Exploring and Comparing the Quality of Public Transportation Services Based on Social Media: Case Study of Greater Jakarta

Hendrik Marantino Napitupulu[a], Yasuhiro Shiomi[b], Ibnu Syabri[c]

[a] *Graduate Program in Regional and Urban Planning, School of Architecture, Planning and Policy Development, Institut Teknologi Bandung, Indonesia and Graduate School of Science and Engineering, Ritsumeikan University, Japan; E-mail: marantino_napitupulu@yahoo.com*
[b] *Graduate School of Science and Engineering, Ritsumeikan University, Japan; E-mail: shiomi@fc.ritsumei.ac.jp*
[c] *School of Architecture, Planning and Policy Development, Institut Teknologi Bandung, Bandung, Indonesia; E-mail: syabri@gmail.com*

Abstract: Personal vehicles are considered a major cause of air pollution and traffic jams in Jakarta. To alleviate these issues, the use of public transportation should be increased. The purpose of this study is to evaluate the quality of service by the two largest public transportation modes in Jakarta: Bus Rapid Transit (BRT) and CommuterLine (KCI). Sentiment analysis and topic modeling were applied to social media data to identify poor service and measures to encourage people to switch to public transportation. The results showed that most users still perceived BRT and KCI to have poor service. The biggest problem with BRT was private vehicles entering BRT lanes, whereas the problems with KCI were the frequency of service and waiting time. These results demonstrate that social media is a useful data source for transportation service providers and for local governments.

Keywords: Social Media, Sentiment Analysis, Topic Modeling, Latent Dirichlet Allocation, Logistic Regression

1. Introduction

Jakarta is the capital city of Indonesia and the most populated, with 11,063,324 inhabitants according to the 2019 population census and a growth rate of 1.19% per year. With a total area 662.33 km^2, the population density of the Special Area of the Capital (DKI) Jakarta was 16,704 people/km^2 in 2019. If the Thousand Islands were removed from the calculation, the population density of DKI Jakarta was 16,882 people/km^2. People traveling within Jakarta as well as from outside Jakarta into the city for work have greatly increased the traffic density, and traffic jams can occur anywhere and anytime. According to the Traffic Index, Jakarta is the 10th most congested city in the world, with a congestion level of 53% (Tomtom, 2020). Private vehicles are a major cause of congestion in Jakarta and make up 77% of the traffic, and public vehicles make up only 23% (Tst/Pmg, 2019). Private vehicles not only contribute to traffic jams, but they are also responsible for the poor air quality in Jakarta (Rds/Agt, 2019). According to the air quality index, Jakarta has the fourth worst air quality and pollution among major cities around the world (IQAir, 2020). In 2018, Jakarta had 139,080,216 motorized vehicles carrying passengers, comprising 16,440,987 passenger cars, 2,538,182 buses, and 120,101,047 motorbikes. This huge number of vehicles contrasts with a total length of 6,652,7 km of available road (national and province roads) covering an area of 46,426,531 m^2 (Statistics Indonesia, 2020).
The era of information technology has made it easier for public transportation providers to get closer to the public and encourage them to use public transportation modes. Apps can be installed on mobile phones that contain information on schedules, prices, and available routes.

1

In addition, providers have official accounts on social media that can be followed by users to find out the latest information regarding the modes they usually use. Besides getting the latest information, social media users can also interact with officers in charge, and they can mention the official ID of the public transportation provider when describing their experiences on a transportation mode. Their comments contain very meaningful information that can be extracted. For example, public transportation service providers can determine which of their services is often complained about or discussed by users. This social media activity can be used to analyze customer satisfaction with the quality of service provided by public transportation providers. In this study, data from social media were analyzed to identify poor service from major public transportation providers in Jakarta, by employing sentiment analysis and topic modeling. The results can be used to develop measures for encouraging people to switch to public transportation and potentially alleviate traffic congestion.

2. Literature Review

For public transportation, macro events such as accidents or breakdowns are commonly captured using hotlines and customer surveys. For micro events that affect service quality such as late arrival and crowded buses, the transportation industry can use social media applications such as Twitter and Facebook (Hoang et al., 2016). Collins et al. (2013) used Twitter to assess commuter satisfaction with Chicago's rail transit and showed that simple aggregation can be used to derive sentiment words. Congosto et al. (2015) and Liu et al. (2012) also used social media to evaluate train disturbances, the time of the disturbance, and the user profile. Hoang et al. (2016) used a sense-making engine on 140,000 Twitter messages related to a bus service to translate the unstructured content into meaningful structured data. The engine had three components: entity extraction, event-type classification, and sentiment mining. Entity extraction involves establishing different micro events in transportation that can be derived from data. Sentiment mining allocates a sentiment value (negative, neutral, or positive) to every micro event tweet. They found that the accuracy of the sentiment analysis was significantly increased when they used the knowledge domain (e.g., regular expressions for extracting entities) and domain-relevant labeled data. Tweet analysis has been demonstrated to be effective in some domains, such as automatic deviation detection to identify power outage events during Hurricane Irene on August 27, 2011 (Pimbert et al., 2004). Lately, attention has been focused on analyzing tweets for radicalization, terrorism, and hate speech (Burby, 2003; Brody et al., 2003).

2.1 Advantages of analyzing social media data

The literature review showed that social media data can potentially be used to develop models for estimating travel demand, managing operation, and long-term planning purposes. However, special caution is required because of the biases associated with social media data. Social media utilization is increasing because of the spread of tablets and smartphones. Thus, such data commonly come with geo-location information, which is valuable for transportation management, planning, and operation purposes (Rashidi et al., 2017).

2.2 Disadvantages of analyzing social media data

A major challenge with using social media data is the extraction of useful information, which requires employing advanced text and data mining methods. Another challenge is identifying suitable models for predicting travel demand at the individual level as opposed to the

aggregate and zone-based levels. A major concern with social media data is the individual-specific information that cannot be shared publicly unless the owner of the data consents (Smith et al., 2012).

3. Methodology

This study analyzed comments on the social media platform Twitter for impressions of the two largest public transportation modes in Jakarta: the bus rapid transit (BRT) system TransJakarta and the rail transit system CommuterLine (KCI). A mixed-methods approach was used to analyze customer satisfaction based on social media data. For the quantitative analysis, logistic regression and latent Dirichlet allocation (LDA) were used for sentiment analysis. For qualitative analysis, a topic model was used to interpret the words and topics. Figure 1 shows the analytical framework of the study.

3.1 Target

This study evaluated responses on social media to public transportation services in Greater Jakarta (i.e., KCI and BRT). The social media users indirectly represent passengers because they could be people who did not take public transportation but mentioned keywords. Greater Jakarta is a megapolis in Indonesia with nine administrative areas: DKI Jakarta, Tangerang City, Tangerang Regency, South Tangerang City, Depok City, Bogor City, Bogor Regency, Bekasi City, and Bekasi Regency.

3.1.1 Bus Rapid Transit (BRT)

TransJakarta is the first BRT system in Southeast and South Asia and has been operating since 2004. TransJakarta was designed as a public transportation mode to support activities in the densely populated capital. TransJakarta has the longest path in the world (251.2 km), and 260 bus stops spread across 13 corridors. It initially operated from 5 am–10 pm, and currently, it operates some corridors for 24 h. TransJakarta operates 1347 buses, which include single and articulated buses. TransJakarta is supported by 13 companies that operate fleets serving each corridor.

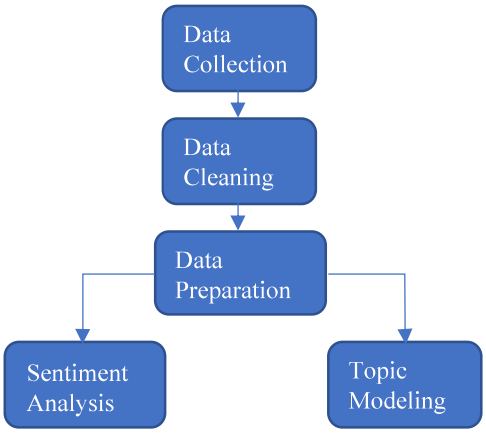


Figure 1. Analytical framework

3.1.2 CommuterLine (KCI)

On July 1, 2013, KCI implemented an electronic ticket system (E-Ticketing) and progressive tariff system as the next stage of its modernization. In December 2019, KCI had 1100 units, and it is continuing to grow. In 2019, KCI served an average of 979,853 users per day on weekdays with a record of 1,154,080 users on a single day. KCI currently serves 80 stations throughout Jabodetabek, Banten, and Cikarang with a route coverage of 418.5 km.

3.2 Sentiment Analysis

Sentiment analysis or opinion mining is focused on extracting people's opinions, sentiments, and emotions toward entities such as issues, services, and their attributes (Liu et al., 2012). Opinion mining extracts the attributes and components of an object from a set of text documents that contain opinions regarding it and determines whether the opinions are positive, negative, or neutral. This is typically performed through supervised machine learning. An algorithm can be developed to generate a mapping function that transforms the input variables (X) to the output variables (Y): $Y = f(X)$.

The mapping function needs to be accurately determined so that the output variable (Y) can be predicted from new input data (X). Supervised learning requires training data to train the algorithm and test data to determine the performance of the trained algorithm when fed new data (i.e., generalization). The trained algorithm is called a model.

We use data other than the data to be analyzed, as many as 1,828 tweets were labeled manually, divided into positive sentiment, negative sentiment, and neutral, where 80% of these data are used for training and other 20% used for testing. In this study, logistic regression was used for machine learning. Logistic regression is applicable when the dependent variable or target is categorical. In particular, Multinomial logistic regression was used, meaning that there were three or more categories without ordering (Khandelwal, 2020). Neutral sentiment are tweets outside the positives and negatives such as questions or information tweets. The confusion matrix is a table that is often used to describe the classification performance of a model when applied to a test dataset (Arora, 2019), and it was used in this study. The accuracy was evaluated according to the precision, recall, and F1 score. The precision is defined as the ratio between true and predicted positives. The recall is defined as the ratio between true and actual positives. When the recall is very high, the precision is very low and vice versa. The F1 score is the harmonic mean of the precision and recall and should be used to evaluate the performance of a model. Cross-validation is a technique used to validate the accuracy of a model that is built according to a specific dataset. K-fold cross-validation is a popular approach where the dataset is divided into K partitions of random size. Then, K experiments are carried out, where each experiment uses one partition as test data and the remaining partitions as training data.

3.3 Topic Modeling

Topic modeling is a type of machine learning where a set of documents is automatically analyzed to identify cluster words for the hidden semantic composition (Pascual, 2019). A topic comprises a group of words that often happen together. A topic model is a statistical approach to identifying abstract topics that occur in a set of documents. Unlike sentiment analysis, topic modeling is unsupervised machine learning, and it does not need a predefined list of labels or training data that have been manually prepared. The lack of training means that the data analysis is quick and easy. However, the precision and accuracy of the output are

not guaranteed, which is why many businesses invest time in training a topic classification model.

The LDA method was used in this study. The plate diagram is shown in Figure 2, where α and β denote the Dirichlet distribution, θ denotes the document-specific topic distribution, Z shows the topic assignment, φ denotes topics, and W denotes observed words. The term "latent" refers to information not previously known and hidden in the data, such as the theme or topic of a document. The Dirichlet distribution represents topics in documents and the distribution of words in topics. Once the Dirichlet distribution is obtained, topics are allocated to documents, and document words are allocated to topics.

LDA is a simple algorithm for topic modeling and works using Gibbs sampling. The joint probability distribution can be obtained by sampling each variable one by one based on the values of other variables (i.e., full conditional probability) (Darling, 2011).

3.4 Data Collection

Data were collected from Twitter. For KCI, all tweets mentioning the keywords "rekancommuters," "commuterline" (Official account of KAI Commuter) and "krlmania" were retrieved. For BRT, all tweets mentioning the keywords "transjakarta," "pt_transjakarta" (Official account of PT. Transportasi Jakarta) and "infobusway" were retrieved. These keywords were chosen because they comprised the official ID of the transportation provider and the user community IDs of each mode with the most followers.

Data were retrieved via the Twitter Application Programming Interface (API). A Twitter account is required to use the API. After the API is registered, four codes are provided: the consumer key, consumer secret, access token, and access token secret. Data were retrieved using the Tweepy library. The data used in this study included the user ID, comments, and timestamps. The four codes received earlier needed to be inputted in the data withdrawal script. Data withdrawal was performed every few days continuously. The collected data drawn were stored in JavaScript Object Notation (JSON) format, which is a lightweight data exchange format. It is easy to read and write for humans and easy to translate and generate for computers.

The data were collected for 6 months starting on October 20, 2019, and ending on April 30, 2020. In total, 150,569 tweets were collected. Figures 3 and 4 show the daily number of tweets on BRT and KCI, respectively. For the sentiment analysis, 1828 tweets were randomly chosen and manually labeled as positive, negative, or neutral for supervised learning. Once the training data were generated, the model was tested and validated. For the topic modeling, the test data were prepared separately from the data used for sentiment analysis. Of all tweets, 80% were used for training and validation, and the remaining 20% were used for testing. Of the 80% used for training and validation, 70% was used for training, and 30% was used for validation.

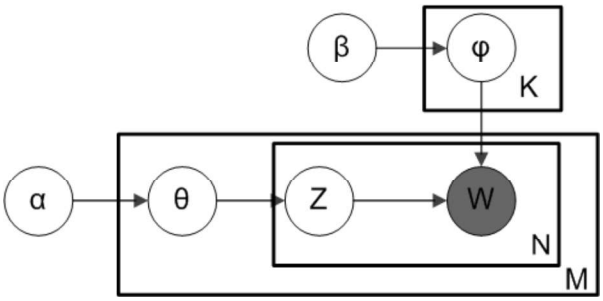


Figure 2. Plate diagram for the LDA model

5

The number of tweets each day varied depending on the problems faced by users and questions asked or negative comments made by users. The number of tweets was also affected by the day and time. For BRT, the following peaks were observed during the six months of data retrieval as shown in Figure 3:

1. March 26, 2020–Routes that previously closed because of COVID-19 were reopened
2. March 16, 2020–COVID-19 pandemic
3. February 25, 2020–Flood
4. January 2, 2020–Flood.

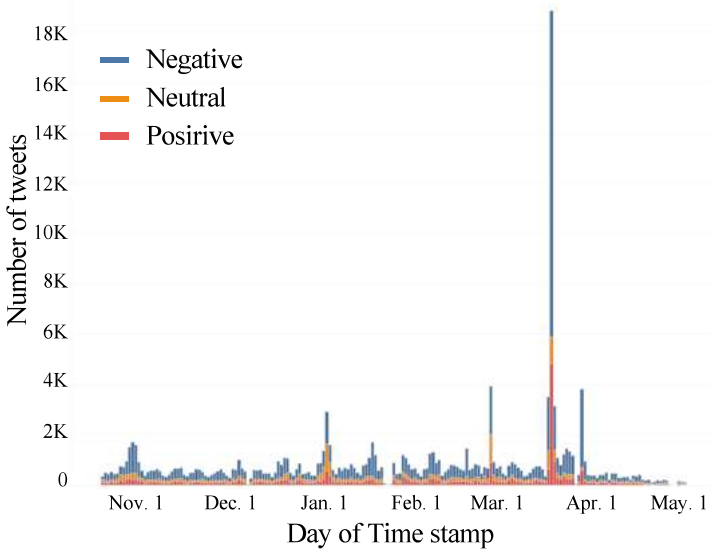


Figure 3. Number of daily tweets on BRT

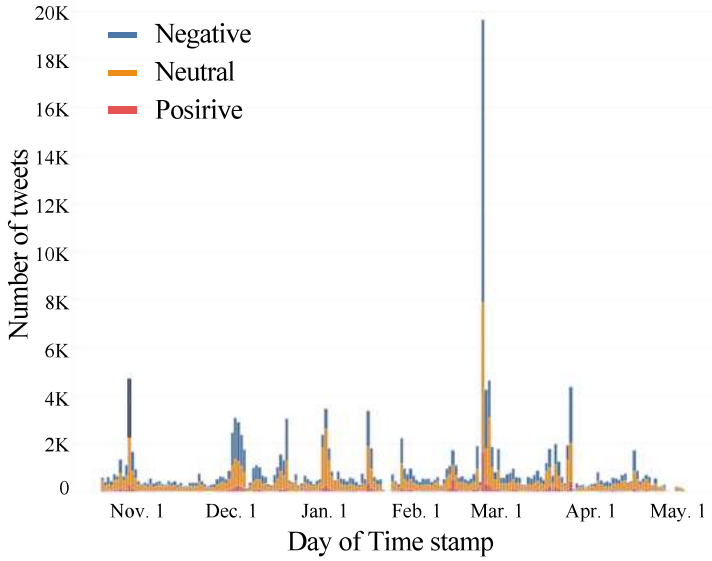


Figure 4. Number of daily tweets on KCI

For KCI, the following peaks were observed:

1. March 26, 2020–Flood

6

2. March 23, 2020–Normal schedule (with a few adjustments)
3. March 16, 2020–COVID-19 pandemic
4. February 23, 2020 –Lever/switch replacement, made worse by flood
5. January 16, 2020 –Improvements in upper flow electricity (LAA)
6. October 29, 2019 –Community fights.

4. Results and Discussion

4.1 Results of Sentiment Analysis

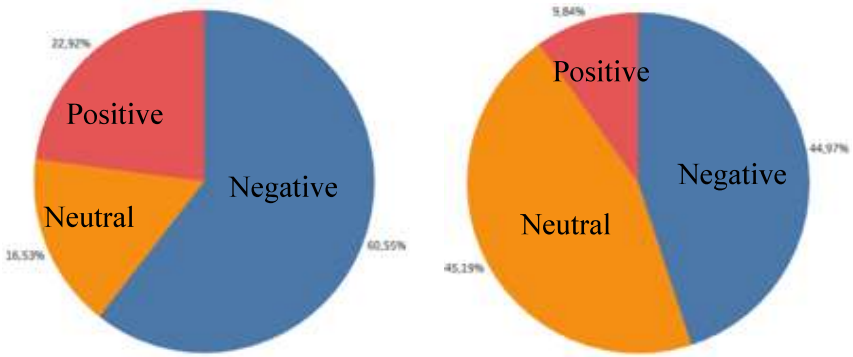
Table 1 presents the classification and accuracy results for the sentiment analysis. Of the 1828 tweets that were labeled manually, 20% (i.e., 366 tweets) were used for testing. The developed model was then applied to all 150,569 tweets to predict the sentiment. The model achieved a precision of 71% at finding negative sentiments. The K-fold cross-validation resulted in an accuracy of 74% (±0.03%).
Based on proportions of sentiment as shown in Figure 5, the following is come for the given

Table 1. Modeling results
(a) Classification results for test data

		Prediction			
		Positive	Neutral	Negative	Total
Observation	Negative	97	21	11	129
	Neutral	22	99	12	133
	Positive	17	8	79	104
	Total	136	128	102	366

(b) Accuracy indices

	Precision	Recall	F1 score	Support
Negative	0.71	0.75	0.73	129
Neutral	0.77	0.74	0.76	133
Positive	0.77	0.76	0.77	104
Accuracy			0.75	366
Macro avg	0.75	0.75	0.75	366
Weighted avg	0.75	0.75	0.75	366



(a) BRT (b) KCI

Figure 5. Sentiment percentages

7

dataset. The proportions of positive tweets and negative tweets are higher for BRT than KCI, and the proportion of neutral tweets for KCI are higher than BRT. Because positive tweets would generally follow negative tweets after accidents get recovered, this finding implies that for BRT negative events tend to happen more frequently than KCI.

Because this study was focused on the quality of public transportation services, although the positive sentiments should not be ignored, the negative sentiments were focused on because they can be used to find complaints. These complaints can be used to determine where services need to be improved.

Tables 2 and 3 summarize keywords indicating negative sentiments toward BRT and KCI, respectively. An example of a negative sentiment toward BRT is the tweet

"Halooooo @PT_TransJakarta gimana neh masa mobil pribadi boleh masuk jalur khusus bus transjakarta? \n\nmau menyalahkan supir mobil pribadi atau petugas yang jaga di jalur transjakarta ya?".

(*"Hello, @PT_TransJakarta, how can private cars enter the special Transjakarta bus lane? Who do you want to blame the private car driver or the officer guarding the Transjakarta route?"*)

Similarly, an example of a negative sentiment toward KCI is

"@CommuterLine min, kereta ketahan lama bgt masuk manggarai, ada apa ya?"

(*"@CommuterLine Min, the old train entered Manggarai station. What's wrong?"*)

Although BRT and KCI had many of the same problems, BRT had some specific issues such

Table 2. Negative words for BRT

Indonesian	English
(Pe)tugas	Officer
Masuk	Enter
Tunggu	Wait
(Pe)Tumpang	Passengers
Antri	Queue
(Ke)Bijak(an)	Policy
Koridor	Corridor
Armada	Bus fleet

Table 3. Negative words for KCI

Indonesian	English
(Pe)Tumpang	Passengers
Masuk	Enter
Benar	True
Pakai	Use
Gerbong	Cart
Lambat	Slow
Tunggu	Wait
Jadwal	Schedule

as queues, drivers, and traffic jams. Many tweets were regarding queuing problems, which occurred because there were still many small bus stops that were unable to accommodate the number of users. For KCI, many of the complaints were associated with the few number of carts, slow service, and waiting time. Many users complained regarding the few carts and recommended adding more, especially during peak hours. Trains that frequently ran slowly did not escape comments. Additionally, KCI trains often waited for long-distance trains to pass because the two services have to share the same track, which generated complaints.

4.2 Results of Topic Modeling

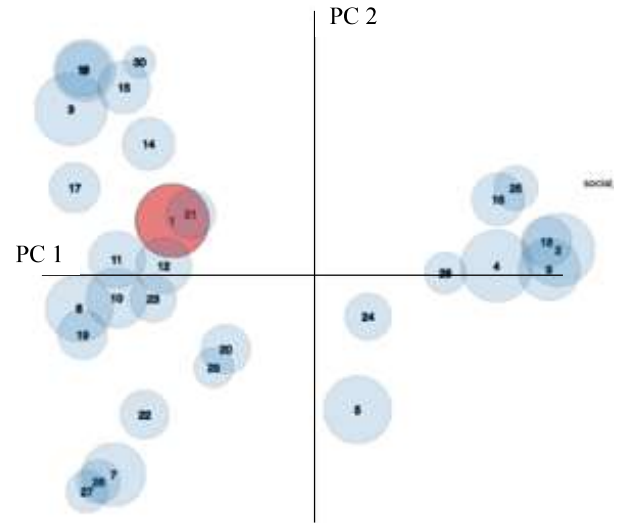
For topic modeling, 30 topics were created according to the highest coherence value, which was 0.500 for BRT and 0.476 for KCI. Topic Coherence measures the score of a topic by measuring the degree of semantic similarity between high-scoring words in that topic. Some topics overlap each other means they have some words in common. The coherence of a topic is defined by its ease of interpretation by the authors. The extracted words in 30 topics for BRT and KCI are summarized in Appendix 1 and Appendix 2, respectively. X and Y axis named as principle components PC1 and PC2 respectively, it's dimension reduction by using several lines or planes, so that it can make it easier for us to interpret the data and see the distribution of data into several clusters (but cluster division is not the main goal).
The 30 topics can be grouped according to similarity and type. For BRT, the 30 topics can be summarized by the following topic groups:

1. COVID-19 and policy
2. Politics
3. Mobile application
4. BRT lane trespass
5. Trending news
6. Integrating mode
7. On the bus violations
8. Schedules and routes
9. Drivers and officers.

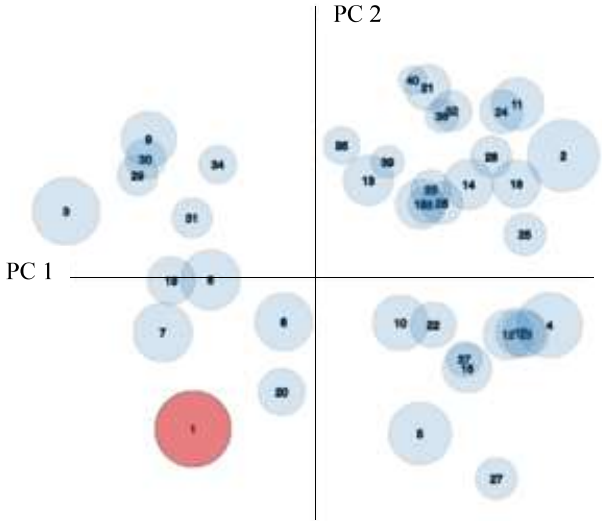
For KCI, the 30 topics were divided into more topic groups than those for BRT:

1. COVID-19 and policy
2. Conditions inside the cart
3. Trending news
4. Politics
5. Pregnancy pin
6. Officers
7. Schedules and routes
8. Online transportation
9. Disturbances
10. Lost
11. Airport train
12. Payment
13. Transportation information apps
14. Private vehicle.

Unique topic groups included BRT are "BRT lane trespass" and "integrating mode". A major problem for BRT is the large number of lane violations, where many private vehicles trespass on the lane meant for TransJakarta. The integrating mode relates to the expansion of TransJakarta to micro/mini transport and their integration with medium rail transit (MRT) and



(a) BRT



(b) KCI

Figure 6. Intertopic distance map

light rail transit (LRT). This intended to encourage private vehicle users to switch to public transportation.

Several topics unique to KCI are “pregnancy pin”, “online transportation”, “disturbances”, “lost”, “airport train”, and “payment”. The pregnancy pin relates to a new program to prioritize expectant mothers. Online transportation is very often installed at all stations waiting for users who want to change modes. Technical problems that often occur include signal failure and track circuit failure. Disruptions with the payment machine also generate complaints. Users also commonly lose or misplace their belongings. Finally, the airport train, which has only been operating for a few years, is in great demand. Previously, apart from taxis, the only public transportation services to the airport were DAMRI buses with varying costs depending on the region.

Figure 6 shows the inter-topic distance maps. The blue circles represent topics and their distance from other topics, and the intersections indicate similarity. The more widespread the

topics that are formed are, the more varied are the topics that are being discussed. For BRT, the topics were unevenly distributed. For KCI, the topics were more evenly distributed, which indicates that more topics were being discussed.

4.3 Discussion

As shown in Figure 3 and 4, the number of tweets varies depending on events, that is, if something happens, the number of negative tweets drastically gets higher. Based on the logistic regression model, negative tweets can be automatically and accurately identified. In addition, it is proved that we can automatically guess and classify the topics of tweets by using Topic model. Based on these methods, we can detect events that occur around the service area of BRT and KCI, and judge if the events are negative or positive by just monitoring tweets with specific hashtags automatically. Under the incident situation, it is essential to provide information as quick as possible to relief the users and promote them to change the behavior. Conventionally, this process needs a lot of manpower and take time and costs. The proposed methods can achieve this process with no manpower and less expensive costs, which can improve the convenience of the use of public transportation and may result in encouraging people to use public transportation. Even though the keyword is for public transportation, but the discussed topics are not always about transportation, as we can see that many Twitter users talked about politics and trending news. From many negative sentiments about public transportation services in Jakarta, it turns out that there are still those who get good appreciation from passengers, such as pins for pregnant women on KCI services or feeder services with micro trans from BRT transportation.

5. Conclusions

In this study, data from social media were analyzed to identify poor service from major public transportation providers in Jakarta, by employing sentiment analysis and topic modeling. First, we collected the tweets data relating to BRT and KCI by using API. Then, we confirmed that the daily number of tweets on BRT and KCI varies according to the special events. By employing the semantic analysis, it is shown that negative tweets can be correctly identified. In addition, it is revealed that we can guess the topic of each tweet by using topic model. Finally, it is concluded that the social media data is useful to improve the convenience of the users of public transportation particularly under the negative situations with incidents.

However, the use of social media has not been able to completely replace the role of manual surveys, especially regarding customer satisfaction. One reason is that many social media users use satirical sentences. In addition, it is still difficult to identify bots and buzzer accounts. For now, social media data are more suitable for supplementing manual surveys because their use still needs further technological development. Although the elderly are underrepresented on social media, some tweets suggest security to help senior citizens left behind by their partners or to add facilities for the mobility-impaired.

The results of this study can be used to help transportation service providers more easily understand the needs of users and identify the problems with their services. The results can help local governments find the response of the public to their policies.

Further research may involve expanding the scope to other transportation modes for a longer study period. The results can be improved by improving the data cleaning process and

naturalization step or by increasing the amount of trained data. The sentiment analysis can be connected to topic modeling, where each topic is rated according to the public sentiment.

Appendix 1. The 10 words composing 30 topics for BRT.

(1) Topic 1

Indonesian	English
Bijak	Policy
Perintah	Command
Batas	Border
Libur	Holiday
Gubernur	Governor
Wfh	Wfh
Kurang	Less/lack
Masuk	Enter
Anies	Jakarta Gov.
Lockdown	Lockdown

(2) Topic 2

Indonesian	English
Corona	Corona
Virus	Virus
COVID	COVID
MRT	MRT
Sebar	Spread
Cegah	Prevent
Antri	Queue
Tular	Contagious
Batas	Border
Kurang	Less/lack

(3) Topic 3

Indonesian	English
Ppd	Government bus company
Panas	Hot
Bocor	Leak
Dingin	Cold
Armada	Fleet
Hujan	Rain
Pakai	Use
Koridor	Corridor
Tumpang	Passengers
Nyaman	Comfortable

(4) Topic 4

Indonesian	English
Waras	Healthy
Ahok	Ex.Jakarta Gov
Kasus	Case
Anies	Jakarta Gov
Korupsi	Corruption
Gubernur	Governor
Reklamasi	Reclamation
Jokowi	Ind. President
Banjir	Flood
Cengkareng	Area name

(5) Topic 5

Indonesian	English
Tugas	Task
Prioritas	Priority
Duduk	Sit
Tumpang	Passengers
Kursi	Seat
Lansia	Seniors Citizen
Plb	Officers
Diri	Stand
Hamil	Pregnant
Anak	Child

(6) Topic 6

Indonesian	English
Kereta	Train
Ojol	Online trans.
Pakai	Use
Macet	Queue
Gojek	Online trans.
MRT	MRT
Angkot	Microtrans
Motor	Motorcycle
Mobil	Car
Bawa	Bring

(7) Topic 7

Indonesian	English
Tunggu	Wait
Waktu	Time
Armada	Fleet
Tumpang	Passengers
Rute	Route
Berangkat	Departure
Jadwal	Schedule
Penuh	Full
Koridor	Corridor
Lambat	Slow/late

(8) Topic 8

Indonesian	English
Tugas	Officers
Tumpang	Passengers
Pintu	Door
Supir	Driver
Masuk	Enter
Jaga	Keep/watch
Turun	Get off
Tutup	Close
Plb	Officers
Atur	Regulations

(9) Topic 9

Indonesian	English
Kereta	Train
Anak	Child
Makan	Eat
Pakai	Use
Duduk	Sit
Tunggu	Wait
Tidur	Sleep
Diri	Stand up
Habis	Empty
Kampus	Campus

(10) Topic 10

Indonesian	English
Trafi	Public trans. App.
Aplikasi	Application
Tunggu	Wait
Arah	Direction
Rute	Route
Pondok	Area name
Error	Error
Layar	Screen
Jadwal	Schedule
Gps	GPS

(11) Topic 11

Indonesian	English
Dirut	Gen. Director
Anies	Jakarta Gov.
Tipu	Trick/gimmick
Angkat	Lift/raise
Direktur	Director
Batal	Canceled
Kasus	Case
Senonoh	Profanity
Gubernur	Governor
Pidana	Criminal

(12) Topic 12

Indonesian	English
Gaji	Salary
Makan	Eat
Bayar	Pay
Anak	Child
Pakai	Use
Libur	Holiday
Potong	Cut
Masuk	Enter
Magang	Internship
Habis	Empty

(13) Topic 13

Indonesian	English
Tumpang	Passengers
Tunggu	Wait
Antri	Queue
Arah	Direction
Penuh	Full
Masuk	Enter
Armada	Fleet
Tumpuk	Pile up
Blok	Area name
Koridor	Corridor

(14) Topic 14

Indonesian	English
Motor	Motorcycle
Mobil	Car
Macet	Jammed
Masuk	Enter
Kendara	Vehicles
Tabrak	Hit
Supir	Driver
Pribadi	Private
Lintas	Cross
Langgar	Break (the rules)

(15) Topic 15

Indonesian	English
Kartu	Card
Pakai	Use
Tap	Tap
Bayar	Pay
Jaklingko	Jak microtrans
Jak_lingko	Jak microtrans
Mesin	Engine
Saldo	balance
Cash	Cash
Top	Top (up)

(16) Topic 16

Indonesian	English
Duduk	Sit
Wanita	Female
Kursi	Seat
Pria	Male
Tugas	Officers
Tumpang	Passengers
Diri	Stand
Prioritas	Priority
Area	Area
Kosong	Empty

(17) Topic 17

Indonesian	English
Operasi	Operation
Harmoni	Area name
Rute	Route
Koridor	Corridor
Arah	Direction
Lebak_bulus	Area name
Kalideres	Area name
Senen	Area name
Pasar	Area name
Rambutan	Area name

(18) Topic 18

Indonesian	English
Stasiun	Station
Rute	Route
Arah	Direction
Turun	Get off
Transit	Transit
Tebet	Area name
Blok	Area name
Pasar	Area name
Tuju	Destination
Jurus	Direction/line

(19) Topic 19

Indonesian	English
Jak	Jak(arta)
Jaklingko	Jak microtrans
Operasi	Operation
Angkot	Microtrans
Henti	Stop
Rute	Route
Stop	Stop
Jak_lingko	Jak microtrans
Tunggu	Wait
Mikrotrans	Microtrans

(20) Topic 20

Indonesian	English
Tugas	Officers
Layan	Service
Tumpang	Passengers
Plb	Officers
Benar	Very/true
Langgan	Customer
Kurang	Less/lack
Lapang	Spacious
Lapor	Report
Bantu	Help

(21) Topic 21

Indonesian	English
Rute	Route
Operasi	Operation
Arah	Direction
Poris	Area name
Juanda	Area name
Senayan	Area name
Grogol	Area name
Gading	Area name
Bundar_senayan	Area name
Stasiun	Station

(22) Topic 22

Indonesian	English
Rute	Route
Operasi	Operation
Arah	Direction
Ragunan	Zoo
Koridor	Corridor
Blok	Area name
Tosari	Area name
Dukuh	Area name
Tunggu	Wait
Ciledug	Area name

(23) Topic 23

Indonesian	English
Anies	Jakarta Gov.
Sehat	Healthy
Gubernur	Governor
Medis	Medic
Lindung	Protection
Pimpin	Leader
Jaga	Keep/watch
Mantap	Good
Tugas	Officers
COVID	COVID

(24) Topic 24

Indonesian	English
MRT	MRT
LRT	LRT
Kereta	Train
Integrasi	Integration
Stasiun	Station
Bangun	Build
Publik	Public
Macet	Jammed
Angkut	Transport
Trotoar	Sidewalk

(25) Topic 25

Indonesian	English
Operasi	Operation
Rute	Route
Koridor	Corridor
Banjir	Flood
Pinang_ranti	Area name
Jambar	Area name
Pluit	Area name
Grogol	Area name
Alih	Diversion
Bsd	Area name

(26) Topic 26

Indonesian	English
Pakai	Use
Masker	Mask
Tugas	Officers
Sedia	Available
Bersih	Clean
Tumpang	Passengers
Hand_sanitizer	Hand sanitizer
Masuk	Enter
Warna	Color
Bawa	Bring/carry

(27) Topic 27

Indonesian	English
Operasi	Operation
Rute	Route
Cibubur	Area name
Blok	Area name
Bkn	Area name
Royal	Premium bus
Summarecon	Area name
Jadwal	Schedule
Royaltrans	Premium bus
Kuningan	Area name

(28) Topic 28	
Indonesian	English
Supir	Driver
Bawa	Bring/carry
Mobil	Car
Tumpang	Passengers
Kiri	Left
Kendara	Vehicles
Kanan	Right
Motor	Motorcycle
Ppd	Gvmnt. bus company
Bahaya	Danger

(29) Topic 29	
Indonesian	English
Tumpang	Passengers
Gendong	Carrying
Bripka_sigit	Police name
Serang_jantung	Heart attack
Ombudsman_aksi	Ombudsman
Prabowo_gendong	Name_carrying
Heroik_personel	Heroic personnel
Satlantas_polda	Regional police
Metro_jaya	Jakarta Regional Police
Ombudsman	Ombudsman

(30) Topic 30	
Indonesian	English
Tumpang	Passengers
Armada	Fleet
Batas	Border
Bijak	Policy
Kurang	Less/lack
Antri	Queue
Operasi	Operation
Tumpuk	Pile up
Tumpu	Fulcrum
Social_distancing	Social distancing

Appendix 2. The 10 words composing 30 topics for KCI.

(1) Topic 1	
Indonesian	English
Masker	Mask
Pakai	Use
Tugas	Officers
Corona	Corona
Tumpang	Passengers
Hand_sanitizer	Hand sanitizer
Masuk	Enter
Sedia	Available
Sehat	Healthy
Virus	Virus

(2) Topic 2	
Indonesian	English
Gerbong	Cart
Panas	Hot
Penuh	Full
Wanita	Female
Masuk	Enter
Rangkai	Series
Dingin	Cold
Tumpang	Passengers
Bau	Smell
Nomor	Number

(3) Topic 3	
Indonesian	English
Tawur	Brawl
Sabar	Patient
Ganggu	Disturbance
Lelah	Tired
Tahan	Restrained
Rui	
Lambat	Slow/late
Benar	Very/true
Habis	Empty
Cepat	Fast

(4) Topic 4	
Indonesian	English
Pasar	Area name
Senen	Area name
Via	Via
Jadwal	Schedule
Berangkat	Departure
Tuju	Destination
Arah	Direction
Rute	Route
Henti	Stop
Transit	Transit

(5) Topic 5	
Indonesian	English
Train	Train
Day	Day
Time	Time
Station	Station
Don	Do not
Pakai	Use
Better	Better
Know	Know
Back	Back
Again	Again

(6) Topic 6	
Indonesian	English
Jadwal	Schedule
Berangkat	Departure
Operasi	Operation
Rangkas	Area name
Rangkasbitung	Area name
Tuju	Destination
Palmerah	Area name
Rute	Route
Kebayoran	Area name
Arah	Direction

(7) Topic 7	
Indonesian	English
Jadwal	Schedule
Aplikasi	Application
Access	Access
Update	Update
Buka	Open
Error	Error
Akses	Access
Ubah	Change
Download	Download
Benar	Very/true

(8) Topic 8	
Indonesian	English
Tahan	Detained
Masuk	Enter
Tunggu	Wait
Lambat	Slow/late
Arah	Direction
Gambir	Area name
Ganggu	Disturbance
Ganti	Change
Tuju	Destination
Juanda	Area name

(9) Topic 9	
Indonesian	English
Jadwal	Schedule
Layan	Service
Lambat	Slow/late
Evaluasi	Evaluation
Transportasi	Transportation
Tumpang	Passengers
Saran	Suggestion
Masalah	Problem
Ganti	Change
Template	Template

(10) Topic 10

Indonesian	English
Pakai	Use
Masinis	Machinist
Ganggu	Disturbance
Kencang	Loud/tight
Volume	Volume
Gerbong	Cart
Speaker	Speaker
Jazz	Jazz
Berisik	Noisy
Keras	Loud/hard

(11) Topic 11

Indonesian	English
Pin_hamil	Pregnant pin
Respon	Respond
Tugas	Officers
Cepat	Fast
Pin	Pin
Bantu	Help
Habis	Empty
Daftar	List/register
Sehat	Healthy
Balas	Reply/response

(12) Topic 12

Indonesian	English
Pintu	Door
Peron	Platform
Masuk	Enter
Tutup	Close
Turun	Get off
Tumpang	Passengers
Eskalator	Escalator
Buka	Open
Tangga	Stairs
Tugas	Officers

(13) Topic 13

Indonesian	English
Jadwal	Schedule
Tumpang	Passengers
Ubah	Change
Tumpuk	Pile
Tunggu	Wait
Berangkat	Departure
Lambat	Slow/late
Penuh	Full
Benar	Very/true
Arah	Direction

(14) Topic 14

Indonesian	English
Pakai	Use
Haram	Forbidden
Headset	Headset
Onta	Camel
Ganggu	Disturbance
Muslim	Moslem
Masuk	Enter
Kafir	Unbeliever
Handphone	Handphone
Ribet	Complicated

(15) Topic 15

Indonesian	English
Operasi	Operation
Rute	Route
Aman	Secure
Banjir	Flood
Ganggu	Disturbance
Arah	Direction
Lancar	Smooth
Rawa_buaya	Area name
Genang	Puddle
Hujan	Rain

(16) Topic 16

Indonesian	English
Operasi	Operation
Benar	Very/true
Batas	Border
Perintah	Order
Transportasi	Transportation
Masuk	Enter
Psbb	Social restrictions
Wfh	Work from home
Henti	Stop
Bijak	Policy

(17) Topic 17

Indonesian	English
Duduk	Sit
Kursi	Seat
Diri	Stand
Prioritas	Priority
Hamil	Pregnant
Tugas	Officers
Tumpang	Passengers
Anak	Child
Gerbong	Cart
Lelah	Tired

(18) Topic 18

Indonesian	English
Video	Video
Benar	Very/true
Viral	Viral
Hapus	Erase
Lapor	Report
Bantu	Help
Hoax	Hoax
Izin	Permission
Medsos	SNS
Takut	Fear

(19) Topic 19

Indonesian	English
Voucher_irit	Eco voucher
Gojek	Online trans.
Voucher	Voucher
Ride_diskon	Discount
Voucher_gojek	Online trans. Voucher
Pakai	Use
Hemat	Economical
Gojek_pastiadajalan	Online trans.
Ojol	Online trans.
Busway	BRT

(20) Topic 20

Indonesian	English
Jadwal	Schedule
Berangkat	Departure
Tuju	Destination
Angke	Area name
Arah	Direction
Operasi	Operation
Rute	Route
Tebet	Area name
Sedia	Available
Ubah	Change

(21) Topic 21

Indonesian	English
Hilang	Lost
Bantu	Help
Handphone	Handphone
Pakai	Use
Bawa	Bring/carry
Temu	Found
Gerbong	Cart
Warna	Color
Turun	Get off
Tugas	Officers

(22) Topic 22

Indonesian	English
Satpam	Security
Tugas	Officers
Tindak	Act on
Lapor	Report
Aman	Secure
Diam	Silent/quiet
Wanita	Female
Kurang	Less/lack
Sigap	Spry
Benar	Very/true

(23) Topic 23

Indonesian	English
Tumpang	Passengers
Batas	Border
Kurang	Less/lack
Jarak	Distance
Social_distancing	Soc distancing
Operasi	Operation
Jaga	Keep/watch
Jadwal	Schedule
Bijak	Policy
Tumpuk	Pile up

(24) Topic 24

Indonesian	English
Turun	Get off
Arah	Direction
Bus	Bus
Tuju	Destination
Bandara	Airport
MRT	MRT
Transit	Transit
Busway	BRT
Halte	Bus stop
Rute	Route

(25) Topic 25

Indonesian	English
Jadwal	Schedule
Lambat	Slow/late
Berangkat	Departure
Tunggu	Wait
Waktu	Time
Ubah	Change
Benar	Very/true
Ganti	Change
Masuk	Enter
Parah	Severe

(26) Topic 26

Indonesian	English
Hijab	Veil
Bawa	Bring/carry
Artis	Artist
Pakai	Use
Cebol	Midget
Koar	Much talking
Pilpres	President elect
Jokowi	Ind. President
Menang	Win
Musisi	Musician

(27) Topic 27

Indonesian	English
Transportasi	Transportation
Motor	Motorcycle
Parkir	Park
Pakai	Use
Mobil	Car
Bawa	Bring/carry
Macet	Jammed
Publik	Public
Kendaraan_pribadi	Pvt. vehicle
Ojol	Online trans.

(28) Topic 28

Indonesian	English
Tua	Old
Anak	Child
Muda	Young
Benar	Very/true
Sebal	Annoyed
Sabar	Patient
Pakai	Use
Diri	Stand
Kasihan	Pity
Marah	Angry

(29) Topic 29

Indonesian	English
Pakai	Use
Kartu	Card
Tap	Tap
Tiket	Ticket
Bayar	Pay
Kmt	Subscription card
Saldo	Balance
Gate	Gate
Linkaja	Payment App.
Tap_out	Tap out

(30) Topic 30

Indonesian	English
Gerbong	Cart
Wanita	Female
Tugas	Officers
Makan	Eat
Tumpang	Passengers
Tegur	Warning
Duduk	Sit
Pria	Male
Atur	Arrange
Ganggu	Disturbance

REFERENCES

Tomtom. (2020). Traffic Index 2019. Retrieved from https://www.tomtom.com/en_gb/traffic-index/ranking/.

Tst/Pmg. (2019), Anies Salahkan Pengguna Kendaraan Pribadi Biang Kemacetan. CNN Indonesia News. Retrieved from https://www.cnnindonesia.com/nasional/20191205123642-20-454410/anies-salahkan-pengguna-kendaraan-pribadi-biang-kemacetan.

Rds/Agt. (2019), Udara Jakarta Terburuk, Pemprov Salahkan Kendaraan Pribadi. CNN Indonesia News. Retrieved from https://www.cnnindonesia.com/nasional/20190714114256-20-411948/udara-jakarta-terburuk-pemprov-salahkan-kendaraan-pribadi.

IQAir. (2020), Air quality in Jakarta. IQAir Web Site, Retrieved from https://www.iqair.com/indonesia/jakarta.

Darling, William M. (2011), A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. Proceedings of the 49th Annual Meeting of the Associationfor Computational Linguistics: Human Language Technologies, pp. 642-647.

Division of Integration Processing and Statistics Dissemination BPS-Statistics of DKI Jakarta Province. (2020). DKI Jakarta Province in Figures. DKI Jakarta: BPS-Statistics of DKI Jakarta Province.

Hoang, T., P. H. Cher, P. K. Prasetyo, and E. P. Lim. (2016). Crowdsensing and analyzing micro-event tweets for public transportation insights. In Proceedings - 2016 IEEE International Conference on Big Data. Big Data 2016.

Congosto, M., D. Fuentes-Lorenzo, and L. Sánchez. (2015). Microbloggers as sensors for public transport breakdowns. IEEE Internet Comput., vol. 19, no. 6, pp. 18–25.

Liu, W., F. Al Zamal, and D. Ruths. (2012). Using social media to infer gender composition of commuter populations. In Proceedings of the when the city meets the citizen workshop, the international conference on weblogs and social media.

Collins, C., S. Hasan, and S. V Ukkusuri. (2013). A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. J. Public Transp., vol. 16, no. 2, p. 2.

Pascual, Federico. (2019). Introduction to Topic Modeling. Blog, MonkeyLearn, https://monkeylearn.com/blog/introduction-to-topic-modeling/.

Pimbert, M. P. (2004). Institutionalising participation and people-centered processes in natural resource management. IIED.

Burby, R. J. (2003). Making plans that matter: Citizen involvement and government action. J. Am. Plan. Assoc., vol. 69, no. 1, pp. 33–49.

Brody, S. D., D. R. Godschalk, and R. J. Burby. (2003). Mandating citizen participation in plan making: Six strategic planning choices. J. Am. Plan. Assoc., vol. 69, no. 3, pp. 245–264.

Rashidi, T. H., A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transp. Res. Part C Emerg. Technol. doi: 10.1016/j.trc.2016.12.008.

Smith, M., C. Szongott, B. Henne, and G. Von Voigt. (2012). Big data privacy issues in public social media. In 2012 6th IEEE international conference on digital ecosystems and technologies (DEST), pp. 1–6.