ISBN : 978-602-14917-9-9

UTILIZATION OF SENTIMENT ANALYSIS USING THE DATA SCIENCE APPROACH TO IMPROVE CUSTOMER SATISFACTION

Ery Setiyawan Jullev Atmadji [1],Nanik Anita Mukhlisoh [2], Risqi Ahmad Sultoni [3]

[1] Jurusan Teknologi Informasi, Politeknik Negeri Jember

[1] Ery@polije.ac.id

Abstract. One of the biggest problem for customer satisfaction is how to understand the user need and the user point of view, to make it visible social media is giving huge impact especially tweeter comment . However, the number of comments submitted is very large and become difficulty to analyse. Besides the comment data on Twitter is an unstructured type of data so that if processing uses a relational database engine the results obtained are not optimal. To deal with these problems, a big data approach is needed in data extraction combined with the comment data processing model. This study uses a combination of big data in data processing and lexicon based to analyse customer comments. Data processing using big data especially with the NoSQL approach is very effective and efficient in conducting searches on unstructured data because the search for big data is based on meta text rather than cardinality between data. While the lexicon based method used depends on the completeness of the dictionary used. The purpose of this study is to analyse comments and share whether they have positive, negative, or neutral sentiments so that they can be used as parameters in decision making in an organization.

Keywoards : Social Media, Twitter, Comment, Lexicon Based, Big Data, NoSQL

Introduction

At present the volume of data available both on the internet and in a company is very large and will increase over time. Sensor data, log files, social media and other sources trigger, bringing volume, speed, and various data that far exceeds the traditional data and warehouse approach (Kusumawati, 2017). In business organizations that have the foresight to utilize new resources in creative ways to achieve unprecedented value and achieve other competitive advantages is a must.

The use of social media to obtain new data sources and get responses from the public who use the services provided by a company is a way to deal with current technological disruption (Francis, L., dan Flynn, 2010). Indonesia is ranked 5th for the number of active users, while for the number of tweets was 4.1 billion in 2016(Wahid & SN, 2016). With a large number of Twitter users making Indonesia a lucrative market destination besides that with a large amount of data and users, to do data processing requires a big data approach..

Sentiment analysis (also known as opinion mining) is the process of extracting information from a text and determining the attitude of the expressions / meanings contained in the text. (Kamal, 2017). Sentiment analysis can be used to conduct evaluations, assessments, attitudes and emotions on a product or service, organization or individual (Liu, 2012). The data obtained is also large and unstructured, the database used to store data must also support unstructured data, conventional databases do not support this type of method, therefore a special database is needed to store and release data with optimal speed. (Doshi et al., 2013). This unstructured database system is called No-SQL, most of the use of No-SQL is used to do predictive analysis on large scale organizations (Sheela, 2016).

ISBN : 978-602-14917-9-9

The tweet data contained on Twitter is very numerous and has various expressions so that it can be used to find out positive sentiments and negative sentiments towards a product or service delivered by users. Positive sentiment is conveyed to assess user satisfaction with something while negative sentiment is conveyed to assess user dissatisfaction (Park, Fink, Barash, & Cha, 2013).

In its development, many PT KAI customers have commented on the quality of service via Twitter. The quality of service from PT KAI can affect comments made by customers. Therefore, service quality must be considered because it will affect the image of the company so that negative sentiments will not appear to the company. Comments made by customers for PT KAI are so numerous that they encourage and motivate this research. Comments delivered regarding services include train facilities, order system and facilities / infrastructure at the station so that it encourages to conduct customer sentiment analysis of PT KAI's services.

Based on these problems and explanations, this research will use the Big Data approach in the form of No-SQL combined with lexicon based. By using the NO-SQL approach, unstructured data processing will be faster and combined with lexicon based, it is expected that data extraction can be done more easily and quickly.

Literature Review

1.1. Sentiment Analysis

Sentiment analysis or what is often referred to as opinion mining is a field of science that analyzes opinions, sentiments, evaluations, judgments, attitudes, and emotions towards entities such as products, services, organizations and individuals in a problems, events, topics, and attributes. There are several different names and assignments but still under the auspices of the sentiment and related to the analysis Sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, emotional analysis, etc. According to the book "*Sentiment Analysis and Opinion Mining*" yang ditulis oleh (Liu, 2012) *Sentiment analysis* in general, it is often used in the industrial and academic world. The term sentiment analysis first appeared by (Nasuka dan Yi, 2003), and the term opinion mining was first put forward by (Dave, Lawrence dan Pennock, 2003), whereas for research on sentiments have appeared before by (Das dan Chen, 2001; Morinaga et al., 2001; Pang, Lee dan Vaithyanathan 2002; Tong, 2001; Turney, 2002; Wiebe, 2000).

Research on people's opinions and sentiments was carried out before 2000. Since then, various reasons have emerged why sentiment analysis has become a very active area of research, among others, first, in the sentiment analysis industry area is developing due to commercial application business. Second, offer many challenging research problems to solve. Third, it has a very large volume of trusted data on social media. Therefore sentiment analysis takes a role in social media which influences linguistics and natural language but also has an important impact on science management, political science, economics, and social science.

Sentiment analysis or opinion mining is one of the branches of research in the domain of text mining which began to be popularized in 2003 (Matulatuwa et al., 2017). *Opinion mining* itself is a computational research of opinions, expressions and sentiments conveyed in writing or textual. Mining Opinion aims to extract a lot of data so that negative and positive sentiments can be identified. Mining extraction data in a document based on attributes and components that have been commented on each data.

With the growth of the big data population, to do sentiment analysis many tools can be used. Many companies use this development by analysing social media. The company now when going to produce a product will consider the political situation, promotion and customer satisfaction with the product or service that has been provided (Thompson et al., 2017).

In a number of studies mentioned that opinion mining focuses on applying to the classification of opinions based on positive or negative polarity. However, at this time many interpret the broad scope of computational opinions, sentiments, and subjectivity in the text. (Pang & Lee, 2006).

ISBN : 978-602-14917-9-9

1.2. *Text Mining*

Text Mining is an intensive knowledge process where users interact with a document that stores the results of the analysis by using an analysis tool at a time. Text mining can be used in various types of data, currently data is stored with various types of sources of data stored in a neatly arranged database to data that is not clearly structured. Text mining can be used on data types that do not have structured data types (Feldman & Sanger, 2006). Unstructured data can be in any type without the need to follow certain formats, rules and paths.

Text mining is much inspired and developed from the features of research conducted on data mining. Many systems adopted by text mining are a specific type of pattern in the core knowledge discovery operation that was first introduced in data mining research. Data mining assumes that data has been stored in a structured format so pre-processing focuses more on two critical tasks, scrubbing and normalizing data. While text mining, pre-processing focuses on the identification and extraction of natural language representations produced by data (Feldman & Sanger, 2006).

Figure 1 Process Text Mining

At the tokenizing stage is the stage of cutting each word in a sentence. The filtering stage is the stage of taking important words from the results of a token using the stop list algorithm (discarding words that are less important) or word list (storing important words). The stemming stage is looking for the root of each word in the results of the filtering stage. Tagging is the stage of finding the initial form of each past word from the results of stemming. Analysing is the stage to determine how far the relationship between words in an existing document (Matulatuwa et al., 2017)

1.3. *Lexicon Based*

Lexicon-based is a suitable method for analysing data in the form of questionnaires, Twitter data, Facebook data and data from various other social media. Lexicon based besides suitable for the data used in this study, lexicon based is also a simple and practical method for using data from social media. (Matulatuwa et al., 2017). The main advantage of lexicon based is portability in its use. When the lexicon based model is applied to the new model, the user does not need to retrain it. Users only need to update the dictionary so that they can present better on the new model, only updating what is needed is not necessary until the basic individual words (Thompson et al., 2017).

1.4. *Extract Keywords*

When the pre-processing process has been completed it will extract the keywords process. Following is the keyword extraction process can be seen in Fugure 2 (Nurfalah & Ardiyanti Suryani, 2017).

ISBN : 978-602-14917-9-9

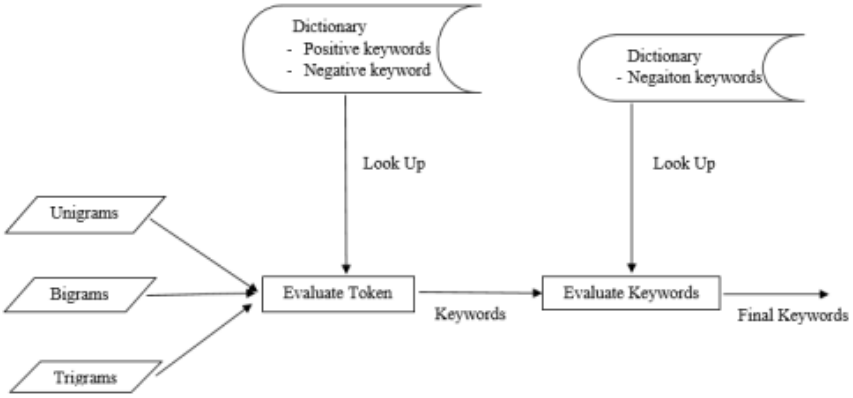

Figure 2 Keyword Extraction Flow Chart

Explanation for the above flow is when the opinion has been done tokenize into the form of unigrams, bigrams and trigrams then the process will be done Explanation for the flow above is when the opinion has been made tokenization in the form of unigrams, bigrams and trigrams then a token process will be conducted. Evaluate token is the process of matching keywords with a positive and negative dictionary. Then, the keywords that have been evaluated will be processed in evaluate keywords and will be matched with the negation word dictionary. At this stage is extracting the words that are the key determinants of positive or negative sentiment. Extraction is done by taking words that will be keywords to determine the type of sentiment. Keywords that have been extracted are then counted to be compared to determine sentiment.

In the steps to extract emoticons it is almost the same as extracting keywords in the previous stage, but because only emoticons are needed. Therefore, each sentence of opinion is searched for emoticons and then separated with other sentences. When finished extracting the emoticon, it is matched with the emoticon dictionary which can then be known the sentiment value.

To determine the level of accuracy of the analysis results is to compare the analysis results of the system with the actual analysis results. The following formulas can be used:

$$\text{Akurasi} = \frac{\text{Jumlah Sistem Mendeteksi Dengan Benar}}{\text{Jumlah Seluruh Kalimat Opini}} \text{ x } 100\% \qquad (1)$$

Architecture And Implementation

1.5. *Architecture*

After a literature study and needs analysis can then be made the design of the system to be built. System design that can facilitate this research to build a system systematically.

ISBN : 978-602-14917-9-9

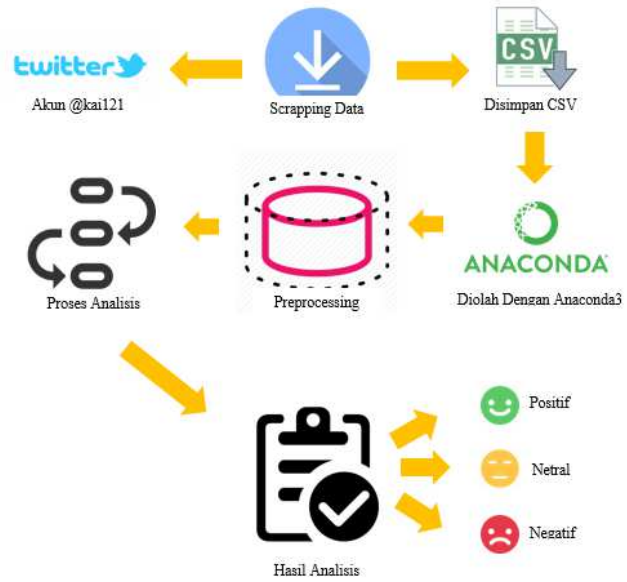

Figure 3 System Overview

Can be seen in Figure 3 is a general description of the system used. In the first process, scrapping data on social media Twitter using the @ kai121 account. Data scrapping is the process of taking data from social media that contains information. Data retrieved contains user, user id, tweet, and time. Scrapping data results are stored in CSV format. Data stored in CSV format is then entered into the Anaconda3 software. Anaconda3 is an IDE (integrated Development Environment) which is software that can be used to help develop a data processing system using python programming language. Then the supporting data in the sentiment analysis system is entered into anaconda3 software. The comment data is then pre-processed before being analysed. Pre-processing data is done by normalizing sentences into sentences that can be processed by the system. Data is then performed tokenization, which is the process of separating sentences into words per word or combination of words. The next stage is the process of analysing the data that has been done pre-processing. Analysis was performed using a positive negative word dictionary, negation dictionary, and negation dictionary. The results of the analysis process are sentiments of each comment data. The analysis results are divided into three sentiments namely positive, neutral and negative.

2. Implementation And Result

Data Acquisition or data acquisition is the process of getting data including customer tweet data. To acquire data, it is done by using tools, namely twint-master. Analysis comments are data obtained from social media Twitter. The number of PT KAI's followers on Twitter social media is greater than on Facebook and Instagram. On Facebook social media the number of followers is 95,993 users. On social media Instagram the number of followers reached 144,000 users. And on Twitter social media the number of followers is 840,000 users. So encouraging to do this research through social media Twitter.

After the Data Acquisition the next process is Pre-processing, Pre-processing is the stage to normalize the tweet data that has been obtained such as eliminating characters that are not important so that it does not affect the results of the analysis. At this stage also performed separation of each word in each sentence or tokenization. Before normalizing, the data that has been obtained is then selected by removing the tweets from the @ kai121 account. The selection is needed because for objectivity in the assessment of PT KAI's services, which uses comments made by PT KAI customers not from the @ kai121 account itself. From the data which amounted to 10,000 then made 5,815 data. The selected data can then be analysed.

ISBN : 978-602-14917-9-9

The results of the analysis of 5815 comments were divided into three types of analysis. Comments were obtained on Twitter social media in the range of August 4, 2018 until October 2, 2018. Analysis conducted included the use of unigram tokens, trigram tokens, and emoticon analysis. The following are the results of the analysis conducted.

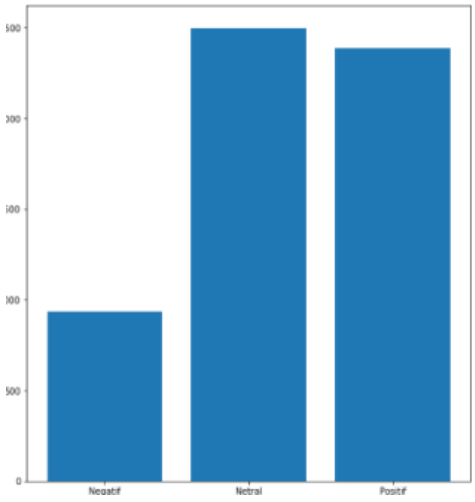

Figure 4 Unigram Token Analysis Results Graph

A unigram token is a token that separates every word in a comment sentence. Sentences that have been separated into word for word are matched with a dictionary and sentiment is determined. It can be seen in Figure 4 that neutral sentiment has the highest number of sentiments followed by positive sentiment, and the lowest is negative sentiment.

```
Jumlah data =  5815 komentar
Sentimen negatif =  933 komentar
Sentimen positif =  2385 komentar
Sentimen netral =  2497 komentar
Benar di analisis =  3051 komentar
Akurasi =  52.467755803955285 %
```

Figure 5 The Amount of Each Sentiment and Accuracy

In Figure 5 it can be seen that the number of comment data analyzed is 5,815 comments. From 5,815 data, there were 933 comments with negative sentiments, 2,385 positive sentiments, and 2,497 neutral comments. Overall the correct data analyzed by the system amounted to 3,051 comments or has an accuracy of 52.468%. Accuracy is obtained by dividing the results of the correct comments analyzed by the amount of data then multiplied by 100%. The following is an accuracy calculation.

$$\text{Akurasi} = \frac{3051}{5815} \text{ x } 100\%$$

$$\text{Akurasi} = 52.4677\ \%$$

From the results of the analysis conducted by the system with different sentiments so as to produce a different level of accuracy. It can be seen in Figure 6 the number of actual sentiments of the analysed commentary

```
In [59]: data['sentimen'].value_counts()

Out[59]: netral     3731
         negatif    1160
         positif     924
         Name: sentimen, dtype: int64
```

Figure 6 Total Every Real Sentiment

There is a difference from the results of the analysis by the system with actual sentiment. In Figure 6 the results of neutral sentiments show the number of 3,731 comments whereas, the results of the

ISBN : 978-602-14917-9-9

analysis using the system yielded 2,497 neutral sentiments. The negative sentiment shows the number of 1,160 comments whereas, the sentiment using the system generates 933 comments. And positive comments indicate the number of 924 comments whereas, using the system produced 2,385 positive comments.

Figure 7 Graph of Error With Unigram Analysis

Data of 5,815 comments that were processed resulted in 3,051 data that could be analysed correctly while 2,764 could not be analysed properly. The factors that cause the system can not detect properly are as follows.

a. Users ask with sentences containing negative or positive keywords, so that the system is read as a sentence that means negative or positive. This factor resulted in 1,159 comments that were not properly analysed or 41.9%.
b. Users submit opinions, enter, or just comments that do not intend to praise or denigrate. However, the comments submitted contain negative or positive keywords. Factors that cause the system can not provide the right analysis results. As a result of these factors produce errors of 1,251 comments or 45.3%
c. Comments submitted by users cannot be detected by the dictionary. In fact, the comments made have negative or positive connotations. This factor is caused by the incomplete positive negative dictionary used so that it cannot detect new words. The factor resulted in a total error of 354 comments or 12.8%

From the results of the analysis produced by the author makes Word Cloud to facilitate reading the results of the analysis that has been done. Word cloud is a display of words that form a picture, the larger the word size, the more frequency the word is used. And conversely, if the word size gets smaller the frequency of the word usage is low.

Conclusion

Based on the results of research and discussion of sentiment analysis of PT KAI customer comments through social media twitter using the lexicon based method, it can be concluded that:

a. Lexicon-based method analyzes using tokenization, namely unigram.
b. The results of the analysis using the two methods resulted in an accuracy rate of 52.47% for each unigram tokenization.
c. Factors that influence the results of the analysis include commenting questions that contain negative or positive keywords, comments that contain only opinions but contain negative or positive keywords, and comments cannot be detected by the dictionary because it is incomplete.
d. The results of the errors of the three factors in the unigram tokenisation were 41.9% of the sentence questions, 45.3% of the sentences of opinion, and 12.8% were not detected by the dictionary.

ISBN : 978-602-14917-9-9

Reference

Doshi, K. a., Zhong, T., Lu, Z., Tang, X., Lou, T., & Deng, G. (2013). Blending SQL and NewSQL Approaches: Reference Architectures for Enterprise Big Data Challenges. *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 163–170. https://doi.org/10.1109/CyberC.2013.34

Francis, L., dan Flynn, M. (2010). *Text Mining Handbook, In Casualty Actuarial Society E-Forum*. (2008), 1.

Kamal, A. F. (2017). *TEXT MINING UNTUK ANALISA SENTIMENT EKSPEDISI JASA PENGIRIMAN BARANG MENGGUNAKAN METODE NAIVE BAYES PADA*. 0–1.

Kusumawati, I. (2017). *Analisa Sentimen Menggunakan Lexicon Based Kenaikan Harga Rokok Pada Media Sosial Twitter*.

Liu, B. (2012). Sentiment Analysis and Opinion Mining Morgan & Claypool Publishers. *Language Arts & Disciplines*, (May), 167. https://doi.org/10.1007/978-1-4899-7502-7_907-1

Park, J., Fink, C., Barash, V., & Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 466–475.

Sheela, L. J. (2016). A Review of Sentiment Analysis in Twitter Data Using Hadoop. *International Journal of Database Theory and Application*, *9*(1), 77–86. https://doi.org/10.14257/ijdta.2016.9.1.07

Wahid, D. H., & SN, A. (2016). Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity Devid. *IJCCS, 10*(2), 207–218.