ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 10 (2023) No. 2, pp. 1-14
DOI:10.33168/JLISS.2023.0201

A Study on Social Media Responses on Road Infrastructure using Sentiment Analysis

Farahnatasyah Abdul Hanan[1], Sofianita Mutalib[1,3], Aizat Mohd Yunus[1], Mohd Fadzil Abdul Rashid[2], Siti Nur Kamaliah Kamarudin[1], Shuzlina Abdul Rahman[1]

[1]School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknology MARA, 40450 Shah Alam, Selangor, Malaysia

[2]Department of Built Environment Studies and Technology, College of Built Environment, Universiti Teknologi MARA Perak Branch, Tapah, Perak, Malaysia

[3]Research Initiative Group Intelligent Systems, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

sofianita@uitm.edu.my (Corresponding Author)

Abstract. This study explores the sentiment of road users on Twitter and provides detailed insights into how the sentiment towards road infrastructure and conditions is perceived in a social media environment. The analysis plays an important role for the government, considering that Selangor is one of the major contributors to the Malaysian economy. The study's results can be used to inform local authorities and urban planners about the current condition of road infrastructure. The sentiment of road users was conceptualized as a multilingual construct, considering both Malay and English words, using a text-mining approach. Tweets were collected through the Twitter application programming language (API) and labeled using VADER and SentiWordNet, which are two commonly used lexicon-based techniques. Several machine learning classifiers were explored, namely Support Vector Machine, Naïve Bayes, and K-Nearest Neighbor algorithms. The Support Vector Machines algorithm with VADER lexicon and 20-fold cross-validation yielded the best performing model with an accuracy of 78.67%. With the exception of the Sepang district, the number of negative polarities exceeds that of positive polarities. It could indicate that the Sepang district road is high quality and has fewer hotspot barriers. Additionally, the insights were presented in a dashboard for better understanding, as the contribution for road infrastructure sentiment analysis with classifiers. Nevertheless, including more cities and states in the study would enhance its generalizability. The findings provide insight for the target users, particularly the authorities managing and maintaining the road, and raise public awareness about Selangor's road infrastructure and conditions.

Keywords: Classification, Lexicons, road infrastructure, Support Vector Machine, visualization.

1. Introduction

Road networks play a vital role in providing efficient movement of people, commodities, and services for a wide range of commercial and social activities. The availability of high-quality road networks, such as expressways, can enhance the speed and efficiency of domestic and international trades by reducing transportation time and costs, whereas the development or provision of high availability local roads provides easy land access and promotes commercial and social activities at the local level (Ng et al., 2019). Therefore, developing good urban road networks is essential to ensuring the city continuously grows sustainably. To this extent, road infrastructure has become the centre of attention among users and is a significant public asset that should be carefully managed during its life cycle. The quality of road infrastructure directly influences the users' satisfaction, safety, health, and security when commuting daily. With the emergence of technology and communication, social media platforms have gained attention from people around the globe. It opens an expansive room for research to examine the users' responses or opinions via social media platforms and consider them in road infrastructure decision-making. Shortly, opinion-based research can be referred as a judgment or belief established on specific things and can be supported with or without fact or expertise and sentiment analysis is also called as opinion analysis or opinion mining (Wankhade et al., 2022).

A massive range of opinions possibly collected to conduct a sentiment analysis - a technique of identifying and classifying the views or sentiments stated in opinionated data to determine if the writer's attitude toward a particular service, product, etc. is expressing positive, negative or neutral feelings (Shahnawaz & Astya, 2012; Shaeeali et al, 2020). Thus, opinions shape the perception and evaluation of reality. Inspired by this, the current paper attempts to comprehend the users' opinions in order to understand the reality or conditions of road infrastructure in Selangor, Malaysia, and visualize them in the form of a dashboard. Moreover, since microblogging is popular among social media because it allows them to post their opinions and discuss viewpoints on the latest issues (Aliprandi et al., 2012; Nasir and Palanichamy, 2022; Kim and Lim, 2022), this study utilizes this platform accordingly.

There exists multiple news related to the severe condition of routes in Selangor, such as potholes, uneven road surfaces, etc. Poor road infrastructure can affect a user's level of compliance, resulting in a risky situation among drivers (Idrus et al., 2016; Khairul et al., 2018). The state of Selangor has documented an increase in death cases from 2013 until 2016 (from 1019 to 1140) but sharply decreased in 2017 with only 627 cases. It shows promising results, but preventing death or injury from the accident should be the target.

Furthermore, the existence and purpose of social media serve diverse purposes and different target audiences. There are a few challenges in using social media platforms, such as one's opinion or viewpoint may include more than one entity-organization, individual, or business. Thus, it is compulsory to recognize the body the author or writer refers to (Kim and Lim, 2022; Sonagi & Gore, 2013; Wankhade et al., 2022). In addition, the data posted by the public is hard to interpret as positive or negative tweets because people in this era, especially Generation Z, post tweets without context. They tend to write quirky or sarcastic remarks. Therefore, to control these issues, an algorithm technique is needed to get high output accuracy. To the best of our knowledge, this is the first study to analyze the sentiment of road infrastructure in Selangor on Twitter, as compared to transportation service which is commonly analyzed.

The followings are the contributions of this study:

- Exploration of Support Vector Machine, Naïve Bayes, and K-Nearest Neighbor Classifiers in analyzing the sentiments of road infrastructure and conditions using a Twitter dataset.
- Creation of dashboard for interactive and up-to-date information and awareness, which can assist authorities or urban planners in decision-making.

The remaining of this paper is structured as follows: Section 2 discusses the related works, while Section 3 describes the study's methodology. Section 4 highlights the results and findings with the discussions, and finally Section 5 concludes the paper with future works.

2. Various Machine Learning Related Works

According to Adilah et al. (2020), one of the studies which applied sentiment analysis was conducted to observe *Gojek* users' opinions towards its transportation services. Indonesia's *Gojek* was founded in 2010, and it provides online transportation services using either motorcycles or cars. Adilah et al. (2020) applied Naïve Bayes' method to get a rule for predicting the sentiments, and the results were reported in a confusion matrix for accuracy, recall, and precision. Another similar research was also conducted by Anastasia and Budi (2016). The main difference implied in this research was the analysis of the results and the addition of the company *Grab* as their source of study. The author retrieved the dataset based on keywords such as "@gojekindonesia," "Gojek", "@GrabID", and "Grab". A total of 19,918 tweets were gained based on the phrase "@gojekindonesia," 34,272 tweets for "Go-jek", for "@GrabID" is 6,503, and lastly, 65,712 tweets for "Grab". This research also applied three methods: Support Vector Machine, Naïve Bayes, and Decision Tree algorithms (Anastasia and Budi, 2016).

Seliverstov et al. (2020) researched traffic safety evaluation in the Northwestern Federal District based on internet users' reviews. The research aimed to ensure the traffic condition in Northwestern Federal District was in good condition, and the author collected the data from Autostrada.info/ru with a total of 1130 text formats. This dataset contained opinions regarding road networks and its requirements in the Northwestern Federal District of Russia. The techniques used in the research were Naïve Bayes and Linear classifier model with stochastic gradient descent optimization. The results of the opinions were classified as positive and negative, in which the best result achieved was 71.94% accuracy (Seliverstov et al., 2020).

Various machine learning methods have been applied in transportation service reviews. Other machine learning techniques should be applied in different research domains for better accuracy and performance. Machine learning is a subset of the larger area of artificial intelligence, which also comprises others like Knowledge Representation, Perception, and Creativity (Pereira and Borysov, 2019). Machine learning explores many algorithms and approaches for automating complicated tasks that are difficult to solve using conventional programming methods (Gopinath et al., 2019). Depending on the data availability and the research problem's suitability, different types of supervised and unsupervised machine learning algorithms should be applied. The supervised algorithm uses labeled data to enable them to apply what they have learned in the past data and predict future events. In contrast, unsupervised algorithms were used when the input data is trained to be categorized or labeled according to its similarities (Varone et al., 2020).

Support vector machine (SVM) is a supervised learning algorithm for analyzing data and recognizing patterns (Nguyen, 2017). They are commonly utilized in classification and regression problems. SVM works with structural risk minimization and dimensional theory, a robust machine learning approach for binary classification (Zhang et al., 2016). The technique works by searching for a hyperplane or decision boundary that divides one class from another, for example, to discover the optimum hyperplane from an infinite number of functions (Bourequat et al., 2021).

The Naïve Bayes (NB) classifier is a group of basic probability classifiers based on the premise that all variables are independent in the presence of a category variable (Xu, 2018). Naïve Bayes classifier is a machine learning algorithm that is simple to use and works well with textual data. Naïve Bayes is also significantly good in terms of space and time complexity. In text document classification, the Naïve Bayes model is used to view the document as an event, and the probability of the non-occurrence of words in the document is checked (Ranjitha, 2019). Some of the previous studies chose SVM and NB classifiers for the comparative studies in sentiment analysis (Haris et al., 2023; Mohd Nasir & Palanichamy, 2022).

The K-Nearest Neighbor approach is among the most straightforward machine learning techniques. This technique aims to acquire the training set and then predict the label of every new instance based on the set's nearest labeled neighbors (Yu and Thandar Nwet, 2020). This technique has two stages; the first is the determination of the closest neighbor, followed by the class using those neighbors (Cunningham and Delany, 2022). The distance between the test data and all training samples will first be calculated using the Euclidean distance formula. If the distance between the training samples and the test samples is less than or equal to the Kth shortest distance, the K-Nearest Neighbor may be considered (Cunningham and Delany, 2022). Conclusively, based on the literature studies conducted, our paper reports on experiments with three commonly used machine learning methods for sentiment analysis based on classification SVM, Naives Bayes, and K-Nearest Neighbor.

3. Methodology

This section presents the methodology of the study. Fig. 1 illustrates the phases involved in this study which are detailed in the following subsections. It began with data extraction from the Twitter platform into the social media responses dataset. Next, the dataset was cleaned and pre-processed using commonly used data pre-processing steps (i.e., removal of duplicates and missing values). Subsequently, the data was labeled accordingly for classification using the cross-validation training and testing approach. Finally, the dashboard was developed for a holistic view of the classification results based on the selected location and sentiments.

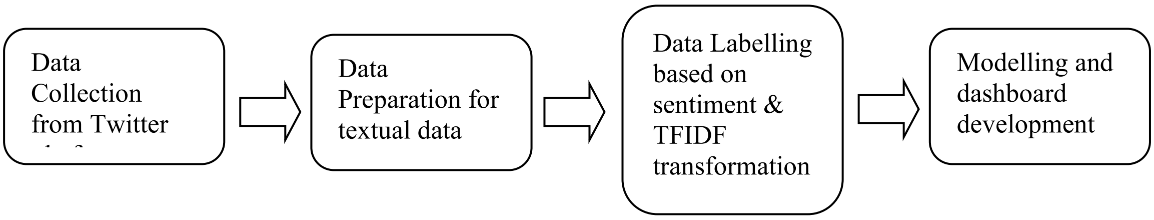

Data Collection from Twitter
Data Preparation for textual data
Data Labelling based on sentiment & TFIDF transformation
Modelling and dashboard development

Fig. 1: Phase of work in the study

3.1. Data Collection and Pre-processing

The author collected the data from an online social media platform (Twitter), and all data collected were related to the condition of Selangor's road infrastructure. Firstly, the author identified all possible road infrastructure keywords as input for collecting raw data through the Visual Code Studio (VS Code) software using Twitter Intelligent Tools (TWINT) module. The Tweets selected were limited to the English and Malay languages only. A sum of 62 number of keywords were identified, including the name of roads in the districts of Selangor (Malaysia) (as shown in Table 1), the condition of roads such as "uneven road surface", "road crack", "pothole" as well as the username account for Malaysian Public Works Department (also known as *Jabatan Kerja Raya* (JKR)). Keywords that do not contain the word "Selangor" were scraped using the geolocation information to make sure that the tweets scraped were centered in Selangor only. The dataset scraped was saved in a separate CSV file according to each keyword identified. The dataset scraped dated within the last five years from January 1 2017, until June 6 2022, and a total of 36 attributes were recorded.

Table 1. Sample of keywords for location related in road condition

Selangor District	**Keywords**
Klang	Jalan Klang, Jalan Kapar, Jalan Dato' Mohd Sidin Jalan Haji Sirat, Jalan Johan Setia, Jalan Kebun Jalan Sungai Pusu, Jalan Sungai Tua
Hulu Langat	Jalan Ampang, Jalan Cheras, Jalan Hulu Langat Jalan Kajang, Jalan Semenyih, Jalan Enam Kaki Jalan Kerja Ayer Lama
Gombak	Jalan Setapak, Jalan Rawang, Jalan Batu Arang Jalan Batu Caves, Jalan Kuang
Hulu Selangor	Jalan Hulu Bernam, Jalan Sungai Tinggi, Jalan Kalumpang Jalan Ulu Yam, Jalan Serendah, Jalan Pertak Jalan Batang Kali, Jalan Tanjung Karang
Kuala Langat	Jalan Banting, Jalan Cheeding Jalan Jugra, Jalan Kelanang
Kuala Selangor	Jalan Api-Api, Jalan Bukit Badong, Jalan Bukit Belimbing Jalan Ijok, Jalan Jeram, Jalan Kuala Selangor
Petaling	Jalan Bukit Raja, Jalan Sungai Buloh Jalan Petaling
Sabak Bernam	Jalan Sabak Bernam, Jalan Pasir Panjang
Sepang	Jalan Dengkil, Jalan Labu

In this research, data preparation methods applied were such as "remove attribute", "attribute generation" and "data translation". Hence some unnecessary attributes were removed, such as attribute id, conversation, created_at, timezone, user_id, username, name, place, language, mentions, urls, photos, replies _count, retweets_count, likes_count, hashtags, cashtags, link, retweet, quote_url, video, thumbnail, near, geo, source, user_rt_id, user_rt, retweet_id, reply_to, retweet_date, translate, and trans_src, trans_dest. The "Location" attribute was an essential attribute for data visualization as it helped give a better visual of the original location of the tweet. The tweets that contain the name of the road were maintained as the value in the "Location" attribute, while missing values were replaced with the "Selangor" value.

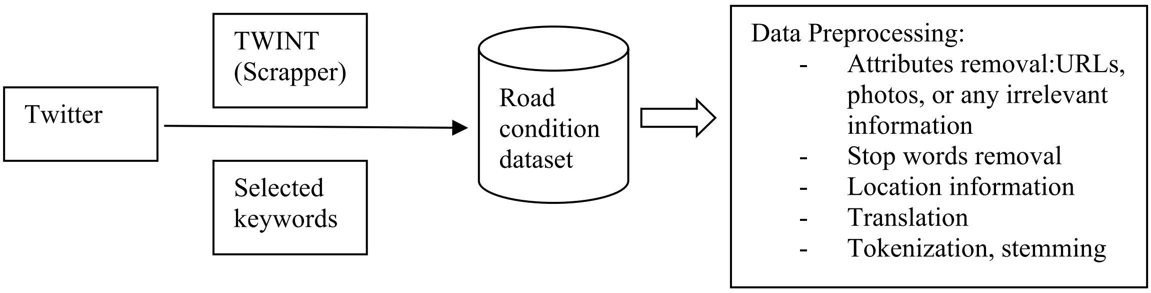
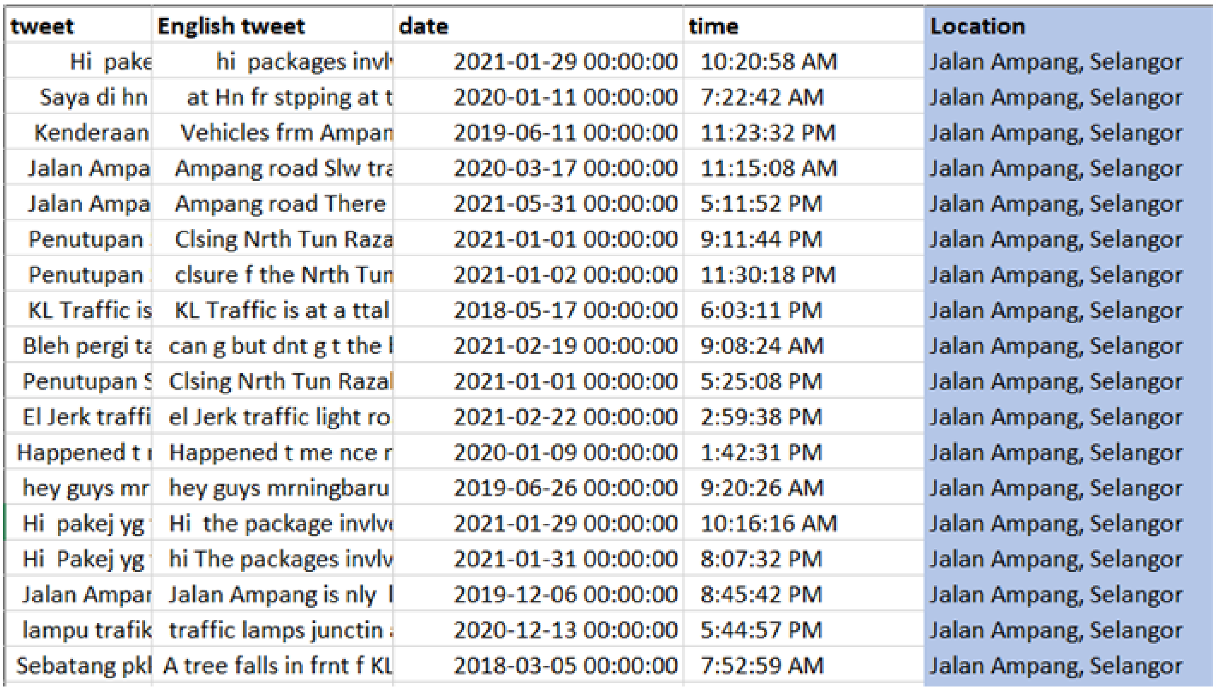


Fig. 2: Task of data preparation in the study

All records were then transferred onto Google Sheets to translate Malay text into English. Data translation was required as this research uses Valence Aware Dictionary for Sentiment Reasoning (VADER) module, which contains English text for text sentiment analysis. The formula =GOOGLETRANSLATE (cell with text, “source language,” “target language”) was applied to enable data translation of Malay text into English. Next, the Excel file was used for further processing, and the sample of the record is shown in Fig.3.



tweet	English tweet	date	time	Location
Hi pake	hi packages invl	2021-01-29 00:00:00	10:20:58 AM	Jalan Ampang, Selangor
Saya di hn	at Hn fr stpping at t	2020-01-11 00:00:00	7:22:42 AM	Jalan Ampang, Selangor
Kenderaan	Vehicles frm Ampan	2019-06-11 00:00:00	11:23:32 PM	Jalan Ampang, Selangor
Jalan Ampa	Ampang road Slw tra	2020-03-17 00:00:00	11:15:08 AM	Jalan Ampang, Selangor
Jalan Ampa	Ampang road There	2021-05-31 00:00:00	5:11:52 PM	Jalan Ampang, Selangor
Penutupan	Clsing Nrth Tun Raza	2021-01-01 00:00:00	9:11:44 PM	Jalan Ampang, Selangor
Penutupan	clsure f the Nrth Tun	2021-01-02 00:00:00	11:30:18 PM	Jalan Ampang, Selangor
KL Traffic is	KL Traffic is at a ttal	2018-05-17 00:00:00	6:03:11 PM	Jalan Ampang, Selangor
Bleh pergi ta	can g but dnt g t the	2021-02-19 00:00:00	9:08:24 AM	Jalan Ampang, Selangor
Penutupan S	Clsing Nrth Tun Razal	2021-01-01 00:00:00	5:25:08 PM	Jalan Ampang, Selangor
El Jerk traffi	el Jerk traffic light ro	2021-02-22 00:00:00	2:59:38 PM	Jalan Ampang, Selangor
Happened t	Happened t me nce r	2020-01-09 00:00:00	1:42:31 PM	Jalan Ampang, Selangor
hey guys mr	hey guys mrningbaru	2019-06-26 00:00:00	9:20:26 AM	Jalan Ampang, Selangor
Hi pakej yg	Hi the package invlv	2021-01-29 00:00:00	10:16:16 AM	Jalan Ampang, Selangor
Hi Pakej yg	hi The packages invlv	2021-01-31 00:00:00	8:07:32 PM	Jalan Ampang, Selangor
Jalan Ampar	Jalan Ampang is nly	2019-12-06 00:00:00	8:45:42 PM	Jalan Ampang, Selangor
lampu trafik	traffic lamps junctin	2020-12-13 00:00:00	5:44:57 PM	Jalan Ampang, Selangor
Sebatang pkl	A tree falls in frnt f KL	2018-03-05 00:00:00	7:52:59 AM	Jalan Ampang, Selangor

Fig. 3: Sample of records with location information

3.2. Data Labelling

The data labeling process was completed using RapidMiner' Extract Sentiment Operator. The dataset was labeled using two lexicon operators called VADER and SentiWordNet3.0. The technique employed a dictionary-based strategy to separate the polarity into positive and negative by utilizing the score from VADER and SentiWordNet3.0. The highest score for positive and negative were 0.85 and -0.85, respectively. Words related to bad road infrastructure, such as *pothole, uneven surface, cracks, crack, leak, jam, hole, holes* and *puddle* were scored -0.85. Since the data were translated using the google translate extension, words like "hollow" and "perforated" referred to *potholes*. In addition, words such as "please" and "report" were also labeled as negative regarding public attempts to notify the local authorities about road maintenance. Subsequently, the sentiment showed problems related to road infrastructure and terrible conditions at the respective location.

Positive words like *convenient, paving, patch*, and *pavement* scored 0.85 to indicate the positive sentiment in the tweets. On the contrary, the term "not" and "no" received the value -0.31. At the same time, truncated words such as "qt" and "x" were manually added to the corpus (due to the inability of google translate to detect the slang words) and scored based on their original word's score ("thank" and "not"). Consecutively, the "Generate Attribute" operator was used to automatically label the records as positive, negative, and neutral using the reference score as in Table 2.

Table 2. Sentiment score and class label assignment

Sentiment Score	**Class Label**
If score > 0	Positive
If score < 0	Negative
If score = 0	Neutral

However, for this research, the records with the class label “Neutral” were then deleted from the dataset and considered irrelevant data. This is because, based on the research observation, most of the records consisted of only generic words related to roads and thus do not hold any significant meaning to the road infrastructure.

3.3. Model Development

This phase comprised several activities, such as the identification of appropriate algorithms for the study, as discussed in the previous section. This study classified the dataset into positive or negative tweets using the chosen supervised machine learning algorithms, namely SVM, Naïve Bayes, and K-Nearest Neighbor (KNN) classification.

The model development was conducted using RapidMiner software, involving six operators. Firstly, "Read Excel file" were used to retrieve the dataset. Next, "Select Attributes" were used for selecting attributes to be included. In this case, attribute English Tweets and Sentiment were selected. Then, the "Set Role" operator was used to set the role of each column. "Attribute Sentiment" was set the role as 'label'. The "nominal to text" operator was used to change the attribute's data type for the models to work.

Subsequently, the "Process Documents from Data" operator created a word vector representing the text data. This operator uses Term Frequency-Inverse Document Frequency (TF-IDF) to choose a subset of the model's related features. TF shows the term frequency of a word in the documents, while IDF provides additional value to the rarely occurring words in the documents. The calculation for TF, IDF, and TF-IDF are shown in equations 1, 2, and 3, respectively.

$$TF = \frac{Number\ of\ times\ word\ appears\ in\ a\ document}{Total\ number\ words\ in\ a\ document} \tag{1}$$

$$IDF = log_e \frac{Total\ number\ of\ documents}{Total\ number\ of\ documents\ with\ word\ in\ them} \tag{2}$$

$$TFIDF = TF \times IDF \tag{3}$$

Based on the mentioned approaches, the dataset was transformed into vector space representation to complete the data preparation phase, which will be a catalyst in improving the model’s performance.

Lastly, the Cross-Validation operator was used to test and train the models. For this analysis, the k-fold cross-validation values were 10 and 20. This analysis used SVM, Naïve Bayes, and KNN models in the default parameter setting. The "Sample" operator was used due to the incapability of the KNN model to run all samples in parallel. Finally, the "Performance" operator was used to measure the accuracy of the models as part of model evaluation.

3.4. Model Evaluation

In the model evaluation phase, the dataset was split into two sections: training data and testing data using the cross-validation method. Both training and testing data were then prepared to evaluate the analysis findings according to its accuracy, precision, and recall (as shown in Table 3). Then, a confusion matrix was produced for results validation and relevancy to the study.

A confusion matrix is required to choose the best model between SVM, Naïve Bayes, and K-Nearest Neighbor. In this study, a 2 x 2 confusion matrix was suitable to be used because it classified positive and negative sentiment. True Positive (TP) are positive subjects that have been correctly labeled as positives, False Positive (FP) are negative subjects that have been incorrectly labeled as positives, True Negative (TN) are negative subjects that have been correctly labeled as negative, and False Negative (FN) are positive subjects that have been incorrectly labeled as negative (Fithriasari et al, 2020).

Table 3. Model evaluation metrics

Metrics	Formula
Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$
Precision	$\frac{TP}{(TP+FP)}$
Recall	$\frac{TP}{(TP+FN)}$

3.5. Dashboard Development

The final stage is the findings display based on an informative PowerBI dashboard. In this stage, all important results from the descriptive analytics model of SVM, Naïve Bayes, and K-Nearest Neighbor will be transformed into a dashboard for easy data viewing and retrieval. The most frequently used words will be displayed in the dashboard using a word cloud. Moreover, a bar chart will be embedded to depict positive and negative sentiments for the overall view. A line chart used shows the trend in tweets over the year. The most crucial dashboard element is a geo map to visualize public opinion on roads in the Selangor region.

4. Results and Findings

4.1. Analysis Bag of Words

The analysis for the bag of words method shows all keywords used after removing irrelevant data. The listed words are closely related and most relevant to the study. Words like road, '*jalan*', and traffic are frequently used in their tweets as concerns about the infrastructure and condition of the roads in Selangor, Malaysia. Furthermore, there is also a list of adjectives which includes "good", "careful", as well as "damaged" that are used to describe the situation of roads. In addition, a group of words, such as Selangor, Klang, Kuala, Cheras, Ampang, and Sungai Buloh, are the areas' names under the location choice for this study. There are also suspend or emergency words used like accident and crash, how the conditions of the roads may lead to negative connotations such as mentioned above. Lastly, there are also words like please and help, which show the level of persuasion by the users in expressing their demand in wanting the infrastructure and condition of the roads in Selangor to be better.

4.2. Analysis of Data Labelling

These experiments used about 17,094 tweets. The output dataset using VADER contains 6733 positive labels and 10360 negative labels. While SentiWordNet 3.0 produces 7171 positive labels and 9815 negative labels. The distribution of positive and negative polarity can be viewed in Table 4, based on nine districts in Selangor. The number of negative polarities is higher than positive polarity for all districts except for the Sepang district. The result may reflect that the Sepang district road has good quality and fewer hotspot barriers. This is because the road in Sepang district might have a good road infrastructure and condition. However, there was not much difference between negative and positive polarity.

Table 4. Number of positive and negative polarity for each district

Districts	**Positive Label**	Percentage	**Negative Label**	Percentage
Gombak	224	28.7%	557	71.3%
Hulu Langat	1148	40.7%	1672	59.3%
Hulu Selangor	307	38.4%	493	61.6%
Klang	541	26.6%	1492	73.4%
Kuala Langat	35	21.3%	129	78.7%
Kuala Selangor	285	24.3%	887	75.7%
Petaling	255	26.4%	710	73.6%
Sabak Bernam	90	37.2%	152	62.8%
Sepang	450	55.8%	356	44.2%

4.3. Percentage of Positive Labels based on District

The top three districts in Selangor with a high percentage of a positive label would be Sepang, Hulu Langat, and Hulu Selangor, with 55.8%, 40.7%, and 38.4%, respectively, as shown in Table 4. In comparison, the top three that have a high percentage of the negative label for the district in Selangor are Kuala Langat, Kuala Selangor, and Petaling, with 78.7%, 73.6%, and 62.8%, respectively. However, all Selangor districts have higher negative than positive labels except for Sepang. Thus, the local authority that handles road maintenance should take this seriously. They must develop proper maintenance plans that keep road infrastructure in good condition and coordinate multiple repairs in nearby areas to avoid traffic jams. Road maintenance is indeed an important issue; however, the fact that negative samples are more than positive samples does not necessarily mean that the road situation is terrible. Users may not be motivated to tweet when their travel goes smoothly, but they may have some motivation when they encounter troubles.

4.4. Comparison Between Model Classifiers

Each model classifier would have its strengths and capabilities for producing results with acceptable accuracy. Table 5 shows a summary of accuracy levels of the SVM, NB, and KNN for users' review of road infrastructure in the study area. The SVM model based on VADER produced the highest accuracy level at 78.67% with cross-validation value k = 20. The accuracy of SVM and Naïve Bayes slightly increased as the number of fold k increased to k = 20. Overall, this study found that all model classifiers using VADER as a lexicon have higher accuracy than SentiWordNet. In other research, VADER was used with SVM and Random Forest, achieving accuracies of 67-69% and 56-65%, respectively (Haris et al., 2022). This study produces a slightly higher accuracy value using the VADER approach. Meanwhile, SentiWordNet was used with SVM and Random Forest with several combinations of n-gram and gained approximately 96-98% accuracy (Mohd Nasir and Palanichamy, 2022), which is better than our results without the n-gram approach.

Table 5. Summary of accuracy for each model

Classifier	Lexicon	Number of folds (k)	Precision	Recall	Accuracy
Support Vector Machine	VADER	10	76.90%	75.75%	78.25%
	VADER	20	77.45%	80.90%	78.67%
	SentiWordNet	10	70.89%	68.55%	70.20%
	SentiWordNet	20	70.94%	68.95%	70.35%
Naïve Bayes	VADER	10	68.91%	59.85%	66.42%
	VADER	20	68.37%	58.80%	65.8%
	SentiWordNet	10	63.76%	49.35%	60.65%
	SentiWordNet	20	63.85%	49.20%	60.68%
K-Nearest Neighbor	VADER	10	79.70%	66.15%	74.65%
	VADER	20	79.91%	65.65%	74.57%
	SentiWordNet	10	66.84%	66.10%	66.65%
	SentiWordNet	20	66.87%	66.50%	66.78%

4.5. Dashboard

For the final output, the overall results of the experiments were transferred into an interactive, user-friendly dashboard dedicated to a better planning and decision-making process. As a primary function, the dashboard begins with clicking the button "Start," redirecting users to the second page of the dashboard - the Introduction page, where all details such as project description, problem statement, and dataset description are displayed. The dashboard was designed according to the available information from the experiments, and it can be dynamically changed subject to the users' requirements in the future. Next, when clicking the Information button, the dashboard redirected to the overall findings related to the reviews in the study. Fig. 4 shows one of the pages when the user views the dashboard. For example, the Word Cloud visualization is the dataset's overall Bag of Words. The next step is to click on the Findings button, which redirects to displaying Districts in Selangor. This page is similar to the overall dashboard on the previous page. However, it is a more detailed view, focusing on each Selangor district. Users can click on the experiment button to redirect toward the experiment page, which shows the experiment used for this study. Users can choose either VADER or SentiWordNet by clicking the box to see the accuracy for each model when using the chosen lexicon, as shown in Fig. 4.

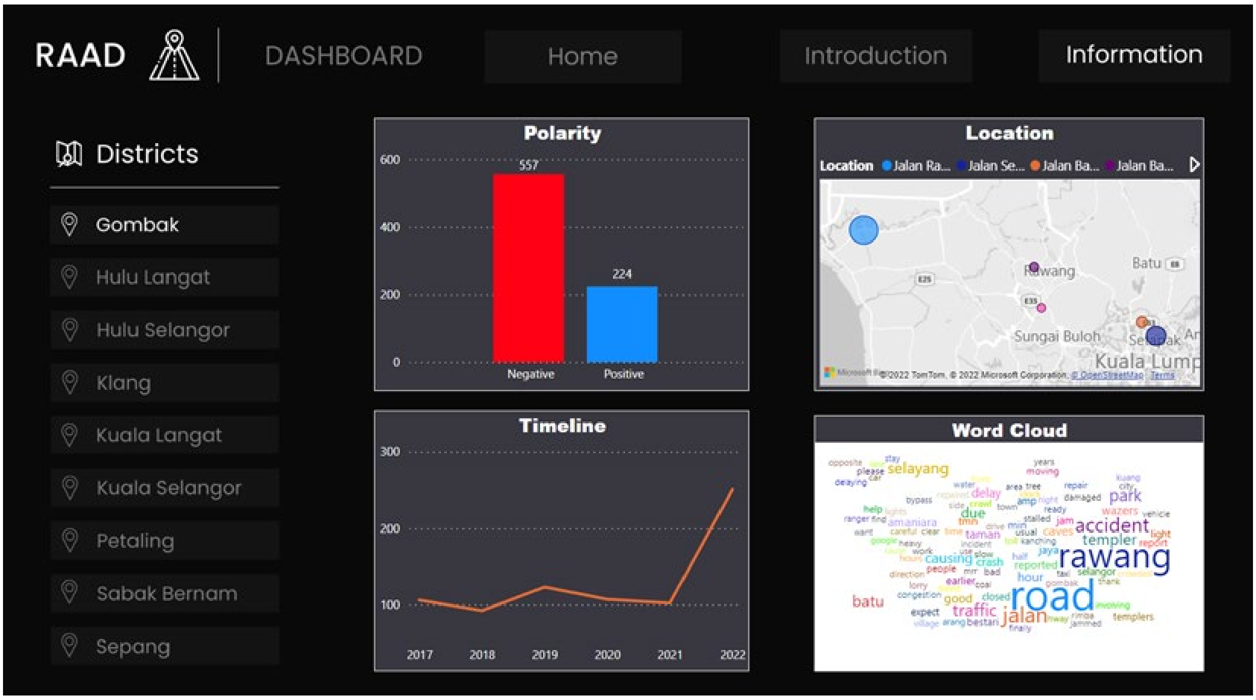


Fig. 4: Information Page with polarity, location and word cloud for samples in the selected location

5. Conclusions and Future Works

This study has presented the work on sentiment analysis of the road infrastructure in Selangor based on the Twitter platform using TWINT. The data extracted is the real issues that the public had discussed, including, but not limited to, problems of traffic lights not working, potholes that are not fixed on the roads, uneven surfaces on the streets, traffic congestion due to road infrastructure, and road accidents. Moreover, a bag of words had been constructed to eliminate all irrelevant data since some of the keywords used were the names of roads in districts of Selangor. Next, the dataset was labeled using the lexicon approach. Both lexicons methods, VADER and SentiWordNet, were used for labeling the datasets to either positive sentiment or negative sentiment. After that, the training and testing phases were done by applying three different models; SVM, Naïve Bayes, and K-Nearest Neighbour. Finally, the results show that the SVM model using VADER achieved the highest accuracy level for labeling the dataset and cross-validation k = 20 with 78.67%. For the users' benefit, an interactive dashboard was created for data visualization using Microsoft Power BI and published as a web page. The target users would be the authorities managing and maintaining the road. It also targets the public as it raises awareness regarding the road infrastructure and conditions in Selangor.

A limitation found in this project is insufficient data due to fewer Twitter users updating or raising their views about road infrastructure in Selangor. As such, public opinion, primarily to obtain information on spot improvements of the road infrastructures, is quite limited. Instead, Twitter users are more focused on tweeting and updating about congested spots. In addition, Malaysian users post on social media using mixed languages, such as Malay and English, in one sentence and tend to use short forms when posting on social media. This problem is difficult to handle, but this research has well-managed it by replacing those words with proper ones. By working on that, future work could include scraping the data on any other online platform, such as Facebook. As the amount of raw data or information increases, opinions on a specific topic become more valid. Also, as the study was undertaken, it was found that keywords related to road infrastructure needed to be added and fine-tuned to gain data. Worthy to note, there is a Malay corpus text called Malaya where data containing Malay words do not need to be translated during data preparation. Because of its wide usage, including Malay stop words and lemmatization, this existing Malay corpus text will aid future researchers in further developing this project. It is a component of future work that future authors must prioritize.

Acknowledgement

The authors would like to express the gratitude to Ministry of Higher Education, Malaysia for the research fund (FRGS/1/2021/ICT02/UITM/02/8) and School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research support.

References

Adilah, T., Supendar, H., Ningsih, R., Muryani, S., & Solecha, K. (2020). Sentiment analysis of online transportation service using the Naïve Bayes methods. Journal of Physics: Conference Series, 1641(1), 012093. https://doi.org/10.1088/1742-6596/1641/1/012093

Anastasia, S., Budi, I. (2016). Twitter sentiment analysis of online transportation service providers. 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 359–365. https://doi.org/10.1109/ICACSIS.2016.7872807

Bourequat, W., & Mourad, H. (2021). Sentiment analysis approach for analyzing iPhone release using support vector machine. International Journal of Advances in Data and Information Systems. 2(1). 36-44. https://doi.org/10.25008/ijadis.v2i1.1216

Fithriasari, K., Jannah, S. Z., & Reyhana, Z. (2020). Deep Learning for Social Media Sentiment Analysis. MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics, 36(2), 99–111. https://doi.org/10.11113/matematika.v36.n2.1226

Haris, N. A. K. M., Mutalib, S., Ab Malik, A. M., Abdul-Rahman, S., & Kamarudin, S. N. K. (2023). Sentiment classification from reviews for tourism analytics. International Journal of Advances in Intelligent Informatics, 9(1), 108-120.

Kamarudin, M. K. A., Wahab, N. A., Umar, R., Saudi, A. S. M., Saad, M. H. M., Rosdi, N. R. N., Razak, S. A. A., Merzuki, M. M., Abdullah, A. S., Amirah, S., & Ridzuan, A. M. (2018). Road traffic accident in Malaysia: Trends, selected underlying determinants and status intervention. International Journal of Engineering & Technology, 7(4.34), 112-117. https://doi.org/10.14419/ijet.v7i4.34.23839

Kim, J-Y. and Lim, C-K. (2022). The Use of Sentiment Analysis and Latent Dirichlet Allocation Topic-Modeling (LDA) on Web Novel Content Quality Factor, Journal of System and Management Sciences, 12 (2), 236-251. https://doi.org/10.33168/JSMS.2022.0211

Mohd Nasir, A. A., & Naveen, P. (2022). Sentiment Analysis of Covid-19 Tweets by Supervised Machine Learning Models. Journal of System and Management Sciences, 12(6), 50-69. https://doi.org/ 10.33168/JSMS.2022.0604

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., By, T. (2012). Sentiment Analysis on Social Media. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 919–926. https://doi.org/10.1109/ASONAM.2012.164

Ng, C. P., Law, T. H., Jakarni, F. M., & Kulanthayan, S. (2019). Road infrastructure development and economic growt.h. IOP Conference Series: Materials Science and Engineering, 512, 012045. https://doi.org/10.1088/1757-899X/512/1/012045

Padraig, C., & Delany, S. J. (2021). K-Nearest Neighbor classifiers – a tutorial. ACM Computing Surveys, 54(2), Article 25. https://doi.org/10.1145/3459665

Pereira, F. C., Borysov, S. S. (2019). Chapter 2 - Machine Learning Fundamentals. In C. Antoniou, L. Dimitriou, F. Pereira (Eds.), Mobility Patterns, Big Data and Transport Analytics (pp. 9–29). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-812970-8.00002-6

Rebala, G., Ravi, A., & Churiwala, S. (2019). Machine Learning Definition and Basics. In An Introduction to Machine Learning (pp. 1–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-15729-6_1

Samsuddin, N., & Masirin, M. I. M. (2016). Assessment of Road Infrastructures Pertaining to Malaysian Experience. MATEC Web of Conferences, 47, 03010. https://doi.org/10.1051/matecconf/20164703010

Seliverstov, Y., Seliverstov, S., Malygin, I., & Korolev, O. (2020). Traffic safety evaluation in Northwestern Federal District using sentiment analysis of Internet users' reviews. Transportation Research Procedia, 50, 626-635. https://doi.org/https://doi.org/10.1016/j.trpro.2020.10.074

Shaeeali, N. S., Mohamed, A. & Mutalib, S. (2020). Customer reviews analytics on food delivery services in social media: A review. IAES Int. J. Artif. Intell 9.4, 691-699.

Shahnawaz, Astya, P. (2017). Sentiment analysis: Approaches and open issues. 2017 International Conference on Computing, Communication and Automation (ICCCA), 154–158. https://doi.org/10.1109/CCAA.2017.8229791

Sonagi, A., & Gore, D. (2013). Sentiment Analysis and Challenges Involved: A Survey. International Journal of Science and Research (IJSR) ISSN, 4, 2319–7064. https://www.ijsr.net/archive/v4i1/SUB15677.pdf

Thein, Y., & Thandar, N. K. (2020). Sentiment analysis system for Myanmar news using k nearest neighbor and naïve bayes. International Workshop on Computer Science and Engineering (WCSE 2020). 1-5. 10.18178/wcse.2020.02.001

Ranjitha, K, V. (2019). Classification and optimization scheme for text data using machine learning Naïve Bayes. IEEE World Symposium on Communication Engineering, WSCE. 33-36. https://doi.org/10.1109/WSCE.2018.8690536

Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55, 5731–5780 (2022). https://doi.org/10.1007/s10462-022-10144-1

Xu, S. (2018). Bayesian Naïve Bayes classifier to text classification. Journal of Information Science, 44(3), 347-356. https://doi.org/10.1177/0165551510000000