VISUALIZATION SYSTEM FOR SENTIMENT ANALYSIS USING TEXTBLOB ON TWITTER

Alfredo Gormantara
Department of Information Technology, Universitas Atma Jaya Makassar, Makassar, Indonesia
Alamat e-mail: alfredo_gormantara@lecturer.uajm.ac.id

ABSTRACT

Sentiment analysis is the classification of texts in several categories, usually positive, negative, happy or sad. In recent years, this technique has been used to analyze online product reviews, online news, general elections, disasters, stock markets, and social media, especially twitter. The results of the sentiment analysis can be used as one of the decisions making considerations. In helping to analyze the results of sentiments, visual analytics are used that can help to navigate data, compare various data sets, and explore data distribution. This research aims to analyze sentiment in Bahasa Indonesia using Tweepy and TextBlob as a python library to access and classify Tweets and with the help of visual analytics to explore and observe the distribution of Tweets based on geographical location, especially in provinces in Indonesia. In addition, this research also provides a comparison of the results of the validation of TextBlob using the Naïve Bayes and SVM algorithms with the results of each accuracy of 62.26% and 69.18% which are higher accuracy compared to SentiWordNet.

Keywords: *Sentiment analysis, TextBlob, Visualization system, Twitter, Bahasa Indonesia*

1. INTRODUCTION

The sentiment is attitude, mind or judgment using feeling. Sentiment analysis is also known as opinion mining which studies people's sentiment toward certain things [1]. By using sentiment analysis, someone could utilize language processing, computational linguistic and text analyzing that classifies text into polarity (i.e. positive, negative or neutral) and emotions (i.e. happy, angry or sad) [2]. This makes the result of sentiment analysis be created as a benchmark for assessing a situation based on the texts that are related to it or can be information that helps predict and assist the decision process.

The use of sentiment analysis result has been carried out in various fields starting from healthcare, politics, economy, social, government, etc. in order to get information in the form of views or public opinions [3]. Based on the results of research conducted by the global GFK and Indonesia Digital Association (IDA) institutions carried out in five major cities in Indonesia throughout 2015, the percentage of news consumption through online media reached 96 percent [4]. Currently, there is a rapid increase in web service, Internet technology, social media such as discussion forums, blogs, online news, online product reviews or responses to politics. These opinions are spread and can be accessed by anyone [5].

According to statistics in 2019, twitter users, based on PT Bakrie Telecom data, have 19.5 million users in Indonesia out of a total of 500 million of its global users [6]. Through the detection of hot topics and analysis of the topic, sentiments can exploit hot topics and appropriate distribution of sentiments throughout the world. Thus, exploring attractive social and economic values are possible. For example in research that uses twitter data analysis to improve company products and services [7]. The results of the sentiment analysis are used as foresight in making decisions. In addition, there are also research that take comments and assess products to extract data sources in sentiment analysis [8]. Some also took film review data as information about people's expressions for the film [9].

To understand the public sentiment from Twitter, a sentiment classification is needed. Some researchers have proposed approaches to sentiment analysis. In terms of analyzing sentiments, many approaches have been used that compare various algorithms such as the machine learning approach using the Naïve Bayes algorithm [10], Support Vector Machine (SVM) [11], Random Forest [12],

Convolutional Neural Network (CNN) [13], and many others. In addition, there are also libraries or APIs (Application Programming Interface) that have been developed such as AlchemyAPI [14], SentiWordNet, Word Sequence Disambiguation (WSD) [15] and a python library TextBlob. All available libraries or APIs usually only support text management in English.

To help analyze and view the results of sentiment, a visual analysis is used. Visual analytics is widely used in the analysis of social media data and contributes in many areas of analysis of exploration data, such as geographical analysis [16], and business prediction [7]. In addition to displaying data, visual analytics allows users to navigate data, compare various data sets, and explore data distribution. With a combination of sentiment analysis can classify data to be positive, negative or neutral. With a combination of sentiment and visualization analysis it is possible to detect distribution of hot topics [17], explore the impact of disasters [18] with the help of geographic visualization [19].

This research proposes sentiment analysis with visual analytics to explore and observe the distribution of Tweets based on geographical location in Indonesia's provinces. The rest of the paper is organized as follows: Section II explain the method used in this research. Section III describes the results and discussion of the trials that have been carried out and obtained. Finally, Section IV conclusion of the research.

2. METHODOLOGY

This research conducted a sentiment analysis on Twitter using TextBlob, which is a library for processing textual data. It provides a simple API for processing the data into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. TextBlob is able to classify, label and calculate the polarity value of the tweets. In addition, to get the tweet data directly, the Tweepy library from python is used [20]. In the final step, validation of the results of TextBlob using WEKA. The methodology proposed in this research is divided into two steps, namely the step of sentiment analysis and visualization. The research framework can be seen in Figure 1 below.

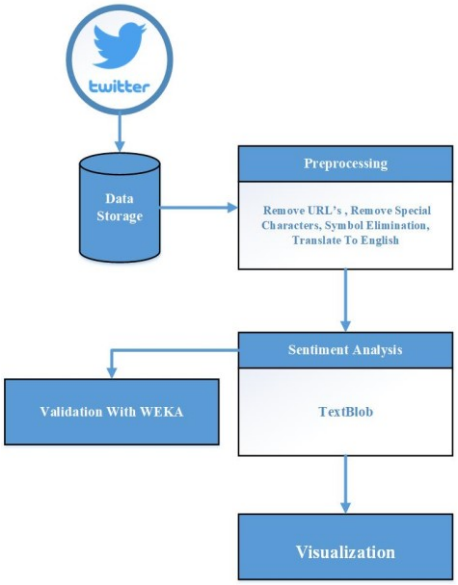


Figure 1. Research Flow

2.1 Data Storage

Twitter data collection was conducted using the Tweepy library from python which can directly access the Twitter API, and then the data are stored into text. In this research, 529 data tweets are stored as system test data with selected hot topics in Bahasa Indonesia.

2.2 Preprocessing

This step removes the URL, deletes the characters or symbols contained in a tweet, and finally translate it to English. The translation is performed using Google API mtranslate in Python.

2.3 Sentiment Analysis

Sentiment analysis was carried out using TextBlob. Data are inputted into the TextBlob library will be changed into polarity value that will determine their sentiment analysis. Where the value ranges from -1 to 1, making the analysis results negative, neutral or positive.

2.4 Visualization

The visualization step is to map the tweets based on their location in Indonesia's provinces so that they can be easier to analyze. The sentiment results will be colored

into the map with colors that will interpret the results of sentiment.

2.5 Validation

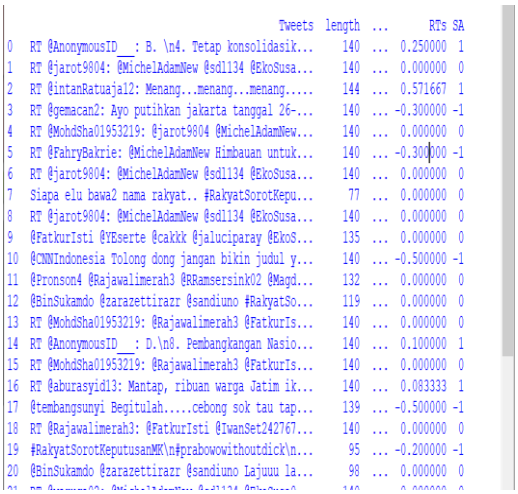
The validation is carried out using WEKA by testing the results of sentiment analysis on several algorithms. The results of the polarity value of the inputted tweets are used as training data for validation. The algorithms used are Naïve Bayes and SVM to compare the accuracy of the results of the polarity values that TextBlob has produced.

3. RESULT

In this section, the discussion will be divided into 3 parts, the first is how to collect data until the sentiment analysis process. The second is the stage of visualization of the results of sentiment analysis. Finally, validation of the results of TextBlob sentiment.

3.1 Tweet Sentiment Analysis

In the initial stages of data collection on keywords that have been determined using Tweepy. In Figure 2 is shown how the data have been collected.



	Tweets	length	...	RTs	SA
0	RT @AnonymousID__: B. \n4. Tetap konsolidasik...	140	...	0.250000	1
1	RT @jarot9804: @MichelAdamNew @sdl134 @EkoSusa...	140	...	0.000000	0
2	RT @intanRatuaja12: Menang...menang...menang.....	144	...	0.571667	1
3	RT @gemacan2: Ayo putihkan jakarta tanggal 26-...	140	...	-0.300000	-1
4	RT @MohdSha01953219: @jarot9804 @MichelAdamNew...	140	...	0.000000	0
5	RT @FahryBakrie: @MichelAdamNew Himbauan untuk...	140	...	-0.300000	-1
6	RT @jarot9804: @MichelAdamNew @sdl134 @EkoSusa...	140	...	0.000000	0
7	Siapa elu bawa2 nama rakyat.. #RakyatSorotKepu...	77	...	0.000000	0
8	RT @jarot9804: @MichelAdamNew @sdl134 @EkoSusa...	140	...	0.000000	0
9	@FatkurIsti @YBserte @cakkk @jaluciparay @EkoS...	135	...	0.000000	0
10	@CNNIndonesia Tolong dong jangan bikin judul y...	140	...	-0.500000	-1
11	@Pronson4 @Rajawalimerah3 @RRamsersink02 @Magd...	132	...	0.000000	0
12	@BinSukamdo @zarazettirazr @sandiuno #RakyatSo...	119	...	0.000000	0
13	RT @MohdSha01953219: @Rajawalimerah3 @FatkurIs...	140	...	0.000000	0
14	RT @AnonymousID__: D.\n8. Pembangkangan Nasio...	140	...	0.100000	1
15	RT @MohdSha01953219: @Rajawalimerah3 @FatkurIs...	140	...	0.000000	0
16	RT @aburasyid13: Mantap, ribuan warga Jatim ik...	140	...	0.083333	1
17	@tembangsunyi Begitulah.....cebong sok tau tap...	139	...	-0.500000	-1
18	RT @Rajawalimerah3: @FatkurIsti @IwanSet242767...	140	...	0.000000	0
19	#RakyatSorotKeputusanMK\n#prabowowithoutdick\n...	95	...	-0.200000	-1
20	@BinSukamdo @zarazettirazr @sandiuno Lajuuu la...	98	...	0.000000	0

Figure 2. Data Collection

The data that has been obtained is displayed using the Pandas library in python. Then the tweet will be translated into English using the mtranslate library. The data collected were 529 tweets which were divided into 297 neutral tweets, 181 positive tweets and 51 negative tweets.

After that, the translated tweet will be cleared by deleting the URL, character, and symbol in a tweet. The preprocessing process can be seen in table 1 below.

Table 1. Preprocessing

No.	Library	Output
1.	Input	#2019Presiden Rapatkan barisan, tugas terakhir kita tanggal 27 Juni 2019, kepada siapa MK berpihak.
2.	mtranslate	# 2019President Get the ranks, our last assignment is June 27, 2019, to whom the Court sided.
3.	Remove URL, characters or symbols	Get the ranks, our last assignment is June to whom the Court sided
4.	TextBlob	0.000

Table 1 shows the input data previously in Bahasa Indonesia was translated to English using mtranslate. After that, the removal of the URL, character or symbol is done at the beginning can only change the context of the sentence or affect the translation. Can be seen in the output of deletions, numbers and punctuation marks and symbols removed. The TextBlob result equals 0 means that the tweet is neutral, because the result from TextBlob if it is smaller than 0 is the same as negative, greater than 0 is positive and 0 is neutral.

In figure 3 shows the total percentage of tweets that have been collected based on the results of sentiment. The results of this percentage describe the conditions for the keywords that have been set in data collection.

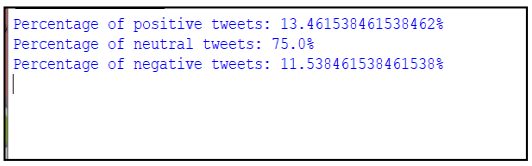


Figure 3. Result Sentiment Analysis

3.2 Tweet Sentiment Visualisation

At this stage visualization of the results of the sentiment analysis that has been obtained is based on geolocation. The geolocation of tweets is obtained from the Tweepy library used previously in data collection. The results of the sentiment

analysis are represented by colors in each province in Indonesia.

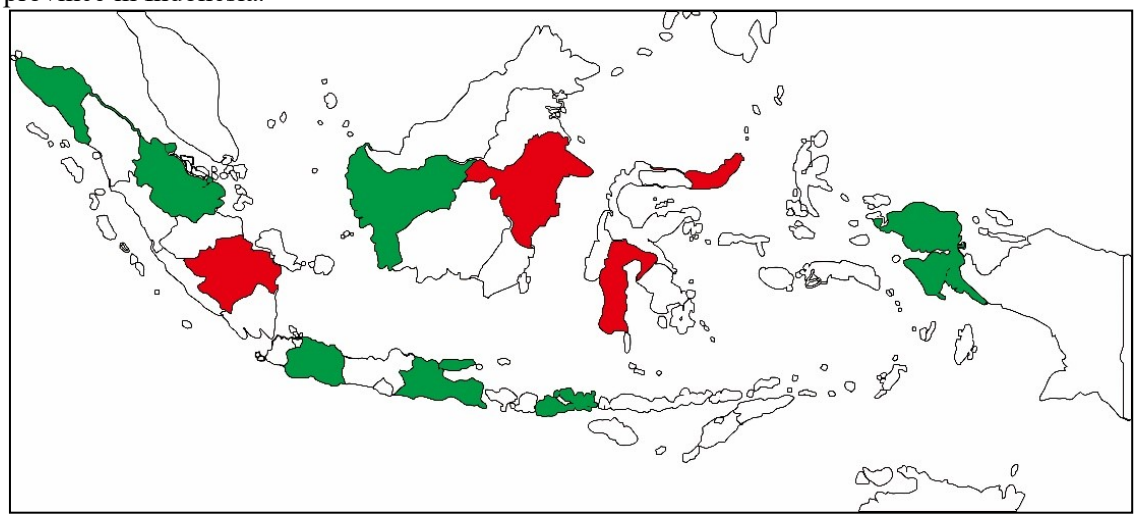


Figure 4. Proposed Visualization

As shown in figure 4, the tweets are grouped by province so that the percentage of sentiment results for each province is obtained. The percentage of the biggest sentiment results as a color choice reference for a province. The color symbolizes the results of the sentiment analysis previously obtained with red was interpreted as negative, green was interpreted as positive and white was interpreted as neutral.

3.3 Validation

To validate the results obtained from TextBlob, the Waikato Environment for Knowledge Analysis (WEKA) is used. When tweets are collected and translated into English, the file is generated in CSV (comma separated value) format. To make it compatible for validation, The CSV file is converted into ARFF. After the data has been uploaded to WEKA, use the existing StringtoWordVector filter with parameters: IDFTransform: true, TFTransform: true, stemmer:SnowballStemmer,stopwordsHandler: rainbow and tokenizer: WordTokenizer. The algorithms used in this stage are Naïve Bayes and SVM with a percentage split of 70%. After conducting the experiment, the accuracy of each analysis is recorded as shown in table 2.

Table 2. Comparison Algorithms

Algorithms	Training Set	Accuracy
Naïve Bayes	529	62.26 %
SVM	529	69.18 %

From Table 2 it can be seen that the resulting accuracy of TextBlob reaches 62.26% with Naïve Bayes and 69.18% with SVM. These results are better than the accuracy obtained from the use of SentiWordNet at 54.75% with Naïve Bayes and 53.33% with SVM [15]. From the results of the experiments above, TextBlob has higher accuracy compared to SentiWordNet, especially for tweets in Bahasa Indonesia.

4. CONCLUSION

This research focuses on the use of the TextBlob library for twitter sentiment analysis in Bahasa Indonesia and the incorporation of geographical analytic visual analysis. In analyzing sentiment results, polarity in lexicon-based methods are calculated based on the dictionary, which consists of semantic scores from certain words obtained through TextBlob. Furthermore, the research also validates and tests the results of sentiment analysis with two algorithms, namely Naïve Bayes and SVM. The results obtained that SVM has a better level of accuracy than Naïve Bayes

with a value of 69.18%. In the future, there will be more developments in libraries or tools for analyzing sentiments, especially for Bahasa Indonesia.

5. REFERENCES

[1] X. Fang and J. Zhan, "Sentiment analysis using product review data," J. Big Data, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0015-2.

[2] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Comput. Sci. Rev.*, vol. 27, pp. 16–32, 2018, doi: 10.1016/j.cosrev.2017.10.002.

[3] S. Shayaa *et al.*, "Sentiment analysis of big data: Methods, applications, and open challenges," *IEEE Access*, vol. 6, pp. 37807–37827, 2018, doi: 10.1109/ACCESS.2018.2851311.

[4] D. Afrianto, "96% Masyarakat Indonesia Konsumsi Berita Online," *Okezone.Com*, 2018. .

[5] P. R. Thanvi, N. S. Sontakke, S. R. Waghmare, Z. S. Patel, and S. Gavhane, "Sentiment Analysis for Political Reviews using AAVN Combinations," pp. 72–74, 2017.

[6] Kementerian Komunikasi dan Informatika, "Pengguna Internet di Indonesia 63 Juta Orang," *Kominfo*, 2019. https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker.

[7] M. A. Paredes-Valverde, R. Colomo-Palacios, M. D. P. Salas-Zárate, and R. Valencia-García, "Sentiment Analysis in Spanish for Improvement of Products and Services: A Deep Learning Approach," *Sci. Program.*, vol. 2017, 2017, doi: 10.1155/2017/1329281.

[8] W. Zhao *et al.*, "Weakly-supervised deep embedding for product review sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 185–197, 2018, doi: 10.1109/TKDE.2017.2756658.

[9] K. Chakraborty and S. Bhattacharyya, "Comparative Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Approach," *Springer Int. Publ. AG*, vol. 723, pp. 311–318, 2018, doi: 10.1007/978-3-319-74690-6.

[10] A. E. Khedr, S.E.Salama, and N. Yaseen, "Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis," *Int. J. Intell. Syst. Appl.*, vol. 9, no. 7, pp. 22–30, 2017, doi: 10.5815/ijisa.2017.07.03.

[11] S. K. Sharma and X. Hoque, "Sentiment Predictions using Support Vector Machines for Odd-Even Formula in Delhi," *Int. J. Intell. Syst. Appl.*, vol. 9, no. 7, pp. 61–69, 2017, doi: 10.5815/ijisa.2017.07.07.

[12] R. Khan and S. Urolagin, "Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 380–388, 2018, doi: 10.14569/ijacsa.2018.090652.

[13] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Syst. Appl.*, vol. 72, pp. 221–230, 2017, doi: 10.1016/j.eswa.2016.10.065.

[14] D. Sorvisto, P. Cloutier, K. Magnusson, T. Al-sarraj, K. Dyskin, and G. Berenstein, "Live Twitter Sentiment Analysis Big Data Mining with Twitter ' s Public Streaming API," pp. 29–41, 2018.

[15] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts," *Math. Comput. Appl.*, vol. 23, no. 1, p. 11, 2018, doi: 10.3390/mca23010011.

[16] P. Singh, R. S. Sawhney, and K. S.

Kahlon, “Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government,” *ICT Express*, vol. 4, no. 3, pp. 124–129, 2018, doi: 10.1016/j.icte.2017.03.001.

[17] Y. Zhao, B. Qin, T. Liu, and D. Tang, “Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog,” *Multimed. Tools Appl.*, vol. 75, no. 15, pp. 8843–8860, 2016, doi: 10.1007/s11042-014-2184-y.

[18] Y. Lu, X. Hu, F. Wang, S. Kumar, H. Liu, and R. Maciejewski, “Visualizing Social Media Sentiment in Disaster Scenarios,” *WWW ’15 Companion Proc. 24th Int. Conf. World Wide Web*, pp. 1211–1215, 2016, doi: 10.1145/2740908.2741720.

[19] N. Sharma, R. Pabreja, U. Yaqub, V. Atluri, S. A. Chun, and J. Vaidya, “Web-based application for sentiment analysis of live tweets,” *Proc. 19th Annu. Int. Conf. Digit. Gov.*, pp. 1–2, 2018, doi: 10.1145/3209281.3209402.

[20] S. Kunal, A. Saha, A. Varma, and V. Tiwari, “Textual Dissection of Live Twitter Reviews using Naive Bayes,” *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 307–313, 2018, doi: 10.1016/j.procs.2018.05.182.