Journal of Big Data

RESEARCH Open Access

Traffic and road conditions monitoring system using extracted information from Twitter

Prabu Kresna Putra[1,2], Rahmad Mahendra[1*] and Indra Budi[1]

*Correspondence:
rahmad.mahendra@cs.ui.ac.id

[1] Faculty of Computer Science, Universitas Indonesia, Kampus UI Depok, 16424 Depok, Indonesia
Full list of author information is available at the end of the article

Abstract

Congested roads and daily traffic jams cause traffic disturbances. A traffic monitoring system using closed-circuit television (CCTV) has been implemented, but the information gathered is still limited for public use. This research focuses on utilizing Twitter data to monitor traffic and road conditions. Traffic-related information is extracted from social media using text mining approach. The methods include Tweet classification for filtering relevant data, location information extraction, and geocoding in order to convert text-based location into coordinate information that can be deployed into Geographic Information System. We test several supervised classification algorithms in this study, i.e., Naïve Bayes, Random Forest, Logistic Regression, and Support Vector Machine. We experiment with Bag Of Words (BOW) and Term Frequency - Inverse Document Frequency (TF-IDF) as the feature representation. The location information is extracted using Named Entity Recognition (NER) and Part-Of-Speech (POS) Tagger. The geocoding is implemented using the ArcPy library. The best model for Tweet relevance classification is the Logistic Regression classifier with the feature combination of unigram and char n-gram, achieving an F1-score of 93%. The NER-based location extractor obtains an F1-score of 54% with a precision of 96%. The geocoding success rate for extracting the location information is 68%. In addition, a web-based visualization is also implemented in order to display traffic information using the spatial interface.

Keywords: Twitter, Text mining, Information extraction, Text classification, Geocoding, Road condition, Traffic situation

Introduction

Traffic disruption is a problem that often occurs in Indonesia, especially in the province of Special Capital District (DKI) Jakarta. According to data from the Central Agency of Statistics (BPS) that was presented in a 2018 land transportation statistics report, there are more than 20 million motorized vehicles in Jakarta [1]. The intense vehicular activity in Jakarta has various consequences, such as congestion, accidents, and disruption of alternate modes of transportation. According to the BPS data, more than 1500 traffic accidents occur annually in Jakarta [1]. In addition, according to the Tomtom Application Report, Jakarta ranks as one of the top ten most congested provinces in the world [2].

The government has made various attempts to reduce traffic disruption in Jakarta. One of these is the development of the Jakarta Smart City information system. The Jakarta Smart City information system uses closed-circuit television (CCTV) data from various sources, including the Transportation Agency (DisHub), Bali Tower, the Public Works Service (PU), and Transjakarta, among others. There are approximately 6000 CCTVs scattered throughout the Jakarta area. The data is sent and displayed in real time on the Jakarta Smart City system portal. However, the current smart traffic management system (STMS) still has several shortcomings. Because Jakarta Smart City relies on CCTV data, the system is highly dependent on the availability of data sources. Despite the existence of thousands of CCTV cameras, some CCTV data cannot be accessed or displayed. As a result, road conditions cannot be monitored optimally. Additionally, the coverage of CCTV data is still limited, and several public areas have not been captured by STMS. Furthermore, although Jakarta Smart City allows users to access video data from CCTVs, it has not provided thorough analyses regarding traffic situations. Consequently, users may still have difficulty determining what is happening at a specific road location.

Social media platforms are systems built on the internet technology that grew out of Web 2.0, which allows the exchange of information between social media users. Social media data can be analyzed using an information extraction approach. This study focuses on the use of Twitter data. There are several reasons why Twitter is used as a source of data in this study. Indonesia ranks 5th in the world for most Twitter users. In January 2022, there were 18.45 million twitter users in Indonesia. A survey finds that around 6000 tweets are sent per second. The abundance of data in Twitter is a valuable source of data to be harnessed in social media analytics. Moreover, Twitter provides an API that enables the researchers collect the data.

The objective of this study was to apply the machine learning models to extract information about traffic conditions in Jakarta. The model was built using social media data and can be utilized to monitor traffic situations and road conditions such as accidents and congestion in Jakarta. We hope that proposed application in this paper can be used as supplementary information for the public policy maker to plan and manage the traffic in Jakarta.

Related work

Several works have identified the potential use of Twitter to monitor traffic situations [3–7].

The study by D'Andrea [3] utilized social media data to obtain information about traffic-related events. This research focused on the Italian region and specifically analyzed traffic jams and traffic accidents. They built an intelligent system using a text-mining approach and machine learning algorithms to detect real-time traffic events by analyzing Twitter activities. The best model in this study was using the Support Vector Machine (SVM) algorithm, with an accuracy score of 95.75% and 88.89% in two-class and three-class classification, respectively.

The study by Gu [5] explored the use of social media data to monitor congestion, including recurring and non-recurring congestion. Recurring congestion is congestion that occurs frequently in a particular location, while non-recurring congestion is caused by external factors such as accidents, construction, climate, and other special events. They built the classification model for categorizing tweets as relating to either traffic incidents (TI) or non-traffic incidents (NTI) using the Naıve Bayes algorithm and the classification model for categorizing TI tweets into five categories using the sLDA algorithm. The first model achieved an accuracy 90.5%, while the second model produces true positive results 51% of the time, which indicates that geocodable TI tweets could be correctly classified by the sLDA classification algorithm. Other finding is that 5% of all the tweets fall into the useful (IT) tweet category and have location information. Of these, 60–70% come from "important users" (IU) such as online media, communities, and drivers, while the rest are from individual users. The results of the temporal analysis also show that the majority of sharing activities on Twitter occur on weekdays and during the day, especially at around noon.

Zhang's study [6] also investigated the detection of traffic accidents using data from Twitter. This study focused on detecting non-recurring congestion caused by external factors, including the accidents due to collisions, vehicular malfunctions, and small fires. The case study in Zhang's work is Northern Virginia, a metropolitan city in the United States. More than 500K Tweets was collected over period of January 2014 to December 2014. Time, date, and location information (in the form of latitude and longitude) were retained for each tweet. They experimented with four classifiers, namely SVM, artificial neural network (ANN), deep neural network (DNN), and long short-term memory (LSTM), for building the accident-classification model. DNN performed the best with an accuracy above 90%. Further validation shows that the accident incident data reported on Twitter is reliable information in accordance with the accident reports submitted to the existing system.

The study by Gutierrez [4] aimed to monitor traffic events in the UK using Twitter data. This study involved several steps in their methodology, including Tweet classification, event type classification, named entity recognition (NER), and temporal and location data extraction. The classification model reach the accuracy between 89% and 95%.

Herwanto's study [7] extracted information related to traffic conditions in Indonesia using Twitter data. This study has three major stages: data classification, entity extraction, and relation extraction. They collected the Tweets using several keywords. Beside that, they also acquired the Tweets from official Twitter account of Jakarta police office, e.g., '@TMCPoldaMetro'. The SVM classifier is used to categorize the Tweets into relevant and irrelevant classes. Entity extraction was performed using the NER method. The entities being recognized were location, time, and status. Then, the relation information was identified from the Tweet, i.e., condition, direction, and detail. The results shows that the accuracy of classification model is over 90% and the NER techniques produce an F1 score of 70%.

Model design

We design a pipeline model in order to monitor traffic situations using Twitter data. The pipeline model is described in Fig. 1. It consists of four standalone models, i.e., a classifier to filter out irrelevant Tweets from crawled data, a classifier to categorize relevant Tweets, a model to extract the location information from the Tweets, and a geocoding

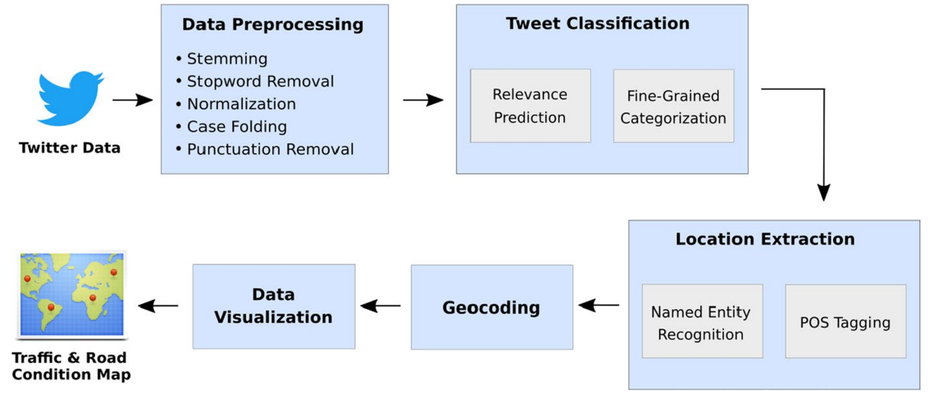


Fig. 1 Pipeline model for extracting traffic information

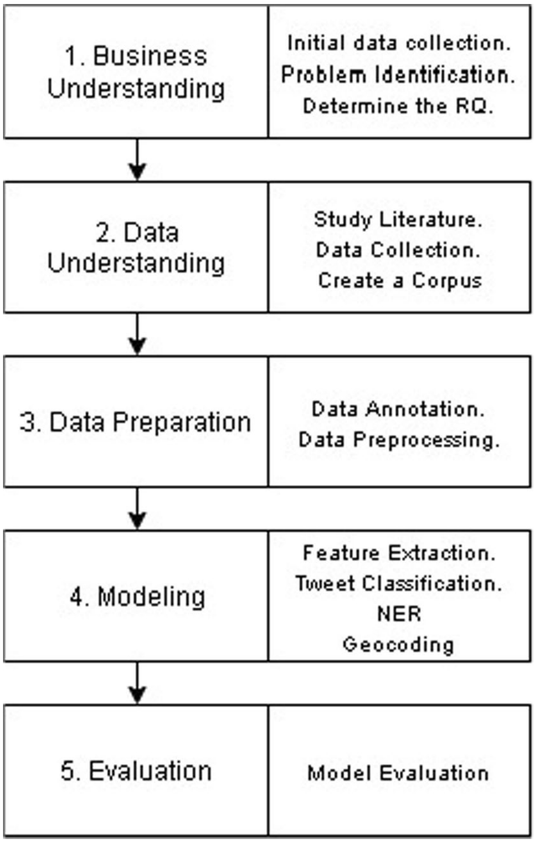

1. Business Understanding
Initial data collection. Problem Identification. Determine the RQ.
2. Data Understanding
Study Literature. Data Collection. Create a Corpus
3. Data Preparation
Data Annotation. Data Preprocessing.
4. Modeling
Feature Extraction. Tweet Classification. NER Geocoding
5. Evaluation
Model Evaluation

Fig. 2 Research methodology

model used to convert text locations into geographic data points in the form of latitude and longitude.

Methodology

This research is conducted using the quantitative method. Figure 2 illustrates the research methodology that is used. First, we identify the problem by observing the current system and related documents. Second, we study the literature related to text mining using Twitter data. After that, we collect and preprocess the data Then, we build the models and evaluate them.

Data collection

We employ data crawling to collect the Tweets. This method requires an application programming interface (API) obtained from the Twitter development website. The data is collected in the period between February to March 2020. The queries used for data collection are "jalan berlubang", "jalan rusak", "kecelakaan", "kerusakan jalan", "konstruksi", "lalu lintas", "lubang jalan", "macet", "mogok", "terbakar", "pembangunan", "perbaikan jalan", "situasi lalu lintas" and "tabrak". The important accounts used are @TMCPoldaMetro, @NTMCLantasPolri, @LewatMana, @infoll, @RadioElshinta, and @PTJASAMARGA. We acquire 280,412 tweets from crawling process. The data collected include the unique tweet number, the full text of the tweet, the coordinates of the device when used to post the tweet, the location listed in the user's profile, and the time at which the profile was created.

Data annotation

The annotation process is carried out to produce labeled data that is needed for training the supervised model. We sample 10,000 Tweets to be annotated. Two annotators perform data annotation in two steps. First, a Tweet is labeled as relevant if the Tweet content is related to a traffic situation. Otherwise, it is labeled as irrelevant. Second, the relevant Tweet is categorized into five event categories. Table 1 explains the definition of label for traffic event type annotation.

After the annotation is complete, the level of agreement is measured between the two annotators. We use the coefficient value of Cohen's kappa to determine the agreement [9]. We obtain that the Cohen Kappa are 93% for the relevance classification and 95% for the traffic event classification. The agreement between the two annotators in those two-step annotation can be interpreted as "almost perfect agreement". The remaining disagreement occurs due to the inaccuracy of the annotator when assigning the label. These differences are then discussed by the two annotators to determine which label is correct.

Table 1 Traffic event label for relevant Tweets

Label	Definition
Congestion	A condition in which the traffic flow that passes on the road under review exceeds the capacity of the road, which results in speeds approaching 0 km/h [8]
Accident	Incidents in road traffic involving at least one vehicle causing injury or damage, or loss to the owner (victim)
Road damage	A condition of damage to roads or traffic instruments
Vehicle problems	A condition in which there is a problem with the vehicle being used so that it cannot function normally and has a disruptive impact on traffic
Roadworks	A condition where an activity or work is carried out on the road

In addition, the location labeling process is also carried out. The annotation is done by single annotator, who is the first author of this paper. For location label, 400 tweets are annotated.

Data preprocessing

The data preprocessing in this study include several tasks, i.e., case folding, punctuation and stopword removal, stemming, and text normalization. We utilize Python Sastrawi library for stemming and stopword removal.

1. Case folding: lowercase the text representation.
2. Punctuation removal.
3. Stemming: strip the affixation of the word and return a word stem.
4. Stopwords removal: remove frequently occurring words or common words in text corpus.
5. Text normalization: transform the non-standard words (e.g., slang word) into a standardized form. In this study, we apply dictionary-based normalization using Salsabila dictionary [10].

Feature extraction

We apply word vectorization to extract the features from the Tweets. TF-IDF vectorization and count vectorization are the two types of feature extraction explored in this study. The vectorization is conducted using the Python sklearn library.

Text classification

The classification algorithms used in this study are Naıve Bayes, Random Forest, Support Vector Machine, and Logistic Regression.

1. Naïve Bayes, is a classifier that uses statistical and probability technique.
2. Random Forest, is a classifier whose structure in the form of a tree-shaped like a flowchart. The decision tree has a node that represents a condition or feature. Each node has a leaf indicating an output class label.
3. Support Vector Machine, a classifier that works by mapping the data point in vector space and separating the data point using a hyperplane.
4. Logistic Regression, is a classifier that uses the logistic function of the regression line formed from the features.

Named entity recognition

Named Entity Recognition method can be used to extract the location information from the Tweets [11, 12]. The location information in this research is extracted using Stanford NER and POS tagger. Stanford NER trained on Singgalang data [13] extracts three types of entity: person, organization, and location. On the other hand, POS tagger trained on Bahasa CSUI data [14] using the Conditional Random Field classifier tags the tokens in Tweets with Part-of-Speech information. We leverage the token tagged with Proper Noun class as the candidate of location.

Geocoding

Geocoding is the process by which ambiguous physical locations are represented with numerical coordinates [15]. The geocoding stage requires two different data sets, namely the data set of addresses to be identified with the location of coordinates and the address database to be used as a reference. Two address databases can be used as references in the ArcGIS software, namely databases from ESRI and spatial databases from other sources, such as government agencies authorized to issue spatial data.

The data geocoding in this study uses the Python programming language and utilizes the ArcPy library. The ArcPy Python library can perform analysis, conversion, management, and automation of geographic data. The geocoding process is carried out in two major stages, namely making a locator using a spatial map and geocoding for the dataset using a locator that has already been made. The geocoding locator created in this study uses a dataset of road name maps of Jakarta, sourced from OpenStreetMap, and has more than 64 000 unique addresses listed.

Evaluation

The experiment to determine the best model uses the k-fold cross-validation setting ($k = 10$). The evaluation metrics used are accuracy, precision, recall, and F1-Score.

Data visualization

To facilitate data visualization, we develop a web-based dashboard. The classification model, location information, and geocoding provide input at this stage. The model is run automatically against operational data and the predicted results of the model are visualized spatially and tabularly. The data visualization process is a series of programs that can retrieve data from previous modeling results and visualize that data in various forms.

The dashboard is built using the PHP (Hypertext Preprocessor), CSS (Cascading Style Sheets), and JS (JavaScript) programming languages. The first step is to create a design or mockup of the web interface. This is based on the monitoring system that was built by the Jakarta government to address the traffic situation of the Jakarta Smart City and on other monitoring systems that have been built by other government agencies. There is no interaction between the user and the system. The interface is divided into two displays. On one side is data whose location information has been successfully extracted, with latitude and longitude values shown in the spatial display of the Geographic Information System. On the other side is location information that could not be successfully extracted because the Tweet did not contain that data.

Experimental result

We conduct three experiments for Tweet classification task. First, we explore the preprocessing effect to model performance. Then, we investigate the feature extraction methods. Finally, we compare the performance among classifiers. In addition, the evaluation is also conducted for the model for location information extraction and geocoding.

Data preprocessing evaluation

Previous works shown that the data preprocessing can affect the classifier accuracy on text classification tasks [16–20]. To investigate the effect of preprocessing in the traffic Tweet corpus, we conduct the experiment with several variation of preprocessed data. We run SVM classifier using count vectorizer on data with four distinguished preprocessing setting, i.e., (1) original data without pre-processing, (2) data after case folding, (3) data with all preprocesing steps (case folding, punctuation removal, stemming, stopwords removal, and normalization), and (3) data with all preprocessing steps except stemming and stopwords removal. Table 2 presents the result of preprocessing experiment.

The result shown in Table 2 implies that, unlike in previous works, most preprocessing steps do not help to improve the classification accuracy. Only case folding increases the accuracy slightly.

Classification model evaluation

We experiment with different text representation to obtain the best classification features. For count vectorizer, we select unigram representation. While, for TF-IDF, we compare unigram, bigram, and trigram representation. In addition, we also explore char gram representation. We conduct feature extraction experiment in which we apply only case folding as the preprocessing step. Table 3 presents the results of model evaluation.

Based on the results presented in Table 3, the Logistic Regression classifier using combination of unigram count vector and char gram TF-IDF vector performs the best among all models tested, achieving the accuracy of 90.72% and 94.14% for Tweet relevance prediction and traffic event prediction type tasks, respectively. On the other hand, using TF-IDF trigram features only obtains the smaller accuracy compared to other features. Combining bigram and trigram features shows the positive trend of accuracy improvement.

Table 2 Evaluation of preprocessing on Jakarta traffic Tweet corpus

	Relevance pred	**Traffic event type pred**
Original	89.10%	93.70%
All preprocessing	88.40%	81.80%
All - {stemming, stopword removal}	89.10%	93.50%
Case folding	89.20%	93.80%

Table 3 Evaluation of classification model on Jakarta Traffic Tweet Corpus

	NB	SVM	LR	RF
Relevance prediction				
Count vector, unigram	87.44%	89.20%	90.55%	88.88%
TF-IDF, unigram	86.54%	90.00%	89.73%	88.96%
TF-IDF, bigram	79.32%	88.34%	88.17%	87.45%
TF-IDF, trigram	62.63%	82.52%	82.73%	82.21%
TF-IDF, char {1,2} gram	82.46%	90.40%	90.00%	86.88%
TF-IDF, bigram + trigram	85.81%	90.23%	90.04%	89.30%
Unigram + char gram	87.35%	89.65%	90.72%	87.49%
Traffic event type prediction				
Count vector, unigram	82.07%	93.79%	94.04%	90.10%
TF-IDF, unigram	81.51%	93.93%	92.35%	89.93%
TF-IDF, bigram	84.72%	89.60%	89.10%	87.64%
TF-IDF, trigram	77.46%	79.17%	79.26%	77.80%
TF-IDF, char {1,2}gram	76.30%	92.66%	91.41%	90.06%
TF-IDF, bigram + trigram	83.51%	93.93%	92.68%	91.81%
Unigram + char gram	81.72%	93.85%	94.14%	89.99%

Table 4 Evaluation of location extraction

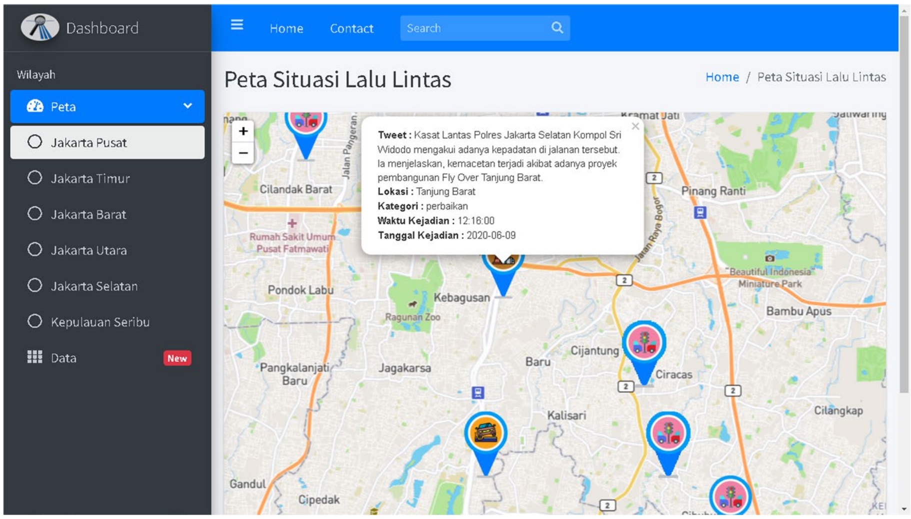
NLP tool	Label	Precision	Recall	F1-score
NER	Place	96.32%	37.88%	54.38%
POS tagger	NNP	32.95%	75.94%	45.96%

Location extraction evaluation

We compare the performance of two methods of extraction of location information in Table 4. Based on the evaluation results, NER can extract the location from the Tweet with precision more than 95%. However, it has low recall ($\leq$ 40%). We suspect that the model cannot generalize the location name from training data in which the sentences are predominantly written in standard Indonesian language, while the our traffic corpus that sourced from Twitter is informally written and may contain misspelling [21]. Conversely, using POS tagger for location extraction achieve much higher recall (76%) but very low precision (33%). The location is tagged as proper noun, but the proper noun can be other entities, e.g., person and organization.

Furthermore, we conduct geocoding experiments using the Jakarta road map sourced from OpenStreetMap. The map used contains geographic information consisting of 64,930 street names. These maps were used as locators, which were then employed to convert 419 true positive of predicted location extracted from NER and 840 true positive of predicted location extracted as proper noun labels from the POS tagger.

When converting the extracted location from Tweet into the pair of latitude and longitude point, 68% data is successfully identified with geographic information, i.e., 286 of the 419 Tweets with extracted location with NER method and 569 of the 840 Tweets with extracted location with POS tagger method. The location information in the rest of 32% data fails to recognize for several reasons. First, the locations are not



Tweet : Kasat Lantas Polres Jakarta Selatan Kompol Sri Widodo mengakui adanya kepadatan di jalanan tersebut. Ia menjelaskan, kemacetan terjadi akibat adanya proyek pembangunan Fly Over Tanjung Barat.
Lokasi : Tanjung Barat
Kategori : perbaikan
Waktu Kejadian : 12:16:00
Tanggal Kejadian : 2020-06-09

Fig. 3 Data visualization

Table 5 Several examples of extracted traffic information from Tweets

Tweet	Relevance	Category	Location
@PTJASAMARGA ada truk terperosok di km 82 tol cipularang arah jakarta'	Relevant	Traffic Accident	Cipularang
@DKIJakarta jalan MT Haryono di atas rel kereta Cawang yang arah Pancoran berlubang. Mohon perbaiki.'	Relevant	Road Damage	MT Haryono
05.36: Truk box mogok di Jalan Kembang Kerep, Kembangan, Jakbar depan SPBU Meruya Ilir	Relevant	Vechicle Problem	Kembang Kerep
16:24 Pejompongan Jalan Penjernihan 1 dari Per4an Karet menuju Slipi / Senayan macet	Relevant	Congestion	Pejompongan
Kecelakaan lalu lintas itu trjadi krn pengemudi tidak taat peraturan.Jgn protes dlu Sob	Not Relevant	-	-

written in a standard manner, so further normalization process is needed to produce exact location name. Second, the predicted location label are incorrect or the locator does not represent the entire location name. Third, the Tweet does not contain any location information. Last, the location is outside Jakarta, so it does not match the locator built from the Jakarta map.

Data visualization

We describes the results of developing a web-based dashboard that serves to visualize data. The main data source of this system is data modeling results that contain information related to disturbing traffic situations. The visualization is displayed spatially on the OpenStreetMap base map using JavaScript from the leaflet and tabularly. This system dashboard is designed for all traffic actors, both users and traffic organizers.

Figure 3 presents information related to current traffic flows. Users can see traffic disruption activities that are occurring on the existing monitoring map, where the traffic activity is displayed using the marker feature. The symbol of the marker represents the type of traffic event that is currently occurring. Users can access more detailed information by clicking on one of the markers on the map. This will bring up tweets related to the traffic event that have been shared by users on Twitter as well as the date, time, location, and category of the incident. Twitter data that is displayed spatially is Twitter data that has been linked to geographic information with geocoding. Meanwhile, tweets that have not been linked to geographic information are displayed in rows and sorted by the time at which they were posted. Table 5 shows examples of Tweets and extracted traffic information.

Conclusion

A daily average of 400–600 tweets related to traffic and road conditions are posted each day. Of that data, 39% is shared by the users who frequently share information related to traffic situations and road conditions, while the remaining 61% are shared by general users. Therefore, it can be concluded that data related to traffic activity and disturbing road conditions on Twitter are available and can be used for traffic monitoring.

The text mining models produced in this study are the classification model to filter out irrelevant data, the classification model to identify event type from relevant data, and the location information extraction model. The first classifier filters the data related to traffic situations and road conditions. The second classifier categorizes the relevant data into five categories of events, namely traffic jams, accidents, problems, repairs, and damage.

Based on the empirical evaluation, the best classifier in our experiment is Logistic Regression using feature union of count and TF-IDF, and applying case folding for data preprocessing the data. The classifier achieves the accuracy of 91% dan 94% for Tweet relevance prediction and traffic event type categorization tasks. On the other hand, location extraction model achieves F1-score in range of 45 – 54%. Only 68% of extracted location from Tweets can be interpreted as geolocation information through geocoding process. This research delivers a dashboard to visualize analyzed data spatially using geographic information from the results of geocoding and tweet location data.

Future work

For future work, we identify several work directions. The low accuracy of the location extraction model is because of lack of publicly available NER data set for Indonesian Twitter domain. We suggest the work on this resource for Indonesian language. On the other hand, most location mentions in Tweets are not written in standard language (e.g., abbreviated or misspelled), so the location normalization is a crucial task. The named-entity linking task for Twitter domain [22, 23] is underexplored for languages other than English.

Currently developed visualization dashboard only displays tweet data related to traffic activities spatially. Future studies may design and implement the dashboard integrated with existing activity monitoring systems developed by the government and the private sector. Other data sources can also be used, such as from other social media or online media, to enrich the information presented by the system. The more data related to the traffic situation the system receives, it is hoped, the better the quality of information users can expect.

The information on the developed system presented in this study is related to problems that occur on the road. Further information, such as alternative routes or preferred modes of transportation, can be provided in future research. Some geospatial analyses such as [24] can be applied to produce more information for the user. Data veracity should be assessed to avoid the extracted information contain fake news and misinformation [25, 26]. This analysis provides additional information regarding traffic predictions. Analysis in the area and around the location where traffic problems occur using several methods of spatial analysis, such as [27]. Multistage classification can also provide additional information, such as congestion repeats.

Acknowledgements
Not applicable.

Author contributions
PKP writing-original draft, data collection, code implementation, design the experiment, perform the experiment, analyze the result. RM writing-editing, conceptualization, analyze the result. IB writing-editing, conceptualization, review and supervision, funding acquisition. All authors read and approved the final manuscript.

Funding
This work was supported by Universitas Indonesia through Grant "Publikasi Terindeks Internasional (PUTI) Q2: NKB-4060/UN2.RST/HKP.05.00/2020"

Availability of data and materials
We crawl tweets from Twitter. The Twitter data will not be shared publicly due to restrictions of the Twitter Developer Policy and Agreement.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Faculty of Computer Science, Universitas Indonesia, Kampus UI Depok, 16424 Depok, Indonesia. [2]National Research and Innovation Agency, Geostech Building, Puspiptek Serpong, Tangerang Selatan, Indonesia.

Received: 23 September 2021 Accepted: 2 May 2022
Published online: 18 May 2022

References

1. BPS: Statistik Transportasi Darat 2018, 2019.
2. Tomtom: TomTom Traffic Index Ranking. Technical report. 2019. https://www.tomtom.com/en_gb/traffic-index/ranking/.
3. D'Andrea E, Ducange P, Lazzerini B, Marcelloni F. Real-time detection of traffic from twitter stream analysis. IEEE Trans Intell Transp Syst. 2015;16(4):2269–83. https://doi.org/10.1109/TITS.2015.2404431.
4. Gutierrez C, Figuerias P, Oliveira P, Costa R, Jardim-Goncalves R. Twitter mining for traffic events detection. In: Proceedings of the 2015 Science and Information Conference, SAI 2015, 2015:371– 378. https://doi.org/10.1109/SAI.2015.7237170.
5. Gu Y, Qian Z, Chen F. From Twitter to detector: real-time traffic incident detection using social media data. Transp Res Part C Emerg Technol. 2016;67:321–42. https://doi.org/10.1016/j.trc.2016.02.011.
6. Zhang Z, He Q, Gao J, Ni M. A deep learning approach for detecting traffic accidents from social media data. Transp Res Part C Emerg Technol. 2018;86:580–96. https://doi.org/10.1016/j.trc.2017.11.027.
7. Herwanto GB, Prasetya Dewantara D. Traffic Condition Information Extraction from Twitter Data. Proceedings 2nd 2018 International Conference on Electrical Engineering and Informatics, ICELTICs, 2018: 95–100 . https://doi.org/10.1109/ICELTICS.2018.8548921.
8. Lubis YA. Analisis Biaya Kemacetan Kendaran di Jalan Setiabudi (Studi Kasus Depan Sekolah Yayasan Pendidikan Shafiyyatul Amaliyyah) (YPSA). Jurnal Warta Edisi. 2016; 48.
9. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22(3):276–82.
10. Aliyah Salsabila N, Ardhito Winatmoko Y, Akbar Septiandri A, Jamal A. Colloquial Indonesian Lexicon. Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018, 2019: 226–229 . https://doi.org/10.1109/IALP.2018.8629151.
11. Taufik N, Wicaksono AF, Adriani M. Named entity recognition on Indonesian microblog messages. Proceedings of the 2016 International Conference on Asian Language Processing, IALP 2016, 2017: 358–361 . https://doi.org/10.1109/IALP.2016.7876005.
12. Rachman V, Savitri S, Augustianti F, Mahendra R. Named entity recognition on indonesian twitter posts using long short-term memory networks. In: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2017, pp. 228–232. https://doi.org/10.1109/ICACSIS.2017.8355038.

13. Alfina I, Manurung R, Fanany MI. DBpedia Entities Expansion in Automatically Building Dataset for Indonesian NER. 2016. https://doi.org/10.1109/ICACSIS.2016.7872784.
14. Dinakaramani A, Rashel F, Luthfi A, Manurung R. Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014, 2014:66–69 . https://doi.org/10.1109/IALP.2014.6973519.
15. Zhang W, Gelernter J. Geocoding location expressions in Twitter messages: a preference learning method. J Spatial Inf Sci. 2014;9(2014):37–70. https://doi.org/10.5311/JOSIS.2014.9.170.
16. Bao Y, Quan C, Wang L, Ren F. The role of pre-processing in twitter sentiment analysis. In: International Conference on Intelligent Computing,2014: pp. 615– 624 . Springer.
17. Hidayatullah AF. The influence of stemming on Indonesian tweet sentiment analysis. Proc Electr Eng Comput Sci Inf. 2015;2(1):127–32.
18. Pradana AW, Hayaty M. The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on indonesian-language texts. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 2019:375–380 2019.
19. Delimayanti MK, Sari R, Laya M, Faisal MR, Naryanto RF, et al. The effect of pre-processing on the classification of twitter's flood disaster messages using support vector machine algorithm. In: 2020 3rd International Conference on Applied Engineering (ICAE), 2020: pp. 1– 6 . IEEE.
20. Mutiara AB, Wibowo EP, Santosa PI, et al. Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. J Big Data. 2021;8(1):1–16.
21. Wibowo HA,Prawiro TA, Ihsan M, Aji AF, Prasojo RE, Mahendra R, Fitriany S. Semi-supervised low-resource style transfer of indonesian informal to formal language with iterative forward-translation. In: 2020 International Conference on Asian Language Processing (IALP), 2020:pp. 310– 315 . https://doi.org/10.1109/IALP51396.2020.9310459.
22. Liu X, Li Y, Wu H, Zhou M, Wei F, Lu Y.Entity linking for tweets. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1304– 1311. Association for Computational Linguistics, Sofia, Bulgaria 2013. https://aclanthology.org/P13-1128.
23. Feng Y, Zarrinkalam F, Bagheri E, Fani H, Al-Obeidat FN. Entity linking of tweets based on dominant entity candidates. Soc Netw Anal Min. 2018;8:1–16.
24. Tian Y, Hu W, Du B, Hu S, Nie C, Zhang C. IQGA: a route selection method based on quantum genetic algorithm-toward urban traffic management under big data environment. World Wide Web. 2019;22(5):2129–51. https://doi.org/10.1007/s11280-018-0594-x.
25. Qazvinian V, Rosengren E, Radev DR, Mei Q. Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1589–1599. Association for Computational Linguistics, Edinburgh, Scotland, UK. 2011. https://aclanthology.org/D11-1147.
26. Buntain C, Golbeck J. Automatically identifying fake news in popular twitter threads. In: 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017: pp. 208– 215 . https://doi.org/10.1109/SmartCloud.2017.40.
27. Wischoff L, Ebner A, Rohling H, Lott M, Halfmann R. SOTIS-a self-organizing traffic information system. In: The 57th IEEE Semiannual Vehicular Technology Conference, 2003. VTC 2003-Spring, vol. 4, 2003: pp. 2442– 2446 IEEE.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.