Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication

1st Kusrini
Magister Teknik Inofrmatika
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
kusrini@amikom.ac.id

2nd Mochamad Mashuri
Magister Teknik Inofrmatika
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
moch.mashuri@outlook.com

***Abstract*— Twitter is one type of social media that is often used. Users use Twitter to convey their tweet to the general public. The number of Twitter users has reached 330 million people worldwide. Besides that, in Twitter there are tweets that can be sentiments. Sentiment itself can be defined as policy, opinion, logic, or mud, etc. Therefore, sentiment analysis is determining the polarity or type of opinion in a predetermined text or subject. NLP (Natural Language Processing) technique is used to support the beginning of a text. The technique used in an analysis is tokenization sentiment, elimination of stop words, and stemming. This study focuses on developing sentiment analysis using lexicons and multiplication polarity. Accuracy results are still smaller than using machine learning. Therefore, this lexicon needs to be improved in terms of its semantics.**

Keywords—twitter, sentiment analysis, NLP, lexicon

I. INTRODUCTION

Social media is now a reference material for companies to see customer behavior [1]. Analysis of tweet data can also be used as a reference to determine the business steps of the company in determining its policies. This analysis is done by looking for opinions or sentiments from several sentences or tweets obtained. Therefore, this stack of text data in Twitter is quite valuable because it stores valuable information. To uncover this information, data mining needs to be done using certain techniques. Mining this data can be done using text mining techniques which can be combined also using the Natural Language Preprocessing approach. Furthermore, important data that has been mined needs to be determined by the type of sentiment. This is done by using analytical sentiments.

Twitter is one type of social media that is often used. Users use Twitter to convey their Twitter to the general public. The number of Twitter users has reached 330 million people worldwide and every second produces 8000 data [2]. The chirp delivered can be in the form of news, opinions, arguments, and several other types of sentences [3]. This causes twitter to be rich in text that has certain data.

In general, someone wants opinions from other people as input to determine decisions. This opinion can be done by asking directly. By asking directly, it takes time and effort to meet people who are believed to ask. Another way is to get opinions from Twitter. Opinions in the form of tweets provided by Twitter with a large amount. However, this opinion must be distinguished based on the type of positive, negative, and neutral opinions. In addition, these tweets have not been grouped according to the categories you want to find. So, it is still widespread and necessary.

How to get opinions according to the desired category can use sentiment analysis. Sentiment itself can be defined as behavior, opinions, emotions, or mud, etc. Therefore, sentiment analysis is the identification of polarity or type of opinion in a text or subject that has been determined [4]. Today, sentiment analysis can be applied to almost all possible domains such as products, services for social activities and political elections, market research, advertising, recommendation systems, email filters, stock market predictions, and so on.

In fact, this analytical sentiment requires a Natural Language Processing (NLP) technique. This NLP technique is used for the initial processing of a text. The technique used in a sentiment analysis is tokenization, stop words removal, and stemming [5]. In this study, we will discuss sentiment analysis using data sets taken from tweeters. Then, text processing is done using several NLP techniques. Furthermore, the determination of sentiment in this study uses integer multiplication rules and matching against lexicon provided.

II. NATURAL LANGUAGE PROCESSING AND SENTIMENT ANALYSIS

A. A Brief of Natural Language Preprocessing

Natural Language Processing (NLP) is a branch of the field of Artificial Intelligence. This field seeks to provide a link for interaction with humans through natural language [6]. This kind of intelligent system requires computational and linguistic technology to build it, and the system processes natural language like humans. The development process at NLP has the steps shown in Figure 1.

The development cycle with NLP starts with collecting text. Text collected in a set is called a corpus. Then, all the text data is analyzed because not all text is processed. Furthermore, the process continues to the initial processing phase. This phase aims to clean and select the appropriate text. Then, change to the feature engineering phase. This phase attempts to get attributes from unstructured text. The existence of features that have been formed, the text becomes structured and ready for calculation to produce sentiment. The method for determining it can use rule based or machine learning. This research is using rule based on lexicon.

B. Sentiment Analysis

Field of study which is part of NLP and aims to interpret people's opinions, on certain topics, about any event, etc. [7] [8]. In text mining it is known as opinion or sentiment analysis. This results in a broad problem zone. There are also various names and have different assignments, eg, Analysis of sentiment, opinion extraction, opinion mining, sentiment mining, influence analysis, subjectivity analysis, mining

978-1-5386-8448-1/19/$31.00 ©2019 IEEE

review, etc. The flow of the processes that occur sentiment analysis is shown in Figure 2.

The process in Figure 2 begins with selecting a subject, then collecting tweets with that keyword and conducting sentiment analysis on the tweet. Tweets can be structured, semi-structured, and unstructured types. Sentiment Analysis Research, can collect tweets using different programming languages like R or python. The initial processing of data is nothing but filtering data to erase incomplete and noisy data.

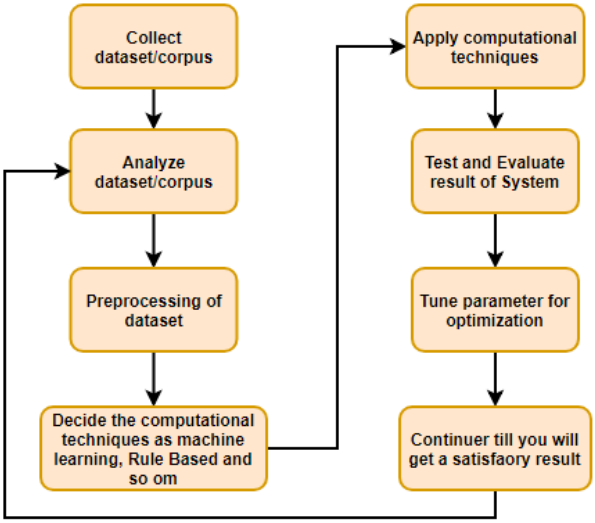

Fig. 1. Development Cylce in NLP

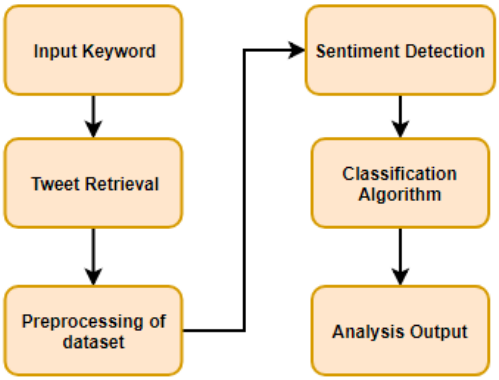

Fig. 2. Sentiment Analysis Workflow

C. *Lexicon Based*

The most important sentiment indicator is sentiment words. These are words that are commonly used to express positive or negative sentiments [9]. For example, good, beautiful, and amazing are words of positive sentiment, and bad, bad, and terrible are words of negative sentiment. Sentiment words are important for sentiment analysis for obvious reasons. A list of such words is called the sentiment lexicon (or opinion lexicon). Over the years, researchers have designed many algorithms to compile such lexicons. Although sentiment words are important for sentiment analysis, just using them is not enough. The problem is far more complex.

In the sentiment classification phase, there are several methods for classifying sentiments. These methods are shown in Figure 3. This classification was initially divided into 2 approaches, namely machine learning and lexicon based approaches. In this study more towards Lexicon which has been prepared dictionary base. This dictionary only contains adjectives that are collected and then have labels.

The following tasks are involved in the pre-processing task: • Deleting retweets (for twitter datasets) • Removing URLs, special characters, Punctuation, Numbers, etc. • Removing Stopwords • Stemming • Tokenization. The word sentiment identification is an important job in many applications of sentiment analysis and opinion mining, such as mining tweets, discovery of opinion holders, and classification of tweets. Sentiment words can be classified into Positive, Negative and Neutral words.

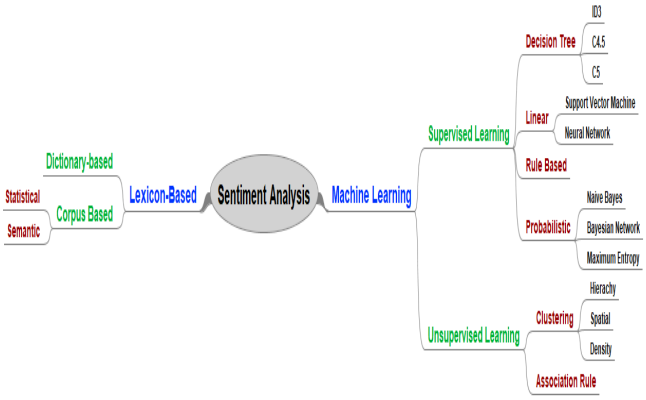

Fig. 3. Sentiment Classification Methods

III. METHODOLOGY

This research design is an experimental [10]. Experimental research is research that manipulates or controls natural situations by making artificial conditions. Making this condition is done by the researcher. Thus, experimental research is research carried out by manipulating the object of research, as well as the intentional control of the object of the research. In addition, in the experimental research there are three important elements that must be considered in conducting this research, namely control, manipulation, and observation. The control variable here is the core of the experimental method, because it is this control variable that will become a standard in seeing whether there is a change, or the difference that occurs due to differences in the treatment given. While the manipulation here is an operation that is deliberately carried out in experimental research.

The flow of this research is shown in Figure 4. This research is experimental or experimental. This research begins with a study of literature on the field of NLP and Sentiment Analysis. Then, data is collected. Data collected in the form of a collection of tweets taken through the use of Twitter API. The data taken is in the form of text stored in the database. Then, proceed to data analysis. The data analyzed was carried out descriptively. The results of the analysis are in the form of a list of adjectives along with manual labeling that is used as lexicon and designing a system of sentiment analysis. The results of this design are then used as a basis for building sentiment analysis in the construction phase of the system. The next phase is the testing strategy. In this strategy, if an error is obtained, it will repeat in the construction process. However, if there are no shortcomings, it will continue to the discussion phase.

In Figure 5. the flow of the sentiment analysis process is shown in getting sentiments towards tweets. The steps that can be passed can be described as follows:

1. Tweet Crawling

A user initially inputs the topic so that there are some tweets that match the topic. The topic was parsed through the

Twitter API to the server so that some tweets that contained words according to the topic were obtained.

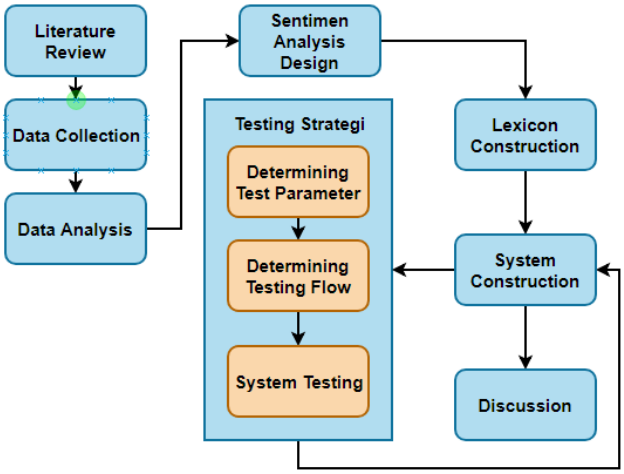

Fig. 4. Research Work Flow

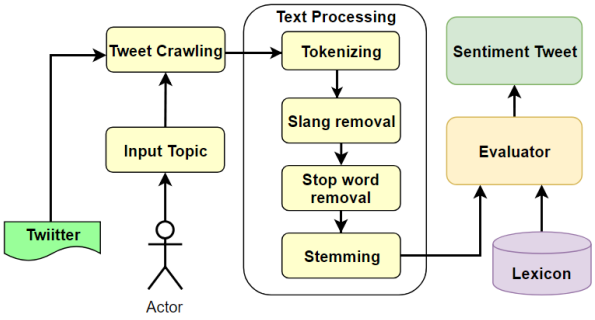

Fig. 5. Sentiment Analysis Work Flow

2. Tokenizing

At this stage, the tweets obtained are broken down into separate words. Additionally, the punctuation and url address that appears when tweets are also deleted.

3. Slang removal

This stage attempts to replace words that have current terms (slang words) with standard words according to language rules. This change process is done by making a list of slang words and their synonyms on the database. Then, match the existing words with the list of slang words in the database. If it matches, it will be replaced with the synonym. If the word does not match, the word is omitted.

4. Stop Word removal

In a sentence there are several words that indicate that the meaning of the word is less meaningful. Stop word in this study has provided a stop word list on the Twitter API.

5. Stemming

This stage aims to get the basic word on a tweet. The basic word is obtained using Porter Stemming which is in the Python library. This is done as a data reduction so that the processing of tweets is not too much data.

6. Evaluator

This sentiment analysis has an evaluator feature. This feature is useful for retrieving data from Lexicon and matching it with the words tweet generated from the text processing stage. This evaluator produced several tweets that have been given sentiment polarity such as positive, negative, and neutral.

The following are examples of tweet tweaks so that the polarity of the tweet is obtained according to the process in Figure 5. For example, the actor or user input the topic of meatballs until the following tweets are obtained:

" u're buying bakso but it's taste less delicious."

Next, there is lexicon with a list of adjectives shown in Table I.

TABLE I.]SAMPLE OF LEXICON

id	Word	Polarity	Score
1	Less	negative	-1
2	delicious	positif	1

This is continued to the tokenizing process so the results are as follows:

U're → slang removal → you

buying → stemming → buy

but → stop word → removed

it's → slang removal → it

less → according to lexicon → negative→ -1

delicious → according to lexicon → positive → 1

So that tweets that are ready to be processed are as follows:

"you buy bakso it taste less delicious"

Because there are 2 adjectives with each having a polarity value, the integer multiplication rule is applied so that the calculation is as follows:

less x delicious → -1 x 1 = -1

The final value obtained is -1 which represents that the tweet has a polarity or mean is negative.

IV. RESULT AND DISCUSSION

After conducting research in accordance with the predetermined research methods, the results of this study were obtained using the Python programming language tool and the library of twitter APIs, which were only learned, done, numpy, scipy, and smart. Then, the twitter API is set to get as many as 100 tweets. The initial step of using the system is done by inputting the topic on the system. Topics that are entered for example: gudheg e so that data is generated as shown in Figure 6. In Figure 6 the number of tweets produced as many as 100 in 65 tweets included in the positive classification. Furthermore, 23 tweets are included in the negative classification while 12 tweets are included in the neutral classification.

In this study 250 English datasets were provided and 112 adjectives were used as lexicon. Then, this is done by testing to find out the recall, precision, and accuracy of this sentiment

analysis. The test results are shown in Table II. .

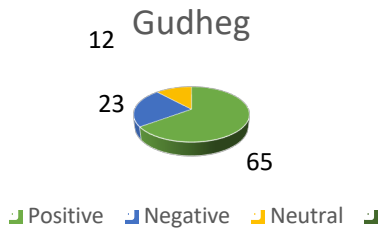

Fig. 6. Amount of Tweet with Polarity

Then, a comparison was made of previous studies using those using the matching learning approach [4]. This is shown in Table III. In Table III. Evaluation is conducted by preparing 250 data tweet. Then, those data give label with appropriate sentiment. Each tweet of those data is done analizing using polarity multiplication and matching using lexicon. The result is performed in Table II.

TABLE II. SENTIMENT ANALYSIS EVALUATION

Testing Category	**Polarity (Sentiment Type)**			**Average**
	Positive	**Negative**	**Neutral**	
Precision	0.62	0.67	0.13	0.4733
Recall	0.62	0.47	0.36	0.4833
Accuracy	0.64	0.66	0.75	0.6833

TABLE III. SENTIMENT ANALYSIS USING MACHINE LEARNING

Dataset	**Algorithm**	**Accuracy**	**Precision**	**Recall**
250 training and 100 Testing	Naïve Bayes	89%	89%	88%
250 training and 100 Testing	Linear SVM	76%	82%	76%
350 training and 150 Testing	Naïve Bayes	84%	83%	84%
350 training and 150 Testing	Linear SVM	79%	85%	80%

V. CONCLUSION AND FUTURE

Sentiment Analysis using the Lexicon approach has lower accuracy than using machine learning. This is because the number of adjectives in Lexicon is still incomplete. This must be completed. In addition, the Lexicon method is simpler than the methods available in machine learning.

Further work on this research is more in the direction of semantic analysis in determining its polarity. The semantic rules of a sentence can affect polarity. Therefore, it is necessary to conduct in-depth research on the Lexicon approach which adds concepts such as Uni-Gram, B-gram, Tagger Post, etc.

REFERENCES

[1] S. Y. Yoo, J. I. Song, and O. R. Jeong, "Social media contents based sentiment analysis and prediction system," *Expert Syst. Appl.*, vol. 105, pp. 102–111, 2018.

[2] A. D. Laksito *et al.*, "A Comparison Study of Search Strategy on Collecting Twitter Data for Drug Adverse Reaction," *2018 Int. Semin. Appl. Technol. Inf. Commun.*, pp. 356–360, 2018.

[3] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," *Inf. Fusion*, vol. 44, pp. 65–77, 2018.

[4] R. Bandana, "Sentiment Analysis of Movie Reviews Using Heterogeneous Features," *2018 2nd Int. Conf. Electron. Mater. Eng. Nano-Technology*, pp. 1–4, 2018.

[5] M. Kanakaraj, R. Mohana, and R. Guddeti, "Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015) Performance*, 2015, pp. 169–170.

[6] J. Thanaki, *Python Natural Language Processing*, no. July. Birmingham: Packt, 2017.

[7] R. Wagh, "Survey on Sentiment Analysis using Twitter Dataset," *2018 Second Int. Conf. Electron. Commun. Aerosp. Technol.*, no. Iceca, pp. 208–211, 2018.

[8] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, 2014.

[9] S. Akter and M. T. Aziz, "Sentiment Analysis On Facebook Group Using Lexicon Based Approach," in *iCEEiCT*, 2016, pp. 8–11.

[10] C. Kothari, *Research methodology: methods and techniques*. 2004.