The Semantic Analysis of Twitter Data with Generative Lexicon for the Information of Traffic Congestion

Subhan[1,*], Eko Sediyono[2], Farikhin[3]

[1] Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia
[2] Faculty of Information Technology, Universitas Kristen Satya Wacana, Salatiga, Indonesia
[3] Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia
*Corresponding author: subhan@mail.unnes.ac.id

ARTICLE INFO

Article history
Received 12 August 2019
Revised 6 September 2019
Accepted 3 October 2019

Keywords
Semantics analysis
Twitter
Generative lexicon

ABSTRACT

This research is closely related to the semantic analysis of Twitter Data with Generative Lexicon for getting information of traffic congestion. This research aims to generate the semantic analysis with Generative Lexicon to obtain structured information about traffic congestion conditions. Semantic analysis is conducted through several stages, namely data acquisition, text segmentation, detection of types and meanings of the words, and (4) semantic analysis. The results of this research, the system can determine the congestion conditions based on the semantic analysis. The system also separates the data of place and time of occurrence of tweets on Twitter.

1. Introduction

Nowadays, microblogging has become popular as a communication medium among internet users including Twitter. In March 2015, Twitter has had more than 320 million users worldwide and every day it gets more than 500 million tweets sent by users (Twitter, 2015). From the data, Indonesia becomes a country having 50 million Twitter accounts (Kompas, 2015).

Social media users increasingly want to share their personal opinions with others on social media, such as product reviews, economic analysis, political polls, and so on. There is a growing interest in finding out people's opinions or submissiveness towards some objects from social media, which can help other users to make decisions and get valuable feedbacks (Pang & Lee, 2008). There are some topics on the research to find out people's opinions there are several types such as opinion extraction, opinion taking, sentiment classification, and summary of opinions. People's opinions are then extracted using keywords and it can be performed by utilizing an Application Programming Interface (API) (Chae, 2014).

Traffic congestion is a major problem faced by many big cities in the world, especially in developing countries. On Twitter, traffic information is widely available. However, the information is often still not structured yet. Sometimes, the traffic information is also broadcast via videos, but they are difficult to access because it requires greater bandwidth. For this reason, it is necessary to process the unstructured information and data from Twitter into the structured one using the data mining process.

Data mining is a process of finding knowledge from very large data. Text data mining is a data mining field that aims to gather useful information from text data. For extracting the text data, natural language processing and then extracting information are done which are useful for certain purposes (Noh & Lee, 2015).

Natural language processing has several steps in exploring the meaning and definitions of a text. This step includes (1) text processing, (2) lexical analysis, (3) syntax analysis, (4) semantic analysis, and (5) pragmatic analysis (Dale, 2010). This research is focused on semantic analysis which aims to study the meaning contained in words, expressions, and overall meaning of the words (Dale, 2010).

To conduct a semantic analysis, there are several methods to be applied, namely logical approaches, discourse representation theory and one of the last methods is Generative Lexicon (Ide, 2013). This method is a development of the previous one of the query-based lexicon-based generative model. The semantic analysis process using Generative Lexicon is expected to transform the unstructured information from Twitter into structured information that is able to be processed by a machine. Based on the facts that have been surfaced, the formulation of the research problem is how to present the latest information on the road conditions for the public using semantic analysis with Generative Lexicon.

2. Theoretical Framework

2.1. Natural Language Processing

Natural language processing is a branch of artificial intelligence that focuses on natural language. Natural language is a language that is generally used by human beings in communicating with each other. The language received by the computer needs to be processed and understood in advance so that the intentions of users can be understood properly. Natural language processing is a language analysis that is divided into several stages, namely syntax, semantics, and pragmatics (Dale, 2010).

In the theory of understanding the natural language processing, there are 3 (three) main stages as follows: (1) **syntax** is an understanding of the order of words in the formation of sentences and the relationship among the words in the process of changing the form of the sentence into a systematic form. The syntax includes the process of setting the layout of a word in a sentence to form a sentence that can be recognized. In addition, the parts of a sentence can also be recognized well. (2) **Semantics**, which means the mapping of the form of syntax structure by utilizing each word into a more basic form and not dependent on the sentence structure. The semantics studies the meaning of a word and how could the meaning of the words form the meaning of a sentence completely; (3) **Pragmatics** explains how the statements are related to the world. To understand the language, the agents must consider more than just the sentences. The agents must look deeply into the context of the sentences, world circumstances, the objectives of the speakers and listeners, special conventions, and the like (Poole & Mackwort, 2010).

To obtain syntax, semantic, and pragmatic studies, there are steps that must be taken as shown in Figure 1. The data in the form of text is given the process of tokenization and sentence segmentation. This step is very important to continue to the next step, including lexical, semantic, and pragmatic analyses. This research will focus on the semantic analysis step.

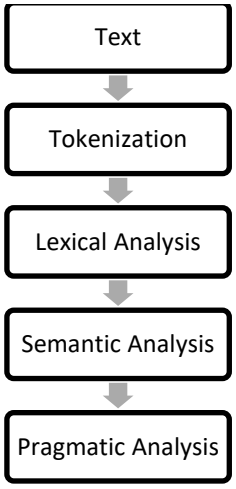

Figure 1. Steps of natural language processing

2.2. Semantic Analysis

Semantics comes from the Greek *semanticos* which means to give a sign, important, from the word *sema* which means sign. Semantics is defined as a study of the meaning of language utterances (Jurafsky & Martin, 2000). The semantic analysis shows the analysis process to study the meaning contained in the words, expressions, and the overall meaning of the words.

For longer texts, the semantic analysis uses natural language processing that includes information retrieval, information extraction, text summarization, data mining, and machine translation. In the shorter words or to the level of one word, semantic analysis is used in understanding the users' requests and matching the users' needs with available data. The semantic analysis is also appropriate to improve web ontology and the representation of science.

Some methods to conduct the semantic analysis include (1) logical approaches, (2) discourse representation theory, (3) Generative Lexicon, (4) Natural Semantic Metalanguage, (5) Object-Oriented Semantics, and (6) Weighted Lexicon-based Generative Model.

This research decides to use Generative Lexicon. The generative lexicon introduces a framework of knowledge representation that offers a rich and expressive vocabulary for lexical information. Generative lexicon can explain the use of creative language; we consider the lexicon to be a key repository holding a lot of information that underlies this phenomenon (Liao, Chen and Wei, 2014).

In the generative lexicon, there are existing lexical standards, namely: (a) **Lexical Typing Structure** which shows the type of words in the language; (b) **Argument Structure**, showing the number of predicates related to arguments or reasons; (c) **Event Structure**, showing the events and sub-events in the sentence; and (d) **Qualia Structure**: providing the explanations. Pay attention to the following lexical representation in figure 2.

$$\begin{bmatrix} \textbf{build} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E}_1 = \text{e}_1 : \text{process} \\ \text{E}_2 = \text{e}_2 : \text{state} \\ \text{RESTR} = < \alpha \\ \text{HEAD} = \text{e}_1 \end{bmatrix} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG}_1 = [1] \begin{bmatrix} \text{animate\_individual} \\ \text{FORMAL} = \text{physobj} \end{bmatrix} \\ \text{ARG}_2 = [2] \begin{bmatrix} \textbf{artifact} \\ \text{CONST} = [3] \\ \text{FORMAL} = \text{physobj} \end{bmatrix} \\ \text{D} - \text{ARG}_1 = [3] \begin{bmatrix} \text{material} \\ \text{FORMAL} = \text{mass} \end{bmatrix} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{create} - \text{lcp} \\ \text{FORMAL} = \text{exist}(\text{e}_2, [2]) \\ \text{AGENTIVE} = \text{build\_act}(\text{e}_1, [1], [3]) \end{bmatrix} \end{bmatrix}$$

Figure 2. Lexical representation

2.3. Twitter API and Its Users

Twitter's popularity has made it a repository of information (Arias, Arratia & Xuriguera, 2013). Data mining from Twitter is used as a material for strategy-making for a company or government agency. The main challenge for this condition is how to obtain valuable information from unstructured data, including the information related to road traffic.

To collect the data from Twitter, it is required the data mining process. Keywords and the use of the Twitter API are first used (Chae, 2014). This API will provide access for the researchers to obtain information submitted by the users. The data obtained at this stage are still irregular so that further processing is needed. The data mining process on Twitter is described in Figure 3 (Oussalah, Bhat, Challis, & Schnier, 2013).

Twitter API is a number of functions that software developers can use to process the data when building software. Twitter API provides several functions to perform certain tasks so that the software developers only call these functions in the software being built. The Twitter API uses the

REST (Representational State Transfer) architecture so that the Twitter API can be used in various data formats such as XML or JSON.

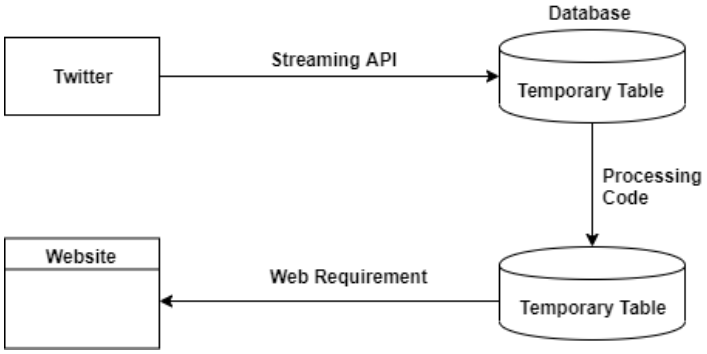

Figure 3. Data mining process on Twitter

The data used in this research are taken from the location-based social networking service of Twitter. The data obtained are a collection of tweets obtained by using the Twitter API with the keywords related to the traffic jam.

3. Method

3.1 Materials

The research material that would be used in this study is the data obtained from Twitter related to the information about the traffic conditions. This material was used for the semantic analysis process using a generative lexicon. Twitter data (tweets) characteristics included: (a) the tweets had only maximum number of 140 characters, (b) the tweets were sometimes followed by a hashtag that was usually used to focus more on the discussion of certain topics, for example, #lalin (*lalin*=traffic); (c) the tweets are usually in the form of repetition or supports for others, or the term retweet, usually detected by the presence of the character @, (d) the tweets could sometimes contain a link. When a tweet was sent, the time information which included hours, date, month and year was also sent automatically.

Traffic information in this research was taken from traffic information in Semarang city. Twitter data related to traffic information in Semarang city was taken from Twitter using the Twitter API.

3.2 Research Instruments

The research instruments included hardware, system software, and other assistive software. The following were some details of the specifications of the devices. The hardware consisted Computer Server with the type processor Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz, memory 6 GB, HDD 500 GB. The system software is Linux Centos 6.8 operating system for the server. Additional software included: Percona Database Server OpenLitespeed Server.

3.3 Research Procedures

During preparing the information system framework, the researcher developed the information system framework that would be created. The system framework to be built was as shown in Figure 4. Based on the system framework in Figure 4, there were several stages to be carried out, namely (1) retrieval of tweets data, (2) text segmentation, (3) detection of word type and (4) semantic analysis. These steps would be explained as follows.

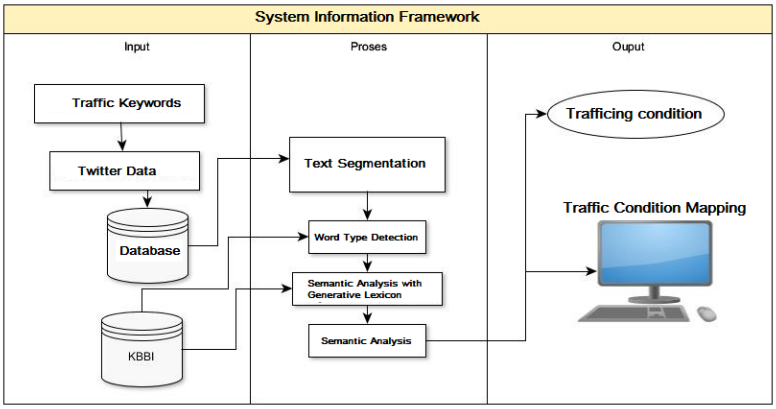

Figure 4. Information system framework

Retrieval of the Tweets Data

The Tweets data were used to get the tweets that matched with the keywords related to the traffic jam. The data would be stored in the local database server for the next processing step. In this process, the thing to be done was to build a system to retrieve the data related to road traffic conditions. To carry out this process, there were several steps to be done:

1. Building applications using PHP to access the Twitter API to retrieve the traffic data in Semarang city
2. Setting the application to retrieve the Twitter data about the traffic on a regular basis automatically. In this case, the data were taken twice a day.
3. Retrieving the traffic information data from the public and from a valid one from the Semarang City ATCS. Semarang City ATCS data can be accessed at @ATCS_KotaSMRG. The traffic data from the public were accessed using the keyword of places installed by the ATCS of Semarang City.
4. The data were then entered into a database server.

Text Segmentation

To hold the text segmentation, there were some steps to be done:

1. Text cleanup, cleaning up the text was done by removing punctuations, URLs, and special strings on Twitter such as a username (@username), re-tweet (RT), and hashtag (#hashtag)
2. Standardizing the forms of letters
3. Cleaning the words that were considered to have no value, such as affixes, suffixes, and conjunctions
4. Splitting the sentences into words

Checking the types of words

To check the types of words, a checking process was done using a dictionary called kateglo.com.

4. Results and Discussion

4.1 Retrieval of the Tweets Data

On the tweet's data, there are several challenges that must be solved. The first is the sampling rate imposed by several Twitter Application Programming Interface (API) combined with a very large number of tweets generated by the user at any time to collect all the tweets that are rather difficult and very dependent on the server used. The second one is related to real-time analysis, and if it is forced inappropriately, it will reduce the number of tweets collected during the analysis process. Search API is used to track the tweets in the past, but this method is quite weak because the API only recovers a number of last tweets. Third, although geolocation information comprised of latitude/longitude, it does not include precise addresses such as locations or positions descriptions.

Fourth, the proper handling of users' text requests will require complex calculations to understand the semantic aspects of the query, instead of the standard words.

The fifth is because the permitted tweet size is only 140 characters per message, the Twitter messages are not necessarily built with good ideas, proper sentences or phrases, but are incomplete, unstructured, and perhaps with many gaps and slang words (which sometimes cannot be found in the dictionary). Sixth, the Twitter data are not only a collection of texts and other attributes but an interactive and dynamic network complete with various connections that are suitable for the users' followers. This will require the identification of the original tweet check. Seventh, large-scale databases together with data analysis and time constraints make the data types and software representations are quite difficult to be processed (Oussalah, et al, 2013).

Based on the challenges above, the tweet's data retrieval is used to get the tweets that match the keywords related to the traffic. These data will be stored in the local database server for the next processing step. In this process, the step to be done is to build a system to retrieve the data related to the road traffic conditions.

The data retrieval uses search technique on Twitter. In this search technique, there are four things that users can do, namely: (1) word-level search that can use hashtags, single keywords, and word combinations, (2) personal based search in the form of user ID based search, users address-based search, and users-based search, (3) location-based search, namely the search which is based on specific latitude, longitude, surrounding locations, and specific distance locations, (4) Date-based search will display the data up to a specific date (Oussalah, et al, 2013). This system uses a person-based search using the username @ATCS_KotaSMRG for the Twitter data from Semarang city.

The data retrieval is done using an automation process for thirty minutes. Every thirty minutes, the system will do the data collection process. To do this task, the system uses the existing crontab facility on Linux. Twitter itself limits the retrieval in every fifteen-minute as many as 180 queries. The thirty-minute break is taken with the consideration that the data available in Semarang city still do not yet reach the 180-query limit.

4.2 Text Segmentation

The text segmentation is done through the following steps:

1. Text Cleanup
   The text cleanup is done by removing punctuations, URLs and special strings on Twitter such as a username (@username), re-tweet (RT), and hashtag (#hashtag).
2. Standardizing the forms of letters
3. The sentences split into the words

4.3 Checking the Types of Words

In the initial planning, to check the word types, a checking process is done using a dictionary called kateglo.com. During the process using a direct connection to the server kateglo.com, there are many requests that could not be served by the server, as described in Figure 5.

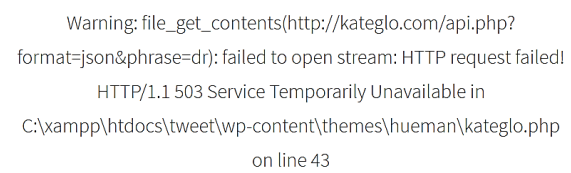

Figure 5. Words Checking using Kateglo.com

The results show that there are many unsuccessful requests, thus the researcher takes data of the Great Dictionary of the Indonesian Language of the Language Center (*Kamus Besar Bahasa Indonesia Pusat Bahasa – KBBI*) directly provided by kateglo.com. By using the direct database, the connection is getting stronger, and a lot of word types checking requests are successfully generated.

The following is an example of checking the word types.

22.00
WIB
Situasi
arus-->Nomina
lalin
di-->Pronomina
simp
Kariadi
ramai-->Lain-lain
lancar-->Adjektiva
dari-->Lain-lain
arah-->Nomina
Kaligarang,
S.Parman
&
Dr.Sutomo

The database of *Kamus Besar Bahasa Indonesia* in this research is used as a data source for checking the word types. Such steps have also been carried out by a team of researchers from Microsoft, namely Hua et al (2015) who used the vocabulary for checking the word types. In this research, the words are detected for their types from each word that exists.

4.4 Results of Semantic Analysis

Based on the results of the semantic analysis with Generative Lexicon, the data processing is performed in the form of a tweet as follows

07.40 WIB Lalin Simp Peterongan terpantau ramai lancar didominasi kendaraan dari arah selatan Jl MT Haryonopic.twitter.com/m3eq6LMcVl

This tweet was then processed using semantic analysis with Generative Lexicon. The analysis results are as follows

1. 0740
2. WIB
3. Lalin
4. Simp
5. Peterongan
6. terpantau
7. ramai
Jenis Kata:Lain-lain
Makna Kata:riuh rendah (tentang suara, bunyi)
8. lancar
Jenis Kata:Adjektiva
Makna Kata:tidak tersangkut-sangkut; tidak terputus-putus
9. didominasi
10. kendaraan
Jenis Kata:Nomina
Makna Kata:sesuatu yang digunakan untuk dikendarai atau dinaiki (seperti kuda, kereta, mobil)
11. dari
Jenis Kata:Lain-lain
Makna Kata:kata depan yang menyatakan tempat permulaan (dalam ruang, waktu, deretan, dsb)
12. arah
Jenis Kata:Nomina
Makna Kata:orang yang menjadi pembantu
13. selatan
14. Jl
15. MT
16. Haryono
pic.twitter.com/m3eq6LMcVl
Tempat:Simpang PETERONGAN
Kondisi: 1.

From the results of the analysis, it is obtained the data that the system can detect the location of the event, for example at the Peterongan intersection. Then, the adjective types of data are detected as traffic conditions.

In this semantic analysis, the information extraction process is carried out to extract traffic jam information in Semarang city from Twitter. The information extraction is done using a semantic analysis-based approach, which is based on the sentence structure. The information extracted is described as follows:

Location Information

In this research, the following methods have been used to detect and extract the location information:

1. Based on the sentence structure: This method is used to extract the location information that shows the relationship between two locations. The words we use are *"menuju", "ke", "arah" or "sampai"*.
2. Using a location dictionary: This method is used to extract the location information that shows the location of a place, using a dictionary that contains a collection of places/intersections in Semarang city.

Date Information

The date of the event is taken from the date information of each tweet used. The retrieval is based on the variable of "create_at" of the tweets.

Time Information

The retrieval of time information is done by extracting from the tweets. If the user does not include the time information, we can use "created_at" variable data

Pictures/images from the information from URL

The URL attached to the tweet usually contains a photo/picture when a traffic jam occurs. Therefore, it is necessary to extract the URL information even though not all of the tweets retrieved have the URL information

The studies related to the traffic jam on Twitter have ever been conducted by some researchers, one of which was conducted by Rodiyansyah (2012). In this research, there are several things done by the researcher such as taking data from Twitter. In this step, both are the same because the data are taken from Twitter, so this step cannot be abandoned. There are several matters that distinguish Rodiansyah with this research, (1) the main focus of the research from Rodiansyah is related to the classification of Twitter posts related to the traffic jam using Naïve Bayes Classification. In this classification process, there is a training process carried out on the data. It is different from the steps taken by Rodiansyah, where the conditions are determined according to the semantic analysis conducted by the system. This semantic analysis is based on data from *Kamus Besar Bahasa Indonesia* which then concludes the traffic conditions. Both studies use the maps as visualization. However, in this study, a layer other than the Google map is given, and the researcher also adds a layer of traffic from Google to see the conditions that exist at the location in question. Visualization of the congestion conditions is displayed in figure 6.

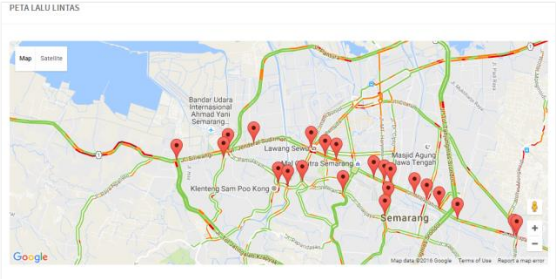

Figure 6. Visualization of traffic congestion

To validate the results, the researcher uses the analysis results of the data, the selection of the types of traffic jams, and images of actual conditions. This form of validation is used to show real data validation on the field. The following are the results of data validation in the system.

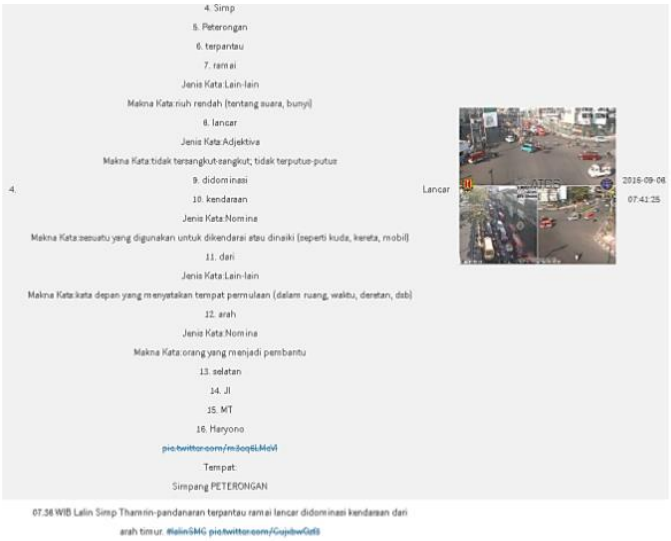

Figure 7. System validation

The results also show that the system has some limitations in detecting the word types. Some words are not found in the system, for the words "*Lalin, Simp"*. The word *Lalin* often appears in the data but not in the dictionary. *Lalin* means traffic in English. The word "Simp" means junctions so that the word *simp* is not found in KBBI. With respect to this, we need to add more vocabularies so that the words that are frequently used can be defined and such types of words can be easily found in KBBI.

5. Conclusion

In the semantic analysis process of Twitter data with Generative Lexicon for traffic congestion information, several steps are carried out, namely (1) automatic data retrieval from Twitter data, (2) text segmentation, (3) checking the types and meanings of words from KBBI, and (4) semantic analysis with generative lexicon. The system is able to detect the place, time and condition of the roads/highways. The detected data can then be visualized using a map with a combination of traffic information provided by Google Maps. With this kind of visualization, people can get precise information on traffic congestion conditions effectively.

References

Arias, M., Arratia, A., & Xuriguera, R. (2013). Forecasting with Twitter data. *ACM Transactions on Intelligent Systems and Technology*, 5(1), 8:1-8:24.

Chae B.K. (2014). Insights from hashtag #supplychain and Twitter analytics: *Considering Twitter and Twitter data for supply chain practice and research*. *International Journal Production Economics*, 165. 247 – 259.

Dale, R. (2010). Semantic analysis, In Indurkhya, N., dan Damerau, FJ. (Eds.). Handbook of natural language processing second edition. Washington DC, Chapman & Hall/CRC, 3-9.

Hua, W., Wang, Z., Wang, H., Zheng, K.., & Zhou, X. (2015). Short text understanding through lexical-semantic analysis. In International Conference on Data Engineering (ICDE).

Ide, N. (2013). *Advances in generative lexicon theory*. Springer: New York.

Jurafsky, D., & Martin, J.H. (2000). Speech and language processing. Prentice-Hall: New Jersey.

Kompas (2015). Pengguna Twitter di Indonesia capai 50 Juta [Twitter users in Indonesia reaches 50 million]. PT. Kompas Cyber Media: Jakarta. Retrieved from http://tekno.kompas.com/read/2015/03/26/16465417/pengguna.twitter.di.indonesia.capai.50.juta.

Liao, X.-W., Chen, H., & Wei, J.-J. (2014). A weighted lexicon-based generative model for opinion retrieval. Proceedings of the International Conference on Machine Learning and Cybernetics. pp.821-826.

Noh, H., Jo, Y., & Lee, S., (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, 42(9), 4348–4360. DOI: 10.1016/j.eswa.2015.01.050

Oussalah, M., Bhat, F., Challis, K., & Schnier, T. (2013). A software architecture for Twitter collection, search and geolocation services. *Knowledge-Based Systems*, 37, 105-120.

Pang, B., & Lee, L., (2008). Opinion mining and sentiment analysis. *Foundation in Information Retrieval*, 2(1-2), 1-135.

Poole, D. L., & Mackwort, A.K.. (2010). Artificial intelligence: *Foundation of computational agent*. Cambridge University Press: New York.

Rodiyansyah, S.F., & Winarko, E. (2012). Klasifikasi posting Twitter kemacetan lalu lintas Kota Bandung menggunakan naïve bayesian classification [Classification of Twitter posts Bandung traffic jams use naïve bayesian classificatio]. *Indonesian Journal of Computing and Cybernetics Systems*, 6 (1), 91-100.

Twitter (2015). Twitter usage, Twitter, San Fransisco, accessed on February 26th, 2016 from https://about.twitter.com/company