JPPI Vol 9 No 1 (2019) 53 - 62

Jurnal Penelitian Pos dan Informatika

32a/E/KPT/2017

e-ISSN 2476-9266
p-ISSN: 2088-9402

Doi:10.17933/jppi.2019.090105

ONLINE REALTIME SENTIMENT ANALYSIS OF TWEETS BY UTILIZING STREAMING API FEATURES FROM TWITTER

ANALISIS SENTIMEN TWEET SECARA REALTIME ONLINE DENGAN MEMANFAATKAN FITUR STREAMING API DARI TWITTER

Bahrawi

Balai Besar Pengembangan Sumber Daya Manusia dan Penelitian Kominfo Makassar
Jl. Prof. Dr. Abdurrahman Basalamah II No.25, Makassar, 90123, Telp/Fax: 0411-4460084
E-mail: bahrawi@kominfo.go.id

Naskah Diterima: 27 Maret 2019; Direvisi : 5 Agustus 2019; Disetujui : 5 Agustus 2019

Abstract

Twitter is one of the social media that has a simple and fast concept. Because it is a short message, news or information on Twitter can be more easily digested. This social media is also widely used as an object for researchers or industry to conduct sentiment analysis in the fields of social, economic, political, or other fields. Opinion mining or also commonly called sentiment analysis is the process of analyzing text to get certain information in a sentence in the form of opinion. Sentiment analysis is one of the branches of the text mining science. Text mining is a natural language processing technique and an analytical method that is applied to the text data to obtain relevant information. Public opinion or sentiment in twitter is very dynamic and fast changing; a real-time sentiment analysis system is needed. It is automatically updated continuously so the changes can always be monitored, anytime and anywhere. This research builds a system to analyze the sentiment from twitter in real-time, automatically, and continuously. The results of the system trial succeeded in drawing data, conducting sentiment analysis, and displaying it in graphics and real-time website, and updated automatically. Furthermore, this research will be developed with a focus on the accuracy of the algorithms used in conducting the sentiment analysis process.

Keywords: sentiment analysis, twitter, opinion mining, text mining.

CC BY NC SA

INTRODUCTION

Latest data from Google consumers behavior written by Kemp reported that Indonesia has 265.4 million population and 50% among them are internet users. Indonesia occupies the sixth position of the most internet users in the world. Indonesia also ranked the 6th largest in the vicinity 3.6 billion total internet access in the world. The results of the We Are Social survey conducted in Singapore in 2017 showed that social media users of the Indonesian population reached 106 million of 262 million total population (Lucy Pujasari Supratman, June 2018).

Today, one of social media which is widely used by the society is Twitter. Twitter has a simple and fast concept. Because the message is short, the news or information on Twitter can be more easily digested by people who read it compared to social media which allows longer and more complex information, such as Facebook or other similar social media (Hardinata, May 2018).

Messages sent via Twitter are usually opinions or can also be called sentiments towards certain subjects. The opinions expressed by people through Twitter can produce data or information that are important and useful for stakeholders. For example in the food or beverage industry from a certain brand, we can collect as many Twitter messages containing opinions or sentiments towards the brand, then make them as the basis of information to change market strategy. Opinion mining or sentiment analysis is a computational study of opinions, sentiments, and sayings that are expressed textually (Liu, May 2012).

Sentiment analysis is one of Natural Language Processing (NLP) fields that builds a system to recognize and extract text-based opinions. Using sentiment analysis, the unstructured information can be converted into more structured data. The data can explain public opinion about products, brands, services, politics, or other topics. Besides, the data can also be used for marketing analysis, product reviews, product feedback, and community services.

Some Twitter sentiment analysis studies used the provided API feature to retrieve data in a period of time, for example Nurhada, et al (Faishol Nurhada, 2013) used the dataset taken from public timeline of tweets by period.

Opinions or sentiments change from the public on social media occurred very quickly. Retrieval of data based on the period is considered irrelevant except to see trends or other specific needs.

Based on that, it is necessary to have a dynamic system analysis which is more real-time and able to keep up with these changes. This study, though, is related to analytical sentiments. However, the focus of this study is building a system that can perform sentiment analysis of twitter content in real time and continuously as long as it is connected to Twitter. The results of the sentiment analysis are displayed in dynamic and real-time graph. The graph changes according to the results of the analysis. Accuracy of the algorithm used in this study is not the main focus.

METHODOLOGY

This research focuses on how to conduct sentiment analysis on Twitter content in a real time and continuous manner. Thus every tweet containing a predefined word will automatically be captured. A sentiment analysis is carried out using a system that

has been previously prepared. The analysis results will be displayed in the form of graphic in web pages. Every tweet underwent the process; hence no tweets are collected by period. Analysis was then conducted and the results were displayed. Generally, there are two approaches that are commonly used in sentiment analysis on internet-based social media. The first is a Machine Learning approach and the second is Knowledge-Based approach. This study applies the second approach, Knowledge-Based with the Lexicon-Based Method.

System Architecture

The system built in this study can be described in the use case diagram as follows:

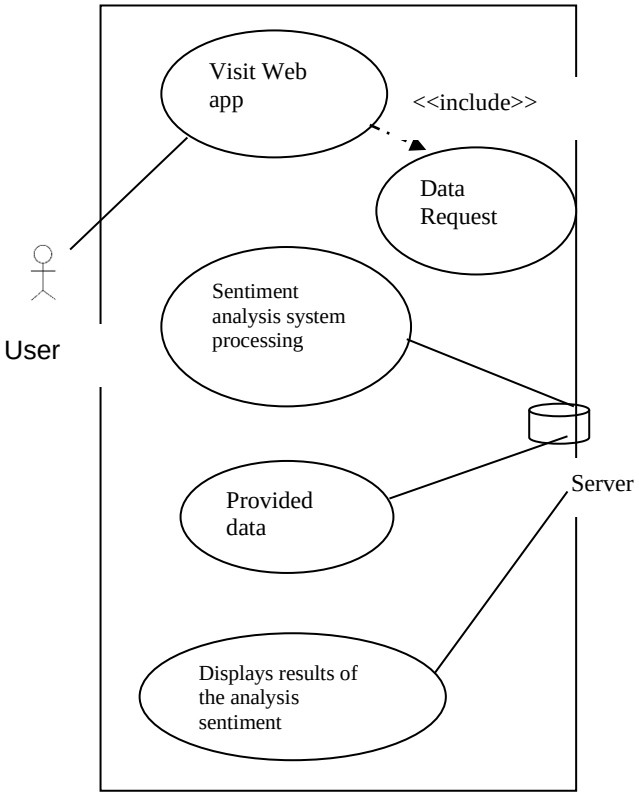

Figure 1. Use Case Diagram

The system is originally designed to be operational with a server. This research, however, did the testing in a regular desktop. The system is written using the Python programming language which already contains existing libraries.

Data Source

Raw data is obtained from Twitter by utilizing the API Stream service provided by Twitter. The raw data is processed until the expected sentiment analysis is produced. Before sentiment analysis of data obtained from Twitter, the pre-processing is carried out. At this stage, tweet data cleaning is carried out in the following steps: folding case, symbols removal, tokenisation, slangword conversion, and stopword removal. This process needs to be done because we can use all the available data.

Crawler Tweepy

Crawling is the process behind a search engine, which is tasked with tracking the World Wide Web structurally with certain ethics. Crawlers are engines or systems that are responsible for crawling. In a simple way, crawling is the process of visiting, reading, finding information from a web, indexing all the words in a document, and adding them to a database using a pre-determined system.

The data used in this study are obtained from the crawling process, taking opinion sentences om Twitter through the Twitter API. The crawler used in this study is tweepy, one of the library modules that runs on Python programming.

Twitter API Stream

Twitter Application Programming Interface (API) is an API service provided by Twitter to make it easier to access programmatic information that is available on the Twitter Web. This study uses the Twitter API stream to download tweets with the desired keywords. From the official Twitter site, API Twitter can be worked as follows:

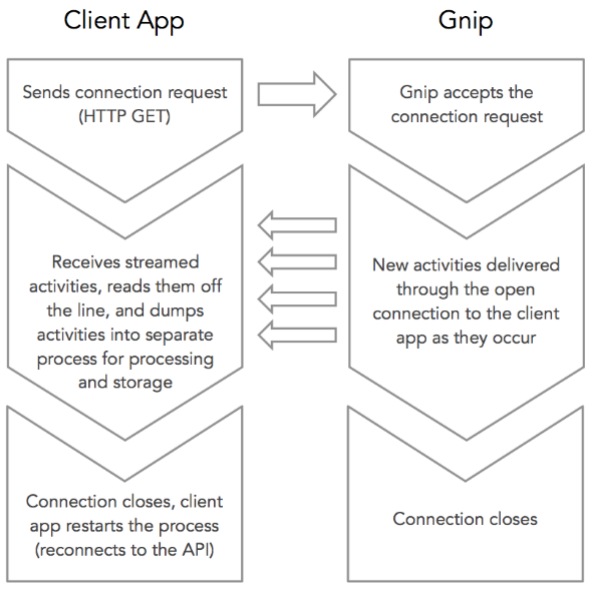

Figure 2. The Interaction of Application

SQLite3 Database

SQLite is a library written in C language that provides a lightweight database and does not require a separate server. The Sqilte3 module is written by Gerhard Haring, providing an SQL interface that complies with the DB-API 2.0 specification (Python Software Foundation, 2019). SQLite has the following features (Refan Andros, 2014):

a. Serverless

SQLite does not require a separate server process or system to operate. SQLite library accesses storage files directly.

b. Zero Configuration

The absence of a server means that it does not require configuration. This makes the SQLite database as easy as opening a file.

c. Cross-Platform

All databases are in a single cross-platform file that do not require administration.

d. Self-Contained

A single library contains the entire database system integrated directly into the host application.

e. Small Runtime Footprint

The system is built less than one megabyte of code and only requires a few megabytes of memory.

This study uses SQLite3 as a database in system testing.

Python 3

Python is an interpreted, interactive, object-oriented programming language. It incorporates modules, exceptions, dynamic typing, and very high level dynamic data types and classes. Python combines remarkable power with very clear syntax. It has interfaces to many call systems and libraries, as well as to various window systems. It is also extensible in C or C++ and usable as an extension language for applications that need a programmable interface.

There are two production-ready versions of Python: 2.x and 3.x. The recommended version is 3.x, which is supported by most widely used libraries. This research uses python version 3.7.2

Sentiment Analysis

Sentiment analysis is an area of science that developed from NLP. It allows the identification of emotions of social media users such as Twitter on certain subjects (Sotiris K. Tasoulis, Aristidis G. Vrahatis, Spiros V. Georgakopoulos, Vassilis P. Plagianakos, 2018). Sentiment analysis, also known as opinion mining, is a field of study that analyzes opinions, sentiments, evaluations, attitudes, and emotions of people towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes (Muhammad Akbar Maulana, 2018).

In this study, every tweet is downloaded through the crawling process. The pre-processing stage is, then,

carried out and the next step, followed by the next step, analyzing sentiment. In this study, the Vader library is used to analyze each existing tweet.

Vader sentiment analysis

Vader sentiment analysis is a python library module that can be used for the sentiment analysis process. On its official website, Valence Aware Dictionary and Sentiment Reasoner (Vader) states sentiment analysis is a lexicon and rule-based sentiment analysis tool specifically tailored to the sentiments expressed on social media, and works well on the text domain (Hutto, June 2014). This study uses the Vader module in the sentiment analysis process.

RESULTS AND DISCUSSION

The output generated from Python programming language application is web-based information page that displays the graph of the tweets contents and the results of the tweets sentiment analysis.

Application Testing

Application testing is done using a desktop computer with an Intel® i3-1.7Ghz processor, 2GB RAM, and a Windows 64 Bit system. Source code is written using the Python programming language by utilizing the Dash framework. The framework is specifically designed to build web-based applications that are widely used for creating web analytics. Figure 3 shows the display of webpage analysis designed in this study.

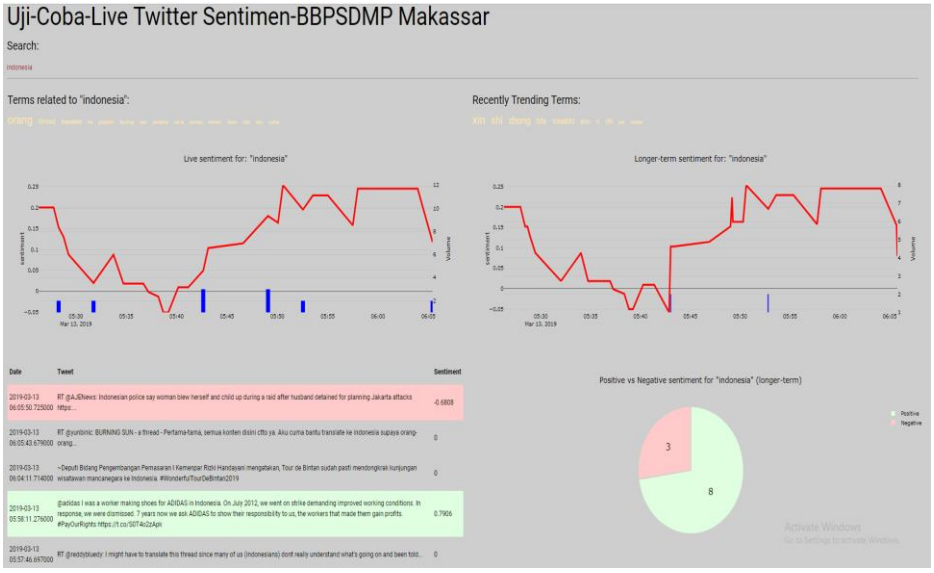

Figure 3. Display of Sentiment Web Page Real-time Tweet Analysis

Information shown in Figure 3 is acquired after the application program is run. As shown in the image, the information is divided into four parts. On the top side are two graphs of which describe the sentiment analysis crawled from social media Twitter. The graphs show the calculation of sentiment based on the algorithm. The sentiments are categorized as negative, positive, or neutral. The top left graph is shows the short-term data, while the right-hand graph shows the long-term data. The table at the bottom left side contains Twitter content, while the bottom right side is a pie chart showing the calculation of the number of tweets entered into the database by sentiment. Tweets

content with neutral tone are not displayed in the pie chart. Graphics, content table, and diagrams appeared on web pages will change automatically according to real-time online tweets streamed.

Figure 4. Search Menu

Figure 4 is a search menu located at the top, below the title on the application page. The search menu is used to enter the words which will be used as keys to search on Twitter

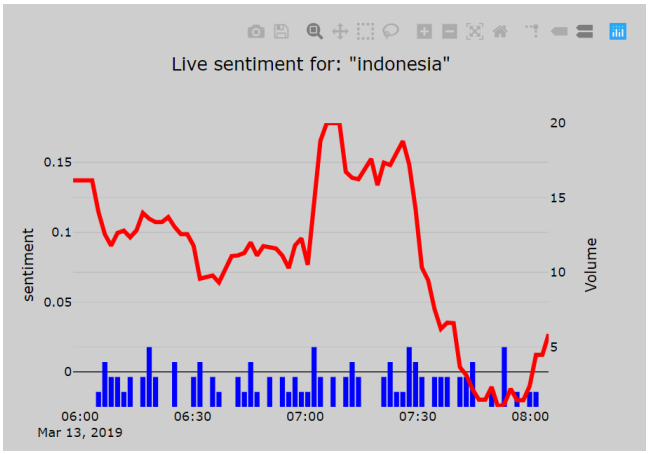

Figure 5. Real-Time Short-Term Analysis Graph

The graph in Figure 5 shows the value or score of the tweet sentiment analysis result. Zero represents the neutral sentiment, number above zero means positive sentiment, and number below zero means negative sentiment.

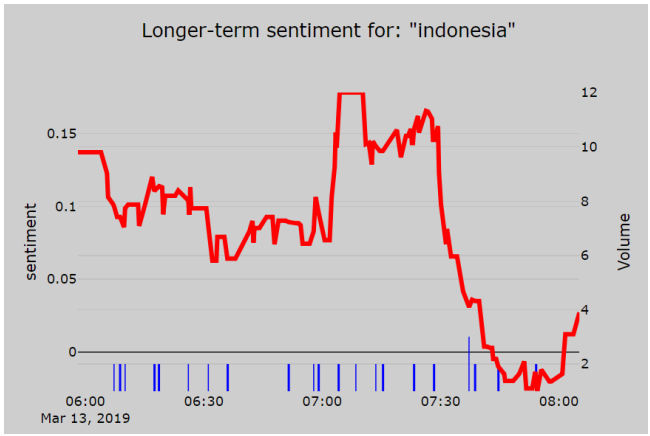

Figure 6. Real-time Long-Term Analysis Graph

Graph in Figure 6 is similar to that of in Figure 5. The difference is only the time period. This graph shows the fluctuations in sentiment with a longer time. The purpose of the graph is to make easier analysis on the sentiment, because it shows the sentiment fluctuation.

Date	Tweet	Sentiment
2019-03-13 08:05:16.704000	RT @DennyJA_WORLD: Fostering Harmonious Diversity in Indonesia, a Muslim World, is very important. Jokowi- Ma'ruf is for Diversity. Please...	0.624
2019-03-13 08:03:47.683000	RT @yunbinic: BURNING SUN - a thread - Pertama-tama, semua konten disini ctto ya. Aku cuma bantu translate ke indonesia supaya orang-orang...	0
2019-03-13 08:02:12.731000	RT @reddybluedy: I might have to translate this thread since many of us (indonesians) dont really understand what's going on and been told...	0
2019-03-13 08:01:49.613000	Sri Lankan Expatriates in Indonesia can build a Brand New Vajira House without coming to Sri Lanka. You can occupy the house as soon as you arrive in Sri Lanka, with a LIFE TIME GUARANTEE. Log on to our web site... https://t.co/4JZJaIIUCt	0.4084
2019-03-13 08:01:11.723000	RT @mputdemenribut: Exposing nct's indonesian fansites yang sedang menjamur belakangan ini. A THREAD	-0.2732
2019-03-13 08:01:07.717000	RT @yoshu_sudarso: I've never seen an Indonesian competes for ice skating before. Now that I've seen one I feel so proud I don't even know...	0.6113
2019-03-13 07:58:11.738000	@CNN Neu toi khong nham, thi tai nan xay ra do dieu ap trong khoang may bay, dan den may bay khong the dat do cao can phai co. Nguyen nhan co ve giong nhu vu tai nan voi Boeing 737 cua Indonesia!	0
2019-03-13 07:57:36.707000	RT @amandrug: typical indonesian in a nutshell https://t.co/D7ZrR1L5bH	0
2019-03-13 07:57:22.737000	RT @amandrug: typical indonesian in a nutshell https://t.co/D7ZrR1L5bH	0
2019-03-13 07:55:46.712000	RT @DennyJA_WORLD: Fostering Harmonious Diversity in Indonesia, a Muslim World, is very important. Jokowi- Ma'ruf is for Diversity. Please...	0.624

Figure 7. Tweets Table

Table XXX consists of three columns containing data tweets. The first column shows the time when tweets were posted. The second column is the Twitter content crawled based on certain specified criteria/included in the search section of application. The third column is the sentiment value of the tweet content.

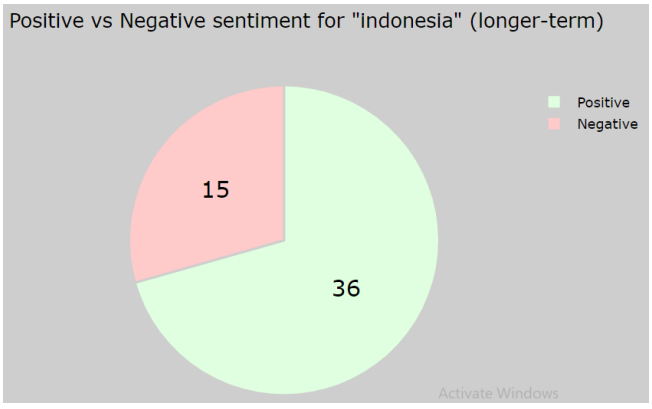

Figure 8. Sentiment Analysis Pie Chart

Figure 8 shows the sentiment analysis result processed by system.

Discussion

Despite a number of constraints, the application runs well. Some constraints are related to the devices used for testing, such as internet speed, memory system capacity, system specifications, etc. The application is run online and real-time. It keeps pulling data as long as the script continues to run and the internet connection network is still connected. Hence, this system should be run on a computer server with higher specification than the desktop used in trial.

By using a fast connection, the possibility of the graph motion will look better. Motion graphics in this research testing is very slow. That condition is most likely because the internet connection speed is very slow, and system memory is low. Internet connection speed and system memory capacity affect to normal motion of the graph. The charts will make it easier to analyze sentiments trending properly.

The accuracy of the sentiment analysis in this study is not measured because the focus is to provide online and real-time graph of sentiment analysis continuously. The tweet content appeared on the monitor screen will be updated continuously in real time according to the criteria or specified keywords input to the system.

Additional scripts are made to delete the data automatically when they have reached the specified limit of database capacity. It is a technique to anticipate stack data on the storage because the script continues to pull data from Twitter.

The development of the next system will be more focused on using stemming or libraries with Indonesian languages such as literary stemming, and so on. It is expected to result the analysis that is as good as possible.

CONCLUSION

According to the research objectives, the application runs to carry out recording, storage, and processing or analyze the sentiment from existing data tweets, and display it in realtime graphics.

Next research will be better if it focuses not only on the sentiment analysis technique but also the accuracy of measurement. The equipments or hardware used for testing must also be improved with higher specifications because the internet speed and system memory capacity are quite influential on the application performance

ACKNOWLEDGMENT

Thank you to the Center for Makasassar Communication and Informatics Ministry's Human Resource Development and Research, fellow colleagues in the Communication and Informatics Ministry's Research and HR Development, as well as other colleagues who have contributed to this research that cannot be mentioned one by one.

REFERENCES

Hardinata, R. (Mei 2018). Ekstraksi Pola Penggunaan Sosial Media Organisasi Keagamaan Menggunakan Metode Word Frequency dan Sentiment Analysis. *Jurnal ICT Akademi Telkom Jakarta*, 26-34.

Lucy Pujasari Supratman. (Juni 2018). Penggunaan Media Sosial oleh Digital Native. *Jurnal Ilmu Komunikasi*, Vol.15 No.1 47-60.

Hutto, C. &. (June 2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. . *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor: MI.

Ira Zulfa, E. W. (2017). Sentimen Analisis Tweet Berbahasa Indonesia dengan Deep Belief Network. *IJCCS Vol.11, No.2, July*, 187~198.

Liu, B. (May 2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Moujahid, A. (2014, July 21). *An Introduction to Text Mining using Twitter Streaming API and Python*. Retrieved from http://adilmoujahid.com: http://adilmoujahid.com/posts/2014/07/twitter-analytics/

Muhammad Akbar Maulana, A. S. (2018). Analisis Sentimen Media Sosial Universitas Amikom Yogyakarta Sebagai Sarana Penyebaran Informasi Menggunakan Algoritma Klasifikasi SVM. *Seminar Nasional Teknologi Informasi dan Multimedia 2018* (pp. 1.2-7~ 1.2-12). Yogyakarta: UNIVERSITAS AMIKOM.

Muljono, D. P. (2018). Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online

Menggunakan Algoritma Naïve Bayes. *Konferensi Nasional Sistem Informasi 2018* (pp. 165-170). Pangkalpinang: STMIK Atma Luhur.

Python Software Foundation. (2019, Mar 6). *Python Documentation contents.* Retrieved from Python Documentation: https://docs.python.org/3/contents.html

Refan Andros, L. (2014). Implementasi Honeypot Dengan Raspberry Pi Sebagai Alat Bantu Pendeteksi Kemanan Jaringan dan Penangkap Malware. *Jurnal Teknik dan Ilmu Komputer*, 12-26.

Sarworsi, A. H. (2009). Aplikasi Web Crawler Untuk Web Konten Pada Mobile Phone. *JUTI Vol.7 No.3*, 127-134.

Sotiris K. Tasoulis, Aristidis G. Vrahatis, Spiros V. Georgakopoulos,Vassilis P. Plagianakos. (2018). Real Time Sentiment Change Detection of Twitter Data Streams. *2018 Innovations in Intelligent Systems and Applications (INISTA)* (pp. 1-6). Thessaloniki, Greece: IEEE.