih ke transportasi umum massal dari penggunaan kendaraan pribadi. Namun, pertanyaan yang masih muncul adalah: apakah sistem transportasi umum tersebut telah optimal? Memahami sentimen publik menjadi hal yang krusial dalam upaya mengoptimalkan pilihan transportasi umum yang tersedia. Penelitian ini bertujuan untuk memahami sentimen publik yang diambil dari media sosial menggunakan metode LSTM. Dengan menganalisis persepsi masyarakat terhadap kondisi transportasi umum massal saat ini, efektivitas sistem tersebut dapat dievaluasi. Menggunakan data sekitar 1000 data dari social media dan google review yang telah diproses, memperoleh hasil penelitian menunjukkan bahwa penggunaan BiLSTM dapat mencapai tingkat akurasi sebesar 94% pada dataset yang hanya memiliki Sentimen Positif dan Negatif, serta 74% pada dataset yang juga mencakup Sentimen Netral. Pada akhirnya, penelitian ini bertujuan untuk dapat menyampaikan dan menyajikan ringkasan mengenai sistem transportasi umum melalui penerapan algoritma NLP dan LSTM, sekaligus menjadi alat evaluasi untuk meningkatkan layanan transportasi umum massal, yang dengan beralih dari penggunakan pribadi ke transportasi umum massal dapat mengurangi kemacetan lalu lintas di Jakarta.

Kata Kunci: *analisis sentimen, transportasi umum massal, kemacetan lalu lintas, NLP, LSTM*

1. INTRODUCTION

Traffic congestion is one of the main issues in Jakarta. Based on the TomTom Traffic Index 2023, Jakarta ranks 30th among the world's most congested cities. Friday evenings from 18:00 to 19:00 are notorious for severe congestion during rush hour, with an average travel time of approximately 33 minutes and 40 seconds to cover a distance of 10 kilometers. The peak congestion in Jakarta during 2023 occurred on March 9th, taking 30 minutes and 10 seconds to travel 10 kilometers, whereas the optimal travel time for this distance is 14 minutes[1]. The congestion primarily results from an overwhelming number of vehicles exceeding the capacity of the roads. According to BPS data in 2022 [2], motor vehicle numbers increased annually by an average of 3.13% from 2018 to 2022, with motorcycles growing by 3.52% annually. The urgency of congestion in Jakarta arises from its role as the economic and governmental center. The economic impact is significant, estimated at 65 trillion IDR based on Ministry of Transportation data for 2023[3]. Also, congestion reduces productivity and increases stress levels, as noted in various international journals.[4][5][6]

In Jakarta, as of 2022, there were 17,304,447 motorcycles and 3,766,059 passenger cars, according to BPS data[2]. The modal share of public transportation in Jakarta, as reported by the Jakarta Provincial Government in 2019, stands at only 21.7%. This means that out of a total of 26,424,851 trips made in the capital, 78.3% are conducted using private vehicles, totaling 20,689,139 trips. One of the primary reasons for this disparity is the inadequate public transportation coverage in many suburban areas surrounding Jakarta.

The government has focused on implementing mass public transportation systems to address congestion issues in Jakarta. The aim is for these systems to reduce traffic congestion by encouraging residents to shift from private vehicles. The Jakarta Provincial Government's Medium-Term Development Plan (Rencana Pembangunan Jangka Menengah Daerah or RPJMD) targets a 60% reduction in private vehicle usage in favor of public transport by 2030. However, the effectiveness of these efforts is questionable. Data indicates that Jakarta's population, exceeding 11 million as of December 2023 according to the Directorate General of Population and Civil Registration (Dukcapil), surpasses the current capacity of public transportation systems. Capacities include KRL serving 1 million passengers daily, MRT handling 180,000 passengers daily, LRT Jabodebek accommodating 130,000 passengers daily, and LRT Jakarta serving 1,300 passengers daily.

Despite efforts to promote public transportation, there has been an increase in its usage: TJ experienced a 43% rise in March compared to the previous year, while daily MRT passengers increased from 91,000 to 102,000. In January 2024 alone, MRT Jakarta carried 3,143,854 passengers, marking a 23.76% increase from January 2023, while LRT Jakarta saw 96,837 passengers, a 33.71% rise over the same period. Transjakarta recorded 30,934,491 passengers in January 2024, a 54.66% increase from January 2023, based on BPS data[7].

Despite these increases, public transportation's modal share remains relatively low at 21.7%, indicating continued reliance on private vehicles, which account for 78.3% of trips in the capital, according to Jakarta Provincial Government data from 2019. This disparity underscores the need for comprehensive sentiment analysis to understand public perception and satisfaction levels towards Jakarta's public transportation systems. Based on recent studies[8] and [9], sentiment analysis is essential for evaluating customer satisfaction with public transportation services in Jakarta. TransJakarta, for instance, has implemented systems to collect passenger feedback, revealing mixed sentiments despite efforts to provide quality facilities. For example, while the service is generally appreciated, concerns persist about the adequacy of bus stop conditions, particularly on Corridor 7 Transjakarta from Kampung Rambutan to Kampung Melayu. These findings illustrate the importance of ongoing sentiment analysis to identify areas for improvement and enhance overall service quality to meet the increasing demand for public transportation in Jakarta.

In 2024, the transformer models inspired the improvement of sentiment analysis and the rating prediction of app reviews. Particularly, the research by Gökberk Eser, employing models like BERT, DistilBERT, RoBERTa and XLM-

RoBERTa over Spotify app reviews in Google Play Store showed conclusive outcomes. DistilBERT performed better among all, the accuracy and recall reached 71.68%, while XLM-RoBERTa achieved the best balance with an F1 score of 69.24%, in predicting the Spotify app rating [10].

Another significant development in 2024 was the creation of hybrid models for emotion classification and sentiment analysis in the Indonesian language. This research by Ahmadian et al. combined BERT or IndoBERT with BiLSTM, BiGRU, and attention models, applied to an Indonesian language dataset. The models achieved 93% accuracy in sentiment analysis and reached 78% for the emotion classification on the IndoNLU (Indonesian Natural Language Understanding) benchmark, representing a significant advancement in the field [11].

In the same year, a novel approach for sentiment analysis in the Bangla language was introduced using the BangDSA dataset, and a new feature metric, skipBangla-BERT. The study by Islam et al. employed 21 different hybrid feature extraction methods, including BOW, N-gram, TF-IDF, TF-IDF-ICF, Word2Vec, FastText, GloVe, and Bangla-BERT, with CBOW and Skip Gram mechanisms. The skipBangla-BERT method outperformed all other techniques across machine learning, ensemble learning, and deep learning approaches. The hybrid CNN-BiLSTM model achieved the highest accuracy, with 90.24% in 15 categories and 95.71% in 3 categories, setting a new benchmark in Bangla sentiment analysis [12].

From previous studies, the SVM method is commonly chosen due to its applicability to smaller datasets and limited computational resources. However, SVM's weakness lies in its limited ability to capture contextual nuances, as SVM treats text as independent feature vectors and does not capture sequence information. In contrast, this research adopts LSTM, designed specifically for sequential data, allowing it to effectively capture long-range dependencies. LSTM utilizes a gating mechanism, the forget, input, and output gates, that regulate the flow of information, and allow it to retain relevant past information over long sequences. This capability is crucial for accurately interpreting sentiment within specific contexts. One limitation of this research is the use of LSTM instead of Transformers like BERT due to the limited dataset size. Given this constraint, LSTM is expected to outperform SVM. Additionally, previous research has extensively utilized LSTM in various sentiment analysis applications, including online food reviews, airline feedback, emotional text classification, product reviews, and movie critics. In this study, LSTM is specifically employed to address a particular issue in Jakarta—analyzing sentiment related to mass public transportation. The objective is to mitigate a specific problem, traffic congestion. This research underscores the utility of NLP in addressing targeted societal challenges.

Private vehicles are the main contributors to Jakarta's congestion, but the question remains: is all the available mass public transportation offered by the government already optimal in reaching and serving commuters in Jakarta? Furthermore, how can NLP techniques be utilized to analyze public sentiments toward Jakarta's mass public transportation system? Therefore, this research objectives are as follows:

1. To understand public sentiments regarding the optimality of mass public transportation in Jakarta using NLP.
2. To identify specific aspects that the users like or dislike about the current services of mass public transportation for continuous evaluation and improvement based on sentiment analysis.
3. To determine the combination of parameters needed to achieve the best accuracy with the LSTM algorithm.

The following section provides a literature review for this journal.

A. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a machine learning technology that allows computers to understand, analyze, and generate human language. It encompasses a range of tasks, including language translation, sentiment analysis, and speech recognition, all aimed at bridging the communication gap between machines and humans. Leveraging linguistic and statistical techniques, NLP processes and extracts meaningful information from both textual and spoken data. Based on a study by Khurana et al., NLP is classified into two parts, which are Natural Language Understanding and Natural Language Generation. The NLU allows

machines to analyze natural language by extracting the concepts, emotions, entities, and keywords. While NLG involves creating meaningful phrases, sentences, and paragraphs from an internal representation [13].

B. Sentiment Analysis

Sentiment analysis, a specialized application of NLP, is dedicated to interpreting emotions and opinions expressed in natural language. Leveraging NLP techniques, sentiment analysis is essential in automating the comprehension of sentiments within various textual data, offering insights into the emotional nuances of human expressions. In a specific study by Parikh et al., an NLP-based approach is adopted for sentiment analysis on data from Twitter, involving crucial steps such as preprocessing, feature extraction, and classification. During feature extraction, sentence tokenization is applied, and the training and testing sets are input into a machine learning algorithm for the sentiment analysis [14].

C. Long Short-Term Memory (LSTM)

In recent years, various approaches to sentiment analysis have been explored, highlighting the strengths of different machine learning models and techniques. For instance, a study in 2017 compared several models including SVM with TF-IDF, Multinomial Naive Bayes (MNB) with TF-IDF, and LSTM networks with different embeddings like Word2vec, self-initialized embeddings, and GloVe. The research utilized the Amazon Fine Food Reviews Dataset and the Yelp Challenge Dataset, finding that LSTM techniques utilizing GloVe and Word2vec embeddings outperformed other methods, indicating the significant impact of the distribution of ratings in the data on model performance [15].

Building on these findings, another study in 2022 focused on emotional text classification using TF-IDF and LSTM models. This study by Alfarizi et al. used a dataset of 18,000 instances divided into training and test sets across six emotional classes, including anger, fear, joy, love, sadness, and surprise. The results demonstrated that the LSTM method achieved a remarkable 97.50% accuracy in emotion classification, significantly outperforming the LinearSVC method, which achieved an accuracy of 89%. This highlights the efficacy of LSTM in capturing emotional nuances in text data [16].

Further research in 2019 by Xu et al. reinforced the superiority of advanced neural network architectures for sentiment analysis. This study compared BiLSTM with traditional models such as RNN, LSTM, CNN, and NB revealing that the BiLSTM-based sentiment analysis method achieved higher precision, recall, and F1 score. These findings underscore the advantages of using bidirectional LSTM networks in accurately predicting sentiment by effectively capturing the context and dependencies in text sequences, thus providing a more robust framework for sentiment analysis [17].

D. Previous Research

1) Implementation of Sentiment Analysis

In a specific application, a research study implemented Word2Vec and LSTM for sentiment classification in hotel reviews. Through an evaluation of 144 parameter combinations, the study identified the most effective scheme, achieving a mean accuracy of 85.96%. This optimal configuration incorporates Skip-Gram as the Word2Vec architecture, employs Hierarchical Softmax for evaluation, and adopts a vector dimension of 300 [18]. The emotion of users in a review is important, research conducted that removed the profanity data in a review shows by judging it as noise data received an accuracy dropped by 2% then using the profanity data in a sentiment classification for review data [19]. A study employing lexicon-based sentiment analysis, specifically the VADER model, effectively forecasted the outcome of the US presidential election by analyzing public sentiments on Twitter [20].

In a Jakarta-based study, the government's use of Twitter as a platform for community interaction was explored, with a focus on sentiment analysis to gauge public opinion on COVID-19 vaccination policies. The research collected and analyzed 1658 tweets directed at the DKI Jakarta Provincial Government's official Twitter account. Two classification methods, naive Bayes and k-NN, were employed using TF-IDF Vectorizer for word weighting. In the comparison of classification methods, the naive Bayes approach emerged with the highest accuracy [21].

2) Implementation of Sentiment Analysis in Public Transportation in Jakarta

The study utilized the SVM method to classify sentiment in Twitter texts related to MRT, LRT,

and Transjakarta transportation. The data source was Twitter, and the SVM method demonstrated an impressive ability to classify the negative and positive sentiments. The classification accuracy achieved was 91.89%, with the distribution of sentiments being 79.2% positive and 20.8% negative [22]. This research employed the Naive Bayes method to analyze community opinions on Transjakarta buses based on Twitter data. The system's accuracy was found to be 73%, which is relatively low. The main reason for this low accuracy relies on the limited training data, as only 62.5% of the total 50 data points, with the remaining 30 data points reserved for testing [23].

A study utilized Google Reviews to create a sentiment analysis model and then evaluate the performance of NB, SVM, and KNN classifiers. The dataset comprised approximately 1,453 reviews. Among the classifiers, the SVM model with an RBF kernel demonstrated the highest performance, achieving an average accuracy of 82%, surpassing the KNN classifier (k=12) at 79% and the Multinomial Naive Bayes algorithm at 75%. Analysis revealed that over the past three years, 64.28% of visitors perceived the infrastructure of Manggarai Station negatively, 27.53% positively, and 8.19% neutrally [24]. This analysis used the SVM method on Twitter data collected via Tweepy. The results showed a high accuracy rate of 92.00%, with precision at 91.00% and recall at 92.00%, based on 2123 data points. The findings indicated that the majority of Jakartans held a negative impression of the bus rapid transit services, with many customers expressing disappointment with the services provided [25].

Several studies regarding the sentiment analysis of the Commuter Line in Indonesia or KRL through data from Twitter, about public opinions particularly complaints, regarding the Commuter Line service. Comparing Machine Learning models such as MNB, RF, and SVM, the research showcases that SVM achieves the highest accuracy at 85% [26]. A study employing sentiment analysis with the Support Vector Machine (SVM) algorithm categorizes public sentiment into positive and negative. Achieving a relatively stable and commendable accuracy score of 0.711 through 5-fold-cross-validation, the research unveils topics with positive sentiments, such as the convenience of free-charged JakLingko public transportation, and negative sentiments, including complaints about JakLingko cards and the perceived low service quality of public transportation [27].

2. RESEARCH METHOD

A. Data Collection

In this research, the dataset was collected manually using online comment extraction from the Social Media platforms and Google Review including Twitter and Instagram comments. Google Collab was used as a tool to develop and train the sentiment analysis model, using Python as the Programming Language and TensorFlow/Keras as the deep learning framework. The dataset contains approximately 800 and 1000 entries, since there are 2 datasets. Figure 1 below is a sample of the dataset.

	date	username	comment	category	source	label	aspect
2	21/06/24 07:34:3	bagus_putra182	Tambah lagi min rangkaianya udh sumpek keretanya karna ga balance antara penumpang sama keretanya ðŸ¥´	KRL	IG	Negative	Comfort
3	21/06/24 14:05:0	hendrysetiadi	@commuterline ya.. bukan cuma saya aja yg merasakan keterlambatan kereta yg menuju tn abang.. tolong disesuaikan schedule nya lagi. Biasa emg tiap jumat selalu banyakan headway nya jauh dibanding schedule2 sebelum jam 17.30. Malah sebelum 17.30 selang antar 1 kereta & kereta berikutnya itu singkat cuma kisaran 2-3 menit. Bener2 saya mohon agar diperbanyak kereta yg jurusan duri tn abang di jam2 pulang kerja.. jangan cuma pentingin yg ke bekasi cikarang doang donk min.. @budikaryas @ditjenperkeretaapian @ditjen_hubdat @kemenhub151	KRL	IG	Negative	Punctuality
4	21/06/24 14:49:3	tegarpermadaa	Minn tambah rangkaian dong yg arah bogorr udah sumpek bgt 8sf mulu klo pulang kantor	KRL	IG	Negative	Coverage
5	22/06/24 05:28:1	desper.ate.in	Min untuk KRL tujuan Tangerang Duri PP itu nggak ada tambahan lagi min soalnya jedanya cukup jauh sekitar 30 menit pada waktu siang hari	KRL	IG	Negative	Punctuality
6	22/06/24 16:14:1	akhmadhafishhh._	min tolong kereta tambahan untuk stasiun juanda , penumpang bludag	KRL	IG	Negative	Comfort
7	23/06/24 03:25:3	noerday71	Baru saja terjadi bbrp menit yg lalu , masinis menutup pintu kereta walau sudah di kasih aba2 bahwa penumpang blm naek hampir membuat celaka para penumpang yg mau naek coba di perhatikan , kejadian di kp bandan kereta yg arah ke kota mohon perhatiannya buat para masinis	KRL	IG	Negative	Comfort

Figure 1. Sample Dataset

These entries encompass user feedback on various aspects of mass public transportation services, such as cleanliness, punctuality, safety, infrastructure, affordability, comfort, coverage, and technology integration. Each entry includes the review date, the usernames, the reviews, the transportation category, such as Transjakarta, MRT, LRT, KRL, and Mikrotrans, and the sentiment label, such as positive and negative for one dataset, and additional neutral sentiment for the other dataset. The dataset was **labeled manually** to prepare it for training with sentiment labels **positive labels, negative labels, or neutral labels** based on the overall sentiment expressed in each review. All labeled data was then saved as Processed Dataset, which was later used for training the model. The flow of the data collection will be shown in Figure 2.

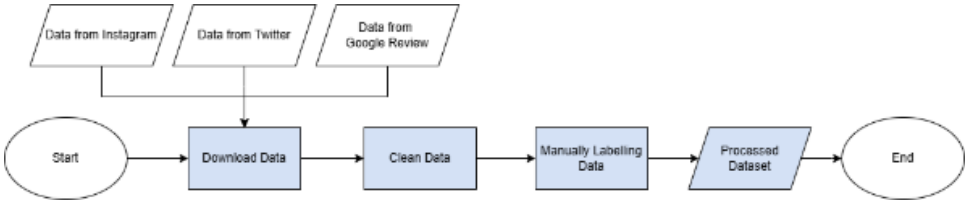

Figure 2. Data Collection

B. *Data Preprocessing*

Several steps are conducted to preprocess the data to become more structured, standardized, and free from noise to be analyzed, listed below: Loading the Processed Dataset as the input, then lowercasing, by converting all text to lowercase letters; removing noise, by removing numbers, symbols, and whitespaces; removing stop words by removing common words that occur frequently in the text but contribute to the analysis; tokenization, by splitting the text into individual tokens or words; and lastly stemming, by reducing words to their root form by removing suffixes.

Below is the following flowchart of this research, shown in Figure 3.

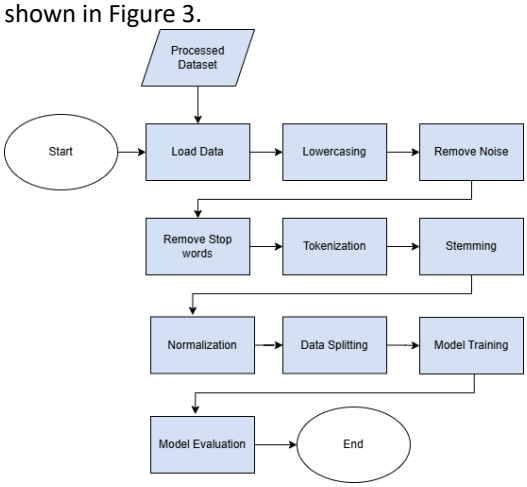

Figure 3. Flow Chart

C. *Data Splitting*

The dataset was split into two subsets: 80% for training the machine learning model and 20% for validation. The training data to optimize the model's parameters, meanwhile, the validation data to evaluate the model's performance metrics such as accuracy or loss, and to fine-tune hyperparameters to achieve optimal results. The data splitting method helps to prevent overfitting.

D. *Data Training*

The LSTM approach will be employed to train the data. LSTM, a specialized type of Recurrent Neural Network (RNN), incorporates additional memory cells designed to manage the long-term dependencies and address the vanishing gradient issue. A sequential model is how the model's layer is arranged, which is sequentially from one layer to the next, forming a linear stack of layers.

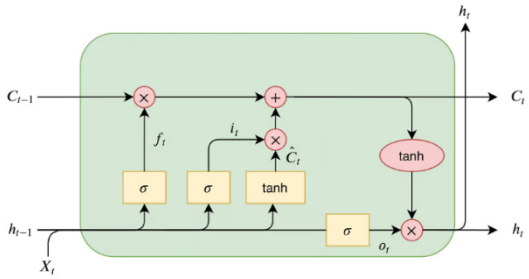

Figure 4. LSTM Model Architecture [28]

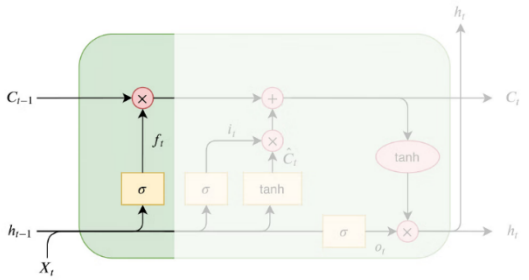

Figure 5. Forget Gate

As shown in Figure 4, the core components of the LSTM model architecture consist of the cell state, forget gate, input gate, and output gate. The cell state serves as the carrier of essential information that traverses through all gates from one cell to the next. As depicted in Figure 5, the Forget Gate stage utilizes the LSTM neural network to identify which elements of the cell state (long-term memory) are relevant, based on the previous hidden state and the new input data. A sigmoid activation function generates a vector where each element ranges between 0 and 1.

$$\boldsymbol{f_t = \sigma\,(W_f \,.\, [h_{t-1}\,, X_t\,] + b_f\,)}$$

Next is the Input Gate, in Figure 6, which decides which of the new input information goes into the long-term memory or the cell state of the network, based on the previous hidden state

and the current input data. With a sigmoid activation function, it filters valuable components to identify important ingredients in the new memory vector.

$$\boldsymbol{i_t} = \sigma\,(\boldsymbol{W_i}\,.\,[\boldsymbol{h_{t-1}}\,,\boldsymbol{X_t}\,] + \boldsymbol{b_i}\,)$$

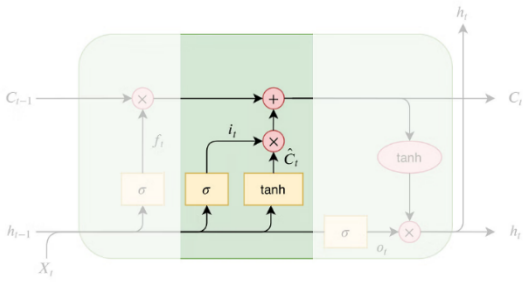

Figure 6. Input Gate

The new memory is activated by this tanh activation function and has been trained to generate the "new memory update vector" by recombining the previous hidden state and the current input data.

$$\hat{\mathrm{C}}_{\boldsymbol{t}} = \tanh\,(\boldsymbol{W_C}\,.\,[\boldsymbol{h_{t-1}}\,,\boldsymbol{X_t}\,] + \boldsymbol{b_C}\,)$$

The final update to the cell state is achieved by combining the "new memory update" with the "input gate" filter. The input gate filter, through pointwise multiplication, controls the output of the new memory update. This ensures that only the most significant parts of the new memory update contribute to the cell state update.

$$\boldsymbol{C_t} = \boldsymbol{i_t}\,.\,\hat{\mathrm{C}}_{\boldsymbol{t}} + \boldsymbol{f_t}\,.\,\hat{\mathrm{C}}_{\boldsymbol{t-1}}$$

The output gate, as shown in Figure 7, is responsible for making the final decision regarding the information to be output as the new hidden state. Specifically, it identifies which parts of a filtered version of the updated cell state are significant enough for output. This process utilizes the network's sigmoid activation function. The output gate takes the new input data and the previous hidden state as inputs, applying sigmoid activation to generate a gating signal. The final hidden state is then produced by pointwise multiplying the output of the sigmoid-activated output gate with the updated cell state, which has been processed through a tanh activation function.

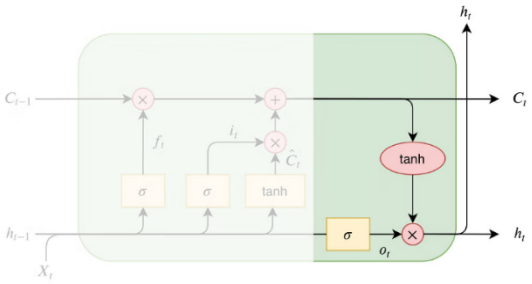

Figure 7. Output Gate

$$\boldsymbol{o_t} = \sigma\,(\boldsymbol{W_o}.\,[\boldsymbol{h_{t-1}}\,,\boldsymbol{X_t}\,] + \boldsymbol{b_o}\,)$$

$$\boldsymbol{h_t} = \boldsymbol{o_t} \times \tanh\,(\boldsymbol{C_t}\,)$$

E. Validation and Evaluation

Standard evaluation metrics such as the Confusion Matrix, Accuracy, Precision, Recall, F1 Score, ROC Curve, and AUC Score were utilized to evaluate the model's performance.

The confusion matrix is a table that represent the performance of a classification model that compares the predicted to the true labels, summarizing its performance through four possible outcomes:

a) True Positive (TP): The number of correctly predicted positive class samples.
b) True Negative (TN): The number of correctly predicted negative class samples.
c) False Positive (FP): The number of negative class samples incorrectly predicted as positive.
d) False Negative (FN): The number of positive class samples incorrectly predicted as negative.

Some commonly used metrics to measure the performance of the classification model, such as:

a. Accuracy, the number of correct predictions divided by the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

b. Precision, the ratio of true positives and total positives predicted.

$$\text{Precision} = \frac{TP}{TP + FP}$$

c. Recall, the ratio of true positives to all the actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

d. F1-Score, the harmonic mean of precision and recall.

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

ROC or Receiver Operating Characteristic curves, shown in Figure 8, plot the true positive rate (TPR) against the false positive rate (FPR) to visualize the performance of a binary classifier across various decision thresholds. The area under the ROC curve, referred to as the AUC score, shown in Figure 9, quantifies the model's ability to differentiate between positive and negative classes over all possible thresholds. An AUC score of 0.5 indicates performance equivalent to random guessing, while a score of 1 represents perfect classification performance.

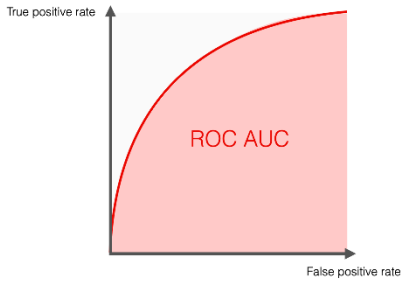

Figure 8. ROC Curves [29]

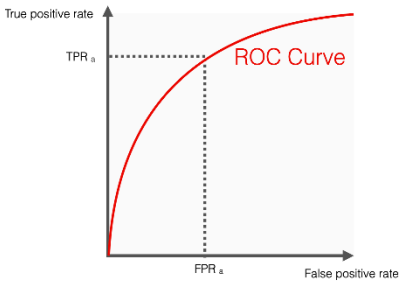

Figure 9. AUC Score

3. RESULT AND DISCUSSION

This research utilized two distinct datasets for sentiment analysis. The first dataset included labels for Positive and Negative sentiments, with a proportion was 47.7% Positive and 52.3% Negative as shown in Figure 10, while the second dataset included an additional Neutral sentiments, with proportions of 35% Positive, 39% Negative, and 26% Neutral, as shown in Figure 11.

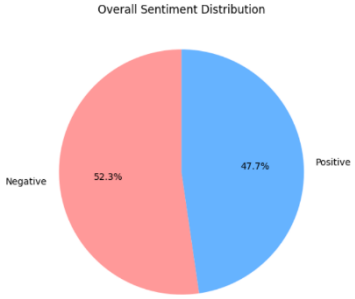

Figure 10. Sentiment Distribution on Dataset 1

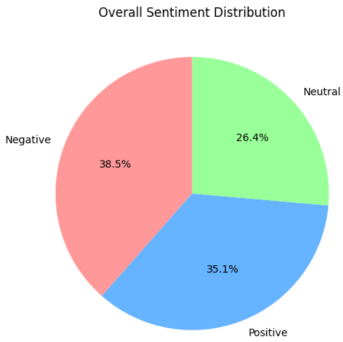

Figure 11. Sentiment Distribution on Dataset 2

Despite the differences in datasets, the same model architecture was applied to both. The model began with an embedding layer that converted input tokens into dense vector representations, continued by two Bidirectional LSTM layers that captured both past and future context. Each LSTM layer was paired with a dropout layer, using a rate of 0.2. A dense layer with 64 units and ReLU activation processed the output before a final softmax layer handled multi-class sentiment classification. With a learning rate of 0.0005 and 0.001, categorical crossentropy as the loss function, the model compiled using the Adam optimizer. Early stopping to avoid overfitting, halting training if validation loss did not improve over three epochs. After 15 epochs, with a batch size of 64 and a 20% validation split, the model achieved 94% accuracy and a loss of 0.24 on the test of the first dataset. While the result of the second dataset, which contains a Neutral label, reached 74% accuracy and 0.68 loss on the test set. All accuracy results were obtained after several parameter tuning iterations.

1) Model Performance

This model employs BiLSTM architecture with an embedding dimension of 128 and features two LSTM layers with 128 and 64 units. The model consists of 8 layers, and each LSTM layer is

followed by a dropout rate of 0.5. It uses a learning rate of 0.001 for balanced training, and a batch size of 64 for efficiency, and is trained over 15 epochs with early stopping. The activation functions include ReLU for the dense layers and softmax for multi-class classification, with the Adam optimizer. The loss function used is categorical crossentropy, suitable for multi-class classification tasks. The result of this model is shown in Figure 12. The training accuracy reached 98% and the **test accuracy was 94%.** With the same architecture model for the second dataset, the training accuracy reached 93% but the test accuracy was only 74% with a loss of 0.67which is relatively high, as shown in Figure 13.

```
Epoch 1/15
9/9 ━━━━━━━━━━━━━━━━━━━━ 17s 825ms/step - accuracy: 0.5086 - loss: 0.6943 - val_accuracy: 0.5441 - val_loss: 0.6871
Epoch 2/15
9/9 ━━━━━━━━━━━━━━━━━━━━ 9s 1s/step - accuracy: 0.5399 - loss: 0.6815 - val_accuracy: 0.6250 - val_loss: 0.6536
Epoch 3/15
9/9 ━━━━━━━━━━━━━━━━━━━━ 7s 725ms/step - accuracy: 0.6762 - loss: 0.6367 - val_accuracy: 0.8309 - val_loss: 0.4921
Epoch 4/15
9/9 ━━━━━━━━━━━━━━━━━━━━ 9s 1s/step - accuracy: 0.9142 - loss: 0.3547 - val_accuracy: 0.8750 - val_loss: 0.3668
Epoch 5/15
9/9 ━━━━━━━━━━━━━━━━━━━━ 7s 732ms/step - accuracy: 0.9443 - loss: 0.1780 - val_accuracy: 0.8676 - val_loss: 0.4793
Epoch 6/15
9/9 ━━━━━━━━━━━━━━━━━━━━ 10s 742ms/step - accuracy: 0.9726 - loss: 0.1056 - val_accuracy: 0.8750 - val_loss: 0.4878
Epoch 7/15
9/9 ━━━━━━━━━━━━━━━━━━━━ 9s 1s/step - accuracy: 0.9871 - loss: 0.0713 - val_accuracy: 0.8529 - val_loss: 0.4624
6/6 ━━━━━━━━━━━━━━━━━━━━ 1s 142ms/step - accuracy: 0.9262 - loss: 0.2470
Test Accuracy: 0.94
```

Figure 12. Accuracy from Dataset 1

```
Epoch 5/15
12/12 ━━━━━━━━━━━━━━━━━━━━ 11s 930ms/step - accuracy: 0.6421 - loss: 0.8952 - val_accuracy: 0.6141 - val_loss: 0.7709
Epoch 6/15
12/12 ━━━━━━━━━━━━━━━━━━━━ 18s 676ms/step - accuracy: 0.7011 - loss: 0.6950 - val_accuracy: 0.6141 - val_loss: 0.7893
Epoch 7/15
12/12 ━━━━━━━━━━━━━━━━━━━━ 11s 901ms/step - accuracy: 0.6751 - loss: 0.6396 - val_accuracy: 0.6250 - val_loss: 0.7647
Epoch 8/15
12/12 ━━━━━━━━━━━━━━━━━━━━ 19s 721ms/step - accuracy: 0.7426 - loss: 0.5999 - val_accuracy: 0.6250 - val_loss: 0.7316
Epoch 9/15
12/12 ━━━━━━━━━━━━━━━━━━━━ 10s 890ms/step - accuracy: 0.7719 - loss: 0.5332 - val_accuracy: 0.6739 - val_loss: 0.7266
Epoch 10/15
12/12 ━━━━━━━━━━━━━━━━━━━━ 18s 690ms/step - accuracy: 0.8340 - loss: 0.4492 - val_accuracy: 0.6685 - val_loss: 0.7440
Epoch 11/15
12/12 ━━━━━━━━━━━━━━━━━━━━ 12s 803ms/step - accuracy: 0.9083 - loss: 0.2977 - val_accuracy: 0.6957 - val_loss: 0.8587
Epoch 12/15
12/12 ━━━━━━━━━━━━━━━━━━━━ 10s 874ms/step - accuracy: 0.9399 - loss: 0.1991 - val_accuracy: 0.6685 - val_loss: 1.0020
8/8 ━━━━━━━━━━━━━━━━━━━━ 2s 229ms/step - accuracy: 0.7189 - loss: 0.6790
Test Accuracy: 0.74
```

Figure 13. Accuracy from Dataset 2

The accuracy results from both datasets are quite significant, possibly influenced by the imbalanced neutral data. Neutral sentiments tend to be ambiguous and overlapping, making it difficult for the model to differentiate them from positive or negative sentiments, and the small size of the neutral class, prevents the model from learning optimally. Balanced data is crucial to ensure that the model is not biased toward the majority class. While undersampling and class weighting techniques were applied in this research, they were not sufficient to fully balanced data. A potential solution to improve this could be combining SMOTE with class weighting.

The previous model also provided valuable insights into the aspects associated with each predicted sentiment, helping us identify areas that require improvement and those that should be maintained. For instance, as shown in Figures 14 and 15, both datasets highlight that the aspects of coverage, punctuality, safety, and technology integration need attention due to the high volume of negative comments.

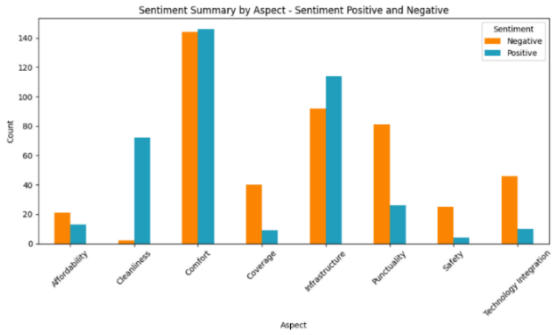

Figure 14. Aspect from each sentiment from Dataset 1

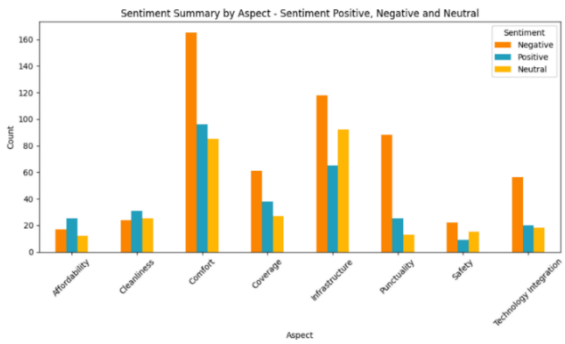

Figure 15. Aspect from each sentiment from Dataset 2

2) Evaluation

From Figure 16, out of 94 actual Negative cases, 86 were correctly predicted, with only 8 misclassified as Positive. Similarly, the model correctly identified 69 out of 75 Positive cases, with only 6 being wrongly classified as Negative. This indicates that the model is highly effective in separating these two sentiments.

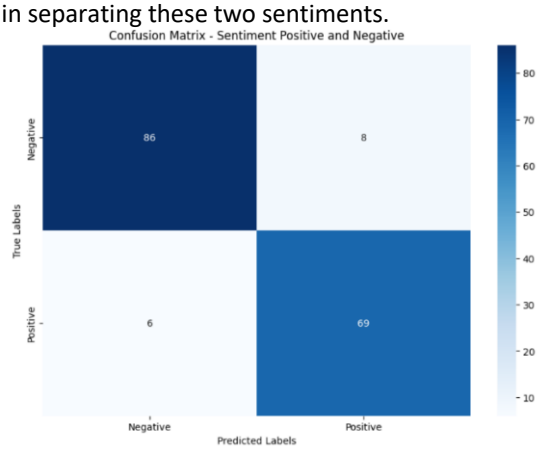

Figure 16. Confusion Matrix from Dataset 1

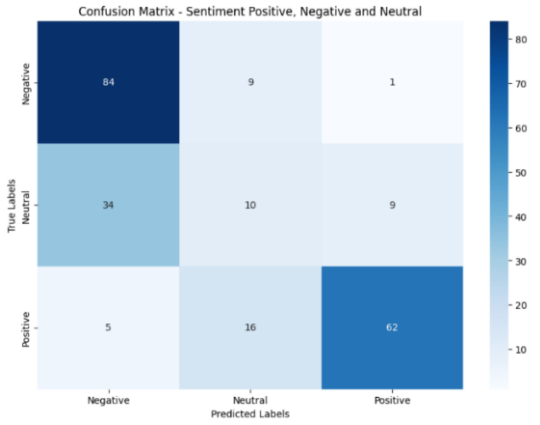

Figure 17. Confusion Matrix from Dataset 2

At the same time, in Figure 17, the confusion matrix reveals that the model performs well when predicting Negative and Positive sentiments but struggles with Neutral classifications. Out of 94 actual Negative cases, 84 were correctly predicted, while only 10 out of 53 Neutral cases were classified correctly, with many being confused for either Negative (34 cases) or Positive (9 cases). Similarly, Positive sentiments were generally well predicted, with 62 out of 83 correctly identified, but a small portion was misclassified as either Neutral or Negative.

The dataset that only consists of 2 classes (positive and negative) as shown in Figure 18, an AUC score of 0.97 for both Class 0 and Class 1 means the model is doing an excellent job at telling the difference between the two classes. A score of 0.97 suggests that the model is nearly perfect at identifying whether a sample belongs to Class 0 or Class 1, indicating that it's making accurate predictions almost all of the time. While on the other hand in Figure 19, the second dataset that has Neutral sentiment, for Class 0, with a score of 0.88, the model could identify it correctly most of the time. For Class 1, with a score of 0.72, the performance is lower, indicating the model struggles more to differentiate this class from others. However, for Class 2 with a score 0.95, the model has almost perfect accuracy in identifying this class.

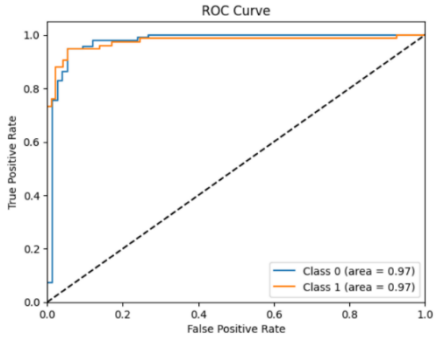

Figure 18. ROC Curve from Dataset 1

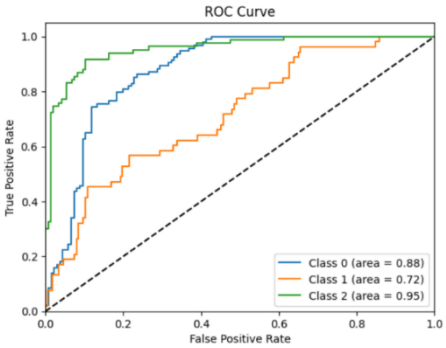

Figure 19. ROC Curve from Dataset 2

The evaluation of the SVM model showed a test accuracy of 55%, or it correctly predicted sentiment for just over half of the test samples. The results varied significantly across different sentiment classes. For the Negative class, the model achieved a precision of 53% and a recall of 76%. The Positive class precision of 68% and a recall of 55%. However, the model struggled with the Neutral class, managing only 36% precision and a mere 19% recall, which highlights the difficulty in accurately identifying

neutral sentiments, as shown in Figure 20. In contrast shown in Figure 21, when applying to the other dataset which without Neutral sentiment, the evaluation showed a much higher test accuracy of 78%, with the Negative class achieving a precision of 78% and a recall of 85%, and the Positive class recording a precision of 79% and a recall of 69%.

Test Accuracy: 0.55

	precision	recall	f1-score	support
Negative	0.53	0.76	0.62	94
Neutral	0.36	0.19	0.25	53
Positive	0.68	0.55	0.61	83
accuracy			0.55	230
macro avg	0.52	0.50	0.49	230
weighted avg	0.54	0.55	0.53	230

Figure 20. Accuracy of SVM Model from Dataset 2

Test Accuracy: 0.78

	precision	recall	f1-score	supp
Negative	0.78	0.85	0.81	
Positive	0.79	0.69	0.74	
accuracy			0.78	
macro avg	0.78	0.77	0.77	
weighted avg	0.78	0.78	0.78	

Figure 21. Accuracy of SVM Model from Dataset 1

The BiLSTM model **outperforms** the SVM model in terms of accuracy on both datasets, indicating that it is more effective at capturing the nuances of sentiment in the text, as shown in Figure 22.

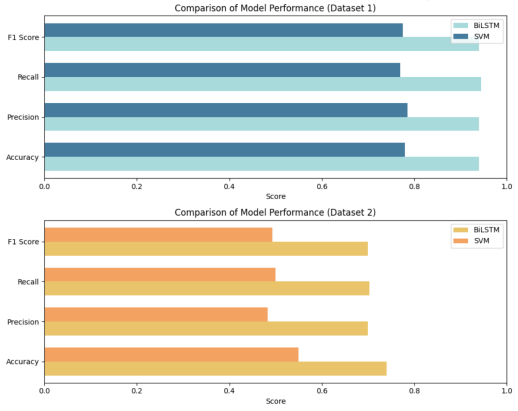

Figure 22. Comparison of Model Performance

The BiLSTM can process sequences in both directions—forward and backward—allowing it to better understand the context and relationships between words. This ability to capture long-term dependencies helps the BiLSTM model make **more accurate** predictions, especially in complex datasets with varying sentiment expressions. On the other hand, SVM (Support Vector Machine) models, while simpler and effective for linear classification problems, **struggle** with the complexity of natural language processing tasks. SVMs are not as good at understanding sequential dependencies, making them less capable of handling subtle patterns in text. Overall, the higher accuracy of the BiLSTM model shows that it's more suitable for tasks involving nuanced language understanding, like sentiment analysis, especially when compared to traditional machine learning models like SVM.

4. CONCLUSION

In this study, around 1000 entries from platforms like Instagram, Twitter, and Google Reviews were compiled into a single dataset to assess public sentiment on various forms of mass transportation in Jakarta, including Transjakarta, MRT, LRT, KRL, and Mikrotrans to analyze the efficiency of these transportation systems through sentiment analysis, with the dataset **labeled as Positive, Negative or Neutral**. In the first dataset which only has Positive and Negative sentiment label, several Long Short-Term Memory (LSTM) model variations were tested, and the best model—a BiLSTM with an output dimension of 128 and learning rate of 0.001—achieved a high testing accuracy of 94%, with a low loss of 0.24. This performance significantly outperformed a SVM model applied to the same dataset, which achieved a maximum accuracy of 78% using a polynomial kernel.

A second dataset, which included an additional Neutral sentiment label, was also analyzed. The same BiLSTM configuration with early stopping over 20 epochs and reached a lower accuracy of 74% and a loss of 0.67. The lower performance, particularly in the Neutral category, can be attributed to the unbalanced nature of the data, with fewer Neutral entries compared to Positive and Negative ones. The SVM model on this dataset performed even worse, with an accuracy of 55%. These results highlight that while BiLSTM models are effective for sentiment analysis in transportation-related datasets, unbalanced data, especially in categories like Neutral, can significantly impact the model's ability to accurately classify sentiment.

Based on the sample dataset and the insights provided by the model, it is evident that the existing mass public transportation in Jakarta still requires improvements as indicated by the high volume of negative comments, especially in

the aspect of coverage, punctuality, safety, and technology integration.

In future work, the focus will be on enhancing the model's performance for neutral sentiments, ensuring more accurate classification. Exploring Transformer models like BERT could be an option to determine whether the model architectures improve sentiment analysis, particularly for neutral sentiments. Also, a larger dataset is needed to enhance the model's generalization. Training on various data from other cities and platforms could further improve the accuracy making the model more practical for real-world implementation.. Additionally, efforts will be directed towards predicting specific aspects related to overall sentiment, allowing for a clearer understanding of which areas need improvement, and which should be maintained. Finally, there's a goal to refine the handling of usernames, enabling richer user analysis that could involve detecting biases or personalizing experiences based on user behavior and sentiment.

REFERENCES

[1] "Jakarta traffic report | TomTom Traffic Index." Accessed: Jul. 12, 2024. [Online]. Available: https://www.tomtom.com/traffic-index/jakarta-traffic/

[2] BPS Provinsi DKI Jakarta, "Transportation Statistics of DKI Jakarta 2022," 2023.

[3] Devi Puspitasari, "Kemenhub Ungkap Kerugian Akibat Kemacetan Jakarta Capai Rp 65 Triliun," DetikNews. Accessed: Jul. 12, 2024. [Online]. Available: https://news.detik.com/berita/d-6795414/kemenhub-ungkap-kerugian-akibat-kemacetan-jakarta-capai-rp-65-triliun

[4] M. Kamruzzaman and Z. F. Rumpa, "The Effect of Traffic Congestion on Employee Productivity in Dhaka Bangladesh," *The International Journal of Business & Management*, vol. 7, no. 5, May 2019, doi: 10.24940/THEIJBM/2019/V7/I5/BM1811-023.

[5] T. D. Weerasinghe, I. Karunarathna, and C. Subhashini, "Effect of Road Traffic Congestion on Stress at Work: Evidence from the Employees Working in Metropolitan Areas of Colombo, Sri Lanka," 2021, Accessed: Jul. 12, 2024. [Online]. Available: https://www.researchgate.net/publication/351281226_Effect_of_Road_Traffic_Congestion_on_Stress_at_Work_Evidence_from_the_Employees_Working_in_Metropolitan_Areas_of_Colombo_Sri_Lanka

[6] Ii. S. D.R, A. Buqhari, and O. Andy, "THE EFFECT OF STRESS AND PRODUCTIVITY DUE TO TRAFFIC CONGESTION AMONG WORK," *Proceeding of International Conference Health, Science And Technology (ICOHETECH)*, 2023, Accessed: Jul. 12, 2024. [Online]. Available: https://journals.indexcopernicus.com/search/article?articleId=3757262

[7] BPS Provinsi DKI Jakarta, "Transportation Statistics of DKI Jakarta 2024." Accessed: Jul. 12, 2024. [Online]. Available: https://jakarta.bps.go.id/publication/2023/11/23/50c5745cdc2f0949e4fc47ec/statistik-transportasi-provinsi-dki-jakarta-2022.html

[8] A. Pristanto, O. T. Lauren, D. Aryani, and S. Sahara, "Analisis Kepuasan Pelanggan Terhadap Fasilitas yang Disediakan oleh Pihak Transjakarta," *Jurnal Manajemen Riset Inovasi*, vol. 1, no. 3, pp. 09–17, May 2023, doi: 10.55606/MRI.V1I3.1162.

[9] M. I. Cahyani, M. Halimah, and B. Bonti, "PENGARUH KUALITAS PELAYANAN TERHADAP KEPUASAN PELANGGAN BUS TRANSJAKARTA PADA KORIDOR 7 (KP. RAMBUTAN – KP. MELAYU)," *JANE - Jurnal Administrasi Negara*, vol. 14, no. 1, p. 71, Aug. 2022, doi: 10.24198/JANE.V14I1.41267.

[10] G. Eser, "Sentiment Analysis and Rating Prediction for App Reviews Using Transformer-based Models," 2024, Accessed: Jul. 12, 2024. [Online]. Available: https://www.researchgate.net/publication/381887825

[11] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Applied Computational Intelligence and Soft*

Computing, vol. 2024, no. 1, p. 2826773, Jan. 2024, doi: 10.1155/2024/2826773.

[12] Md. S. Islam and K. M. Alam, "Sentiment analysis of Bangla language using a new comprehensive dataset BangDSA and the novel feature metric skipBangla-BERT," *Natural Language Processing Journal*, vol. 7, p. 100069, Jun. 2024, doi: 10.1016/J.NLP.2024.100069.

[13] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/S11042-022-13428-4/FIGURES/3.

[14] S. M. Parikh and M. K. Shah, "Analysis of various sentiment analysis techniques of NLP," *Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*, pp. 1323–1331, Feb. 2021, doi: 10.1109/ICICV50876.2021.9388525.

[15] J. Barry, "Sentiment Analysis of Online Reviews Using Bag-of-Words and LSTM Approaches," 2017, Accessed: Jul. 12, 2024. [Online]. Available: https://www.yelp.com/dataset/challenge

[16] M. I. Alfarizi, L. Syafaah, and M. Lestandy, "Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory)," *JUITA: Jurnal Informatika*, vol. 10, no. 2, pp. 225–232, Nov. 2022, doi: 10.30595/JUITA.V10I2.13262.

[17] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.

[18] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews," *Procedia Comput Sci*, vol. 179, pp. 728–735, 2021, doi: 10.1016/J.PROCS.2021.01.061.

[19] C. G. Kim, Y. J. Hwang, and C. Kamyod, "A Study of Profanity Effect in Sentiment Analysis on Natural Language Processing Using ANN," *Journal of Web Engineering*, vol. 21, no. 3, pp. 751–766, Mar. 2022, doi: 10.13052/JWE1540-9589.2139.

[20] D. K. Nugroho, "US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis," *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, pp. 136–141, Jan. 2021, doi: 10.1109/CONFLUENCE51648.2021.9377201.

[21] A. Winanto and C. Budihartanti, "Comparison of the Accuracy of Sentiment Analysis on the Twitter of the DKI Jakarta Provincial Government during the COVID-19 Vaccine Time," *Journal of Computer Science and Engineering (JCSE)*, vol. 3, no. 1, pp. 14–27, Feb. 2022, doi: 10.36596/JCSE.V3I1.249.

[22] D. A. Kristiyanti, R. Aulianita, D. A. Putri, L. A. Utami, F. Agustini, and Z. I. Alfianti, "Sentiment Classification Twitter of LRT, MRT, and Transjakarta Transportation using Support Vector Machine," *2022 International Conference of Science and Information Technology in Smart Administration, ICSINTESA 2022*, pp. 143–148, 2022, doi: 10.1109/ICSINTESA56431.2022.10041651.

[23] B. D. Meilani, R. K. Hapsari, and I. F. Novian, "Classification of community opinion on the use of the Transjakarta bus based on twitter social network using naïve bayes method," *IOP Conf Ser Mater Sci Eng*, vol. 1010, no. 1, Jan. 2021, doi: 10.1088/1757-899X/1010/1/012030.

[24] N. T. Hidayat and Suharjito, "User Satisfaction SentimentAnalysis for Mass Transportation Infrastructure (Case Study of Manggarai Station)," *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2023*, pp. 14–19, 2023, doi: 10.1109/ICITACEE58587.2023.10277391.

[25] Z. Nurthohari, D. I. Sensuse, and S. Lusa, "Sentiment Analysis of Jakarta Bus Rapid Transportation Services using Support Vector Machine," *2022 International Conference on Data Science and Its*

Applications, ICoDSA 2022, pp. 171–176, 2022, doi: 10.1109/ICODSA55874.2022.9862903.

[26] M. O. Pratama *et al.*, "The sentiment analysis of Indonesia commuter line using machine learning based on twitter data," *J Phys Conf Ser*, vol. 1193, no. 1, Apr. 2019, doi: 10.1088/1742-6596/1193/1/012029.

[27] F. Ladayya, D. Siregar, W. E. Pranoto, and H. D. Muchtar, "Analisis Sentimen pada Program Transportasi Publik JakLingko dengan Metode Support Vector Machine," *Jurnal Statistika dan Aplikasinya*, vol. 6, no. 2, pp. 381–392, Dec. 2022, doi: 10.21009/JSA.06221.

[28] "The Ultimate Guide to Building Your Own LSTM Models." Accessed: Jul. 12, 2024. [Online]. Available: https://www.projectpro.io/article/lstm-model/832

[29] "How to explain the ROC AUC score and ROC curve?" Accessed: Jul. 12, 2024. [Online]. Available: https://www.evidentlyai.com/classification-metrics/explain-roc-curve