www.bircu-journal.com/index.php/birci
Budapest International Research and Critics Institute-Journal (BIRCI-Journal)
Humanities and Social Sciences
ISSN 2615-3076 Online)
ISSN 2615-1715 (Print)
Budapest Institute

Paper Review Text Mining Twitter

Samuel Dananjaya Raharjo[1], Gunawan Wang[2]
[1,2]*Binus Graduate Programs, Universitas Bina Nusantara, Indonesia*
samuel.raharjo@binus.ac.id, gwang@binus.edu

Abstract

Research on data analysis from social media such as Twitter has been carried out in recent years. This cannot be separated from the fact that Twitter is one of the most popular social media used by social-networkers. One of the services provided by Twitter is an API (Application Programming Interface) which allows developers to get Twitter data directly for further processing. This paper aims to review the papers on twitter data mining that have been published. The contribution of this paper is to provide information on the extent of the research that has been done on Twitter data mining to obtain a mapping that will be used as the next research plan. This review paper does not choose the best technique or method and does not provide an opinion on an analysis that has been carried out from previous research. From this review paper, it can be seen that a research activity can be carried out using twitter text data, with data acquisition techniques and text analysis methods in a text mining approach.

Keywords
sentiment analysis; twitter; data mining, crawling; paper review

I. Introduction

Twitter has grown into a microblogging site thatpopular in the category of social network applications. Twitter text content that contains a maximum of one hundred and forty characters does not prevent this service from being a reliable social networking media. This is possible because of the short and direct nature of Twitter messages, making it easier for users to convey the desired information. Social media is an example of a relatively recent development of information technology (Marbun et al, 2020). Communication through social media promises a comfortable state of communication, where someone who cannot compose words can be someone who is very poetic, with a very relaxed appearance and state, someone can carry out communication activities with others, lecturers, or someone when we communicate with it must take care of all things, appearance and style of language, but communicating through social media do not have to pay attention to it, sit back with a cup of coffee and use casual clothes a person can carry out communication activities (Marlina, 2020).

The use of twitter has been widely used for various purposes ranging from personal messengers, product and service promotion media, and even used as various media to give official messages from an authority. The various benefits provided by Twitter have become a social networking media that is quite efficient and effective for conveying short but fast messages.

Twitter application support for application developers has been given very significantly. Through its API function, Twitter data can be accessed, developed and/or stored for further processing. Various studies conducted in the form of Twitter data analysis have also been carried out with various approaches in the data mining framework, so that from the analysis obtained more valuable information from twitter text collections or "corpus".

DOI: *https://doi.org/10.33258/birci.v5i3.6842*

Of the many studies that have been published, it is known that there are at least more than tens of papers that review Twitter's data analysis activities for data mining. These papers have generally been published in worldwide conferences as well as electronic journals. The papers used for this review paper are the results of a search for papers that review the use of twitter data extraction for sentiment analysis.

The presentation of the results of this review paper will be divided into several sections of the discussion, namely Introduction (section 1), discussion of Twitter Data Acquisition Techniques (Data Capturing) (section 2), Data Analysis Techniques (section 3), Utilization of Twitter data mining results (section 4) and ends with conclusions and research plans that will do (section 5).

II. Review of Literature

2.1 Techniques to Get Twitter Data

Twitter contains short messages that are distributed through the site micro-blogging by users, which is limited to 140 characters per submission. The content of a sentence or text on Twitter is multi-character (can consist of numbers, letters or symbols) with a sentence structure that is free according to the user's wishes.

Twitter text can consist of several parts including emoticons, URLs, RT for retweets, @ for mentioning other users, # for hashtags used in opening twitter topics. Between Twitter users who are connected with other users (followers) can see each other the text of the message that is conveyed by a Twitter user to other users, known as Tweets.

Twitter offers three main methods for accessing and getting tweet data through the Twitter API (Application Programming Interface), including the Streaming API, REST (Representational State Transfer) API and the Search API. All of these methods are open, accessible to the public with the terms and conditions set by the Twitter site such as limits on the number of tweets that can be retrieved, the type or range of data you want to retrieve and so on.

The Twitter API provides access to tweet data from specific time range, from a specific user, with the word certain key, or from a certain geographic area, but it doesn't provide feature to extract structure from tweets, and does not provide an overview of aggregated data twitter on different topics (e.g., frequency of tweets about a particular topic from time to time).

The Streaming API relies on a continuous network between Twitter and the receiving host designed to support the volume of data transfers. The Streaming API allows users to issue continuous requests for twitter data over HTTP with the selected Keyword, Location, or User Id.

In contrast, the REST API follows unique Client-server requests and patterns in response communicating on request relationships between Twitter and hosts that are created dynamically on a per request basis. Furthermore, Twitter will provide API data in JSON (JavaScript Object Notation) format or in an interchange format similar to XML document representation.

Various techniques were developed to obtain data twitter more specifically according to needs user. These include TwitterZombie, a data app twitter crawling built with Search API technique and capable of pulling up to 1,500 corpus in a single process [2]. This data capture model captures (crawling) twitter data to be stored in a MySQL database as illustrated in Figure 1.

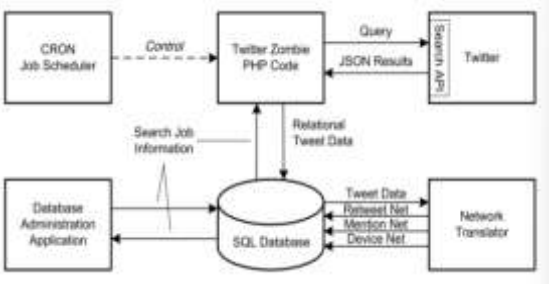

Figure 1. *Twitter Zombie System Diagram*

The use of the "Or" operator and the negation of "-" to "@" and "#" for users and text topics allowed the development of this model to get the data they wanted. The frequency of the data collection process is set at minute intervals so that several processes can be carried out in the data collection period. In addition to getting data, TwitterZombie produces a visualization of the results of retweet Network, reply to network and knows the type of application used by users from all tweet data obtained.

Besides TwitterZombie, another model was developed using the Streaming API called TweeQL for the programmer interface and TweetInfo for the user interface [3]. The technique that runs on TweeQL is to create STREAM to pull twitter data based on keywords, UDFs (User-defined Functions) which break the corpus into words in a per token format so that data obtained in the form of text, username, userid, location, latitude, longitude. The TweeQL architecture is described in Figure 2.

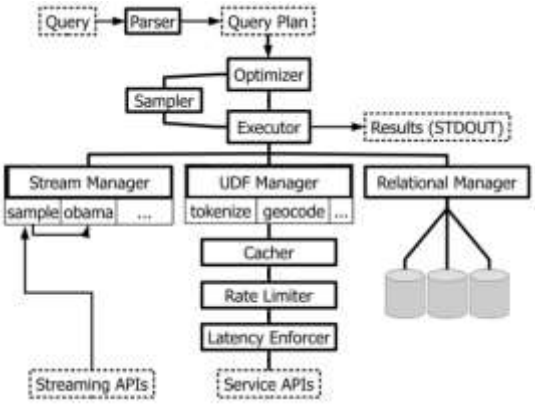

Figure 2. *TweeQL architectural components.*

With TweeQL, the classification of tweet words is generated based on the tokens obtained and the geolocation is also generated from the existing tweet data. UDF results developed using Exponentially Weighted Moving Mean (EWMA) analysis techniques can be used to determine trends or detect an event from processed twitter text.

One of the data capture models developed using the REST API is Twython which works to get twitter data according to extraction needs such as based on topic, user id or tweet data with a certain date range [4]. The REST API on Twython is developed with Python language which is able to generate data extraction about user profiles including city and country. From the city and country information, spatial data coordinates can be obtained, with the help of the Restful API developed by Yahoo!. The input data in the form of cities and countries are then converted into latitude and longitude coordinates (longitude and latitude) in XML form. With a combination of techniques and API features as well as web services from Yahoo!, classified twitter data is obtained, along with a description of

the interval for the arrival of the next tweet that retweets and the location where the tweet came from. In the research, the test was carried out using the determination of RMSE (root-mean square error).

III. Result and Discussion

3.1 Data Analysis Technique

a. Pre-processing Data

The pre-processing activity is the stage of work that will significantly contribute to the next activity within the Data Mining framework, namely the analysis of existing twitter data. Appropriate pre-processing activities will make data analysis work more accurate so as to produce quality information from data mining work.

Table 1. Pre-processing phase in various Text Mining researches

Researcer	Data Set	Preprocessing
Iman Raeesi Vanani. 2019	2.811.774 tweets	Make all letters in the dataset lowercase, remove emoticons, url and html tags, remove punctuation, remove stop words, remove frequently occurring words, remove rare words
Olga Monica , et al. 2019	5.600 tweets	Delete attribute language, location, number of author status, author's favorite, number of author's friends, number of author's favorite
Ashish Kumar Rathore, et al. 2020	302.632 tweets	Remove stop word, tokenization, stemming, and identifikasi n-gram.
Ankita Rane , et al. 2018	14.640 tweets	Remove irrelevant words, remove stopwords, lemmatization, remove non-English words.
Khalisa Virra, et al. 2019	812 training data and 203 test data	Tokenizing to make sentences into words and remove punctuation marks. Filter to remove stopwords

From Table I above, it can be seen that the twitter data preprocessing activities that are commonly carried out are filtering to remove common stop-words (“a”, “the”, “an” and so on), removing Retweet (RT) codes, eliminating duplicate tweets. remove emoticons, and delete tweets from different languages. In addition, tokenization is commonly done to make sentences into words.

b. Twitter Data Mining

Data mining jobs use special algorithms to complete various functions. The algorithm looks for the model that best fits the characteristics of the data being considered. The types of models that are known are predictive and descriptive models. Predictive models are used to make predictions, for example to predict road congestion, predict stock prices and so on. Some of the functions in predictive models are classification, regression and time series analysis.

Descriptive model used to identify patterns in the data. Some of the functions in the descriptive model are clustering, association rules and visualization. To perform Twitter

data processing with the Data Mining approach, generally several researchers use different methods, for example by using the Scoring method, Clustering, Classification or other methods as shown in Table 2.

Table 2. Data Mining Methods on Twitter Data Analysis

Paper	Metode				Results of data analysis
	Scoring	Clustering	Classification	Other	
Iman Raeesi Vanani. 2019	v				Sentiment score: 0.6696
Olga Monica , et al. 2019		v		v	Supporters of the Gerindra political party with an @herilatif account scored more than 40 degrees, while supporters of the PDIP political party with a marierteman account of less than 40 degrees
Ashish Kumar Rathore, et al. 2020			v	v	Post launch hyundai creta get 83 percent positive sentiment results
Ankita Rane, et al. 2018			v		The Random Forest algorithm obtained the best precision value of 85%, the best recall of 86.5%, and the best F-Measure of 86.5%.
Khalisa Virra, et al. 2019			v		Naïve Bayes algorithm is the best classification method with 83.54% accuracy

From various researches on Twitter data mining, as summarized in Table 2, in general the dominant model method used is the prediction model of the Classification method, by comparing several algorithms to obtain relatively high accuracy results so that the algorithm is feasible to be prioritized to be applied in mining Twitter data.

3.2 Benefits of Data Mining Results

Of the various studies using twitter data, each of them presents the purpose of using information in various forms depending on the purpose of the twitter data analysis. However, the use of Twitter data mining greatly contributes to various applications for determining detection, trend and sentiment analysis as described in Table 3.

Table 3. Twitter Data Mining Utilization

Category	Author	Tpe of twitter data processed	Utilization results
Trend	Iman Raeesi Vanani. 2019	About tweets a customer sends to a company to indicate a problem with a product, service or	The results show that there is a meaningful correlation between the trends of inbound tweets from customers toward international brands and their positive or negative

		opinion that might lead to improvement and the response an international brand provides to customers	sentiments.
Detection	Olga Monica, et al. 2019	The tweet contains the names of the political parties in the 2019 Indonesian elections, such as "Gerindra"	Find groups that have more influence on political parties to predict the winner of the presidential election
Trend	Ashish Kumar Rathore, et al. 2020	Tweets that contain attention to a product to be launched. The tweet contains 3 pre-launch and post-launch products from various brands such as "Hyundai Creta"	Shows that emotions play an important role in determining the subjective nature of product attributes. The various emotional contagions and their significance indicate a shift in the behavioral intention of users towards a new product taking into account its market performance.
Sentiment	Ankita Rane, et al. 2018	Data consisting of tweets for 6 major US airlines	Classification of twitter users' sentiments about US airlines for companies in order to improve service to users
Sentiment	Khalisa Virra, et al. 2019	Tweet about RUU PKS	Predict whether tweets posted via twitter support RUU PKS

From Table 3 it is known that the use of twitter data for data mining analysis has a category of utilization in detection, as trend indicator and sentiment analysis. Sentiment analysis can be used in a new product launch strategy to find out the trend or popularity of the community towards the brand's product.

Based on the above research on data from Twitter, sentiment analysis can be carried out to obtain grouping opinions on an object. Such as knowing how much positive support is for the support for the PKS Bill, and seeing the benchmarks for services carried out by airlines. In addition, Twitter data analysis containing sentiment can be used to predict the winner of the general election

IV. Conclusion

From research on review papers related to the use of Twitter data, researchers can conclude that Twitter provides various methods that can be used openly to obtain twitter data capture or known as corpus tweet. So that the corpus tweet obtained can be used in further analysis, it is necessary to do pre-processing so that the data is presented better, free from useless data and in accordance with the needs of the analysis. For this, proper pre-processing is needed in various ways and appropriate application algorithms. Furthermore, various analytical techniques will be carried out on data mining work in order to produce previously unknown information.

Based on this paper review, the authors plan research using twitter data which contains information about comparing sentiment analysis on minimarket coffee shops. For the data capturing process, the author will build a Twitter crawler application that works periodically and is able to overcome the limitations of data retrieval through the Twitter API. Data mining techniques will be carried out using various methods and algorithms to get the best accuracy. The results of the analysis will be used to compare public responses about the best coffee shop using the CRISP-DM step.

References

A. Black, C. Mascaro, M. Gallagher, dan S. P. Goggins. (2012). "Twitter Zombie : Architecture for Capturing , Socially Transforming and Analyzing the Twittersphere," ACM Gr. 2012, ACM, Sanibel Island, FL.

A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, dan R. C. Miller. (2011). "Processing and Visualizing the Data in Tweets," vol. 40, no. 4.

H. Wang, D. Can, A. Kazemzadeh, F. Bar, dan S. Narayanan. (2012). "A System for Real-time Twitter Sentiment Analysis of 2012 U . S . Presidential Election Cycle," no. July, pp. 115–120.

Marbun, D. S., et al. (2020). The Effect of Social Media Culture and Knowledge Transfer on Performance. Budapest International Research and Critics Institute-Journal (BIRCI-Journal), Volume 3, No 3, Page: 2513-2520.

Marlina, et al. (2020). Disclosure of Communication in the Facebook and Impact Social Media on Worship Activities in Dakwah Faculty Students and Science of Communication Media of North Sumatera State University (UINSU). Budapest International Research and Critics Institute-Journal (BIRCI-Journal), Volume 3, No 3, Page: 2142-2148.

Monica, Olga, Firda Wahyu Wahida, and Hanif Fakhruroja. (2019). "The Relations between Influencers in Social Media and the Election Winning Party 2019." 2019 International Conference on ICT for Smart Society (ICISS), Vol. 7, pp. 1-5, IEEE.

R. D. Perera, S. Anand, K. P. Subbalakshmi, dan R. Chandramouli. (2010). "Twitter analytics: Architecture, tools and analysis," Milcom 2010 Mil. Commun. Conf., pp. 2186–2191, Oct.

Raeesi Vanani, Iman. (2019). "Text analytics of customers on twitter: Brand sentiments in customer support." Journal of Information Technology Management 11.2, pp. 43-58.

Rane, Ankita, and Anand Kumar. (2018). "Sentiment classification system of Twitter data for US airline service analysis." IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC). Vol. 1, pp. 769-773, IEEE, 2018.

Rathore, Ashish Kumar, and P. Vigneswara Ilavarasan. (2020). "Pre-and post-launch emotions in new product development: Insights from twitter analytics of three products." International Journal of Information Management 50, pp. 111-127.

S. K. Wasan, V. Bhatnagar, dan H. Kaur. (2006). "The Impact of Data Mining Techniques On Medical," vol. 5, no. October, pp. 119–126.

Virra, Khalisa, Rachmadita Andreswari, and Muhammad Azani Hasibuan. (2019). "Sentiment Analysis of Social Media Users Using Naïve Bayes, Decision Tree, Random Forest Algorithm: A Case Study of Draft Law on the Elimination of Sexual Violence (RUU PKS)." 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC), pp. 239-244, IEEE.