KOMPUTASI: JURNAL ILMIAH ILMU KOMPUTER DAN MATEMATIKA
VOL. 21 (2) (2024), 37-43, p-ISSN: 693-7554, e-ISSN:2654-3990

Sentiment Analysis Of Online Loans On Twitter Using Lexicon Based Methods And Support Vector Machine (SVM)

Cita Suci Saputri[1], Arie Qur'ania[2], Irma Anggraeni[3*]

[1,2,3] Department of Computer Science, Faculty of Mathematics and Natural Science, Universitas Pakuan, Bogor, West Java, 16143, Indonesia

Abstract

Technological developments are increasingly rapid and moving towards digital, which in the end technology can also help people who are experiencing economic problems, namely with online loan services. Even though there are many conveniences provided by online loan services, of course not all people give positive comments because there are quite a few negative comments about this service.One of the social media that is widely used by the public to provide comments about online loans is Twitter. Sentiment analysis is a data processing process to obtain information about whether an opinion sentence tends to be positive, negative or even neutral. This research contains sentiment analysis towards Online Loans on Twitter using the Lexicon Based and Support Vector Machine methods. From the results of this research, the accuracy for SVM was 82.36%. From these results it can be concluded that the use of the Lexicon Based and Support Vector Machine methods is considered quite good and effective for classifying sentiment

Keywords: *Sentiment Analysis; Online Loans; Lexicon Based; Support Vector Machine*

1. Introduction

Online loans are financial service provider facilities that operate online or without having to meet in person. This application or information technology based loan service is a type of Financial Technology (Fintech) Implementation. This is characterized by the use of internet media as a means of transactions when carrying out banking activities [1]. However, the existence of this online loan service certainly raises pros and cons for the public, because apart from legal online loans, now there are also many online loans that are illegal or unofficial and not registered with the Financial Services Authority (OJK). Legal online loans make it easy to get the funds needed with an easy process, while illegal online loans are often detrimental and have a bad impact on many people, especially those who are consumers. This is what makes people pros and cons about online loan services so that online loans or loans have become a much discussed topic at the moment because of the many cases of fraud and giving unreasonable interest to borrowers. One of the social media that is widely used by people to provide comments about online loans is Twitter. Data generated from Twitter can also be very useful if analyzed, because this data can be extracted into important information through opinion mining. Opinions regarding any news or product launch and even certain types of trends can be well observed on Twitter [2].

Sentiment analysis or what can be called opinion mining is the process of understanding, extracting and processing textual data automatically to obtain sentiment information contained in opinion sentences regarding an issue or object by someone, whether it tends to be positive, negative or even neutral. opinion. Lexicon Based is a method commonly used for sentiment analysis on social media, this is because this method is quite practical to use. The method used is Support Vector Machine using Lexicon Based Features as a feature update in addition to using the TF-IDF feature [8] Lexicon Based uses a dictionary as a language or lexical source [3].

Support Vector Machine or commonly abbreviated as SVM is a classification technique that uses 2 points (vectors) which then form a dividing line (if 3 or more dimensions are the dividing side) [4]. Based

[*]**Corresponding author**. *E-mail address:* irmairhamna@unpak.ac.id
Received: 24 June 2024, Accepted: 26 July 2024 and available online 31 July 2024
DOI: https://doi.org/10.33751/komputasi.v21i2.5260

on several studies conducted using the Lexicon Based and Support Vector Machine methods, the results obtained are quite good, therefore in this journal we will discuss Online Loan Sentiment Analysis on Twitter using the Lexicon Based and Support Vector Machine methods.

2. Methods

In this research, to describe the work process of a system, this is by using a flow diagram or program flowchart. Flowcharts are a way of writing algorithms using line notation. A flowchart is a picture or chart that shows the sequence or steps of a program and the relationships between processes and their statements [15]. The method used in this research is using data mining or what can be called KDD (Knowledge Discovery and Data Mining). Knowledge Discovery In Database is a method for obtaining knowledge from existing databases. In the database there are tables that are interconnected/related. The results of the knowledge obtained in this process can be used as a knowledge base for decision making purposes [12].

The research method used uses the following stages which can be seen in Figure 1:

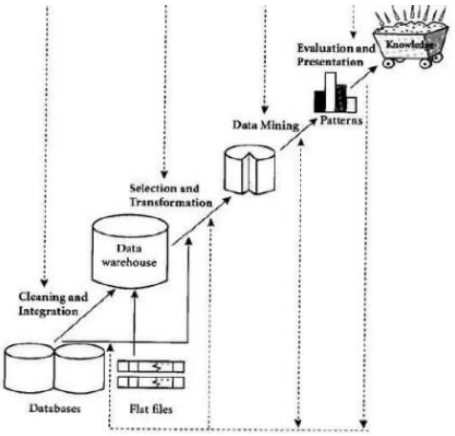

Figure 1. Research methods

2.1. Sentiment analysis

Sentiment analysis is the process of understanding, extracting and processing textual data with the aim of obtaining information. This is done to find out opinions about a problem or can be used to identify trends in the market. There are many benefits of sentiment analysis from various points of view, including that it can be used to obtain a general picture of public perception of service quality, monitor a product, predict sales, politics and investor decision making [5].

2.2. Lexicon Based

Lexicon Based is a feature of words that have positive and negative sentiments based on a lexicon dictionary. The process of labeling sentiment data is carried out by a lexicon based dictionary by calculating the sentiment score. In the lexicon dictionary, every word that contains sentiment has a value/polarity that determines whether the word is a positive or negative word. After knowing which words are positive or negative, then count each word that contains sentiment in the sentence and add up the opinion value. The opinion value determines whether the word has a positive or negative sentiment [6].

Classification using the lexicon method is a classification of tweets based on positive words and negative words in the tweet data results that have been cleaned at the text preprocessing stage. The stages in the Lexicon Based method are automatically contained in Appendix 3. The dictionary used in this research is Inset Lexicon (Indonesian Sentiment Lexicon) [9].

Inset lexicon or Indonesian Sentiment Lexicon has been tested quite well for sentiment analysis on Indonesian language data. Inset Lexicon itself has quite a lot of vocabulary, including 3,609 positive words and 6,609 negative words in Indonesian, each of which has its own weight value or polarity score with a weight range between -5 to +5 [13].

This research carries out labeling automatically using a lexicon dictionary and uses 3 classes in labeling sentiment, namely positive, negative and neutral as in Table 1.

Table 1. Determination of Sentiment Class

Class	Skor
Positive	> 0
Neutral	0
Negative	< 0

2.3. Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF (Term Frequency-Inverse Document Frequency) weighting is a process for transforming data from textual data into numerical data for weighting each word or feature. TF - IDF is a statistical measure used to convey how important a word is in a document. TF is the frequency of occurrence of a word in each given document indicating how important the word is in each document. DF is a frequency document that contains these words showing how common they are. IDF is the inverse of the DF value [10].

2.4. Support Vector Machine

Support Vector Machine or usually abbreviated as SVM is a classification technique that uses 2 points (vectors) which then these 2 points will form a dividing line (if 3 or more dimensions become the dividing side). The dividing line or side formed from these two vectors is called a hyperpane. If SVM is translated into Indonesian, it means a machine that uses vectors as supports/markers to divide data into 2 groups. This method consists of a training process to learn the system and testing to obtain classification results [11].

SVM is a classification using 2 vector points which later these 2 points will form a dividing line called a hyperplane [4]. To find the hyperplane you can use the equation:

$$F(x) = w.x + b \text{ atau } \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) + b \qquad (1)$$

W : The hyperplane parameter to be sought (the perpendicular line between the hyperplane line and the support vector point)
X : Support vector machine input data points
A : The weight value of each data point
Xx, x_i : Kernel functions
B : The sought hyperplane parameter (bias value)

Then look for the linear function with the equation:

$$g(x) = Sign\,(f(x)) \qquad (2)$$

2.5. Confusion Matrix

There are four terms that represent the results of the classification process in the confusion matrix, namely True Positive, True Negative, False Positive, and False Negative [7]. The configuration matrix commonly used for 2 class calculations is shown in Figure 2.

Figure 2. Confusion Matrix

Wich one :

1. True Positive (TP) = The number of positive comments documents and it is true.

2. True Negative (TN) = The number of documents is negative and that is correct.
3. False Positive (FP) = The number of positive comments documents and it is wrong.
4. False Negative (FN) = The number of negative comments documents and it is wrong.

Calculating the confusion matrix in the case of binary classification generally only has 2 classes. However, in this study there were 3 classes, so a 3 x 3 confusion matrix calculation was used.

3. Result and Discussion

The topic raised in this research is online loans or commonly known as pinjol. The aim of creating this system is to provide information to the public to be more careful in making online loans and not to be fooled by unofficial online loans. Data collection in this research was carried out by crawling from Twitter.The tools used for crawling data are the Tweepy library in the Python programming language, the keywords used to pull data are "pinjol" and the programming language used is Python. The tweets data obtained through crawling is 7631 data, then duplicate data will be deleted so that the data is reduced to 1728 data. For non-standard words, words that are short and incorrectly written, a normalization process will be carried out, such as the word "tdk" being changed to "tidak" and many other words so that the data used is more relevant data to proceed to the classification stage.

Apart from that, there are still many numbers, punctuation marks, emoticons and other words that are less important to be used as features. Therefore, a preprocessing process is carried out with the aim of eliminating noise. The preprocessing process is carried out so that the data used is clean from noise, has smaller dimensions and is more structured. So when the data is clean, the data will be ready for further processing [14]. Data preprocessing stages are case folding, tokenizing, filtering, stemming. After preprocessing, the labeling process is done by assigning labels to sentences or documents into classes. Data labeling is carried out using the Lexicon Based algorithm by calculating the positive sentiment score minus the negative sentiment score in the sentence and to determine the score for each word in a sentence, it is matched with a positive and negative dictionary, namely the Inset Lexicon. If the polarity score ≥ 0 then the polarity result of the word is categorized as a positive sentence, if 0 it is neutral and if the polarity score ≤ 0 then the sentence is negative. Then classification is carried out using Support Vector Machine. Based on the analysis, sentiment results were obtained for each class which collected 274 positive sentiment data, 1270 negative data and 184 neutral data. To see the comparison of data from each sentiment class, a graph of the classification results was created which can be seen in Figure 3.

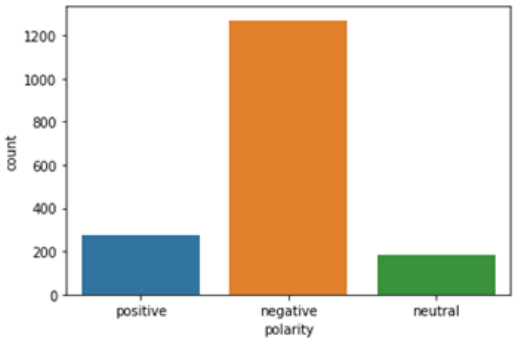

Figure 2. Sentiment Classification Results Graph

Apart from that, visualization is also displayed in a pie chart to find out the percentage of each sentiment. The percentage results for each sentiment were obtained with a total of 73.5% for negative, 15.9% for positive and 10.6% for neutral. The sentiment comparison graph can be seen in Figure 4.

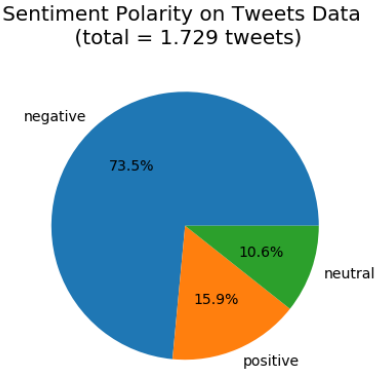

Figure 3. Sentiment Comparison Chart

From the results of the research carried out, it can be seen that the results of the classification graph above show that the average person is tempted by online loans. However, there are still many people who fall into the trap of illegal or unofficial online loans, resulting in losses which can be seen from the number of negative sentiments shown in the graphs in Figures 3 and 4. However, there are still some people who are more careful when it comes to lending and borrowing. For example, preferring official or legal online loans, finding out first about the application or website that will be used for online loans, not being tempted by advertising words such as "low loan interest" which turns out to be just a form of fraud and there are still many other.

Next, we can see the words that appear most frequently in the tweet data using wordcloud. Wordcloud of the entire tweet data. Wordcloud displays words based on the frequency of occurrence of the word, where the more words used in a tweet or comment, the larger the size of the word. The wordcloud of all tweet data can be seen in Figure 5.

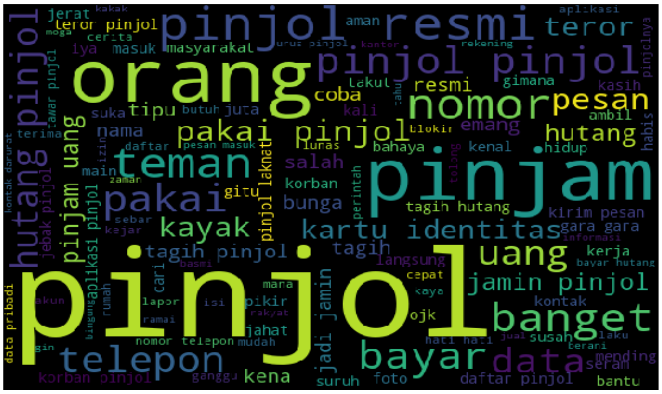

Figure 4. Wordcloud Overall Data

Based on the classification results using the Lexicon Based and Support Vector Machine methods, it produces quite good accuracy as in Figure 6.

```
SVM RESULT
Accuracy Score = 0.8236994219653179
Precision = 0.8117012040485274
Recall = 0.8236994219653179
F-score = 0.8106065202774013
```

Figure 5. Support Vector Machine Classification Results

After classification, the performance of the support vector machine algorithm shows quite good results. The average accuracy result was 82.36%, precision 81.00%, recall 82.00%. The F1-Score value is 81.00%, which is the result of a comparison of the average between the precision value and the weighted recall value. The results of the classification method were validated using the Confusion Matrix as in Figure 7.

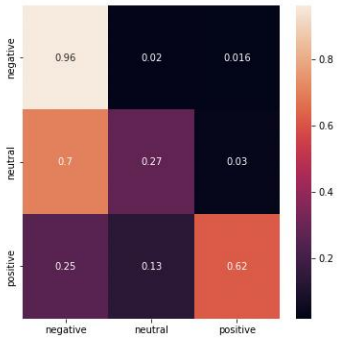

Figure 6. Confusion Matrix Test Results

From the classification results, negative accuracy values are more numerous than positive and neutral accuracy values. Because in the tweet data with the keyword "PINJOL" more negative words appear than positive words, resulting in many sentences with negative sentiment. The negative words in question are terror, bill, debt, victim and many other negative words which are visualized on the wordcloud. However, apart from negative sentences, there are still -+ 20% sentences that contain positive sentiment. This is because quite a few positive words appear in the tweet data that has been crawled. The positive words in question include send, official, receive, profit and others.

4. Conclusion

Based on the research results, it can be concluded that the use of the Support Vector Machine method using 1728 cleaned datasets is quite good in carrying out classification. By using automatic labeling using the Lexicon Based method, then continuing with classification using the Support Vector Machine method and measuring accuracy using the Confusion Matrix to get an accuracy result of 82.36%. With the comment data obtained, there are more comments with negative sentiments, meaning that more loan consumers and even the public feel disadvantaged and do not agree with the existence of online loan services. Because quite a few people are deceived by online loans, people have to be more careful and more careful when choosing online lending and borrowing services.

References

[1] Utami, D.S. & Erfina, A. 2021. Analisis Sentimen Pinjaman Online di Twitter Menggunakan Algoritma *Support Vector Machine* (SVM). Seminar Nasional Sistem Informasi dan Manajemen Informatika. 299-305

[2] Zuhdi, A.M., Utami, E. & Raharjo, S. 2019. Analisis Sentimen Twitter Terhadap Capres Indonesia 2019 Dengan Metode K-NN. Jurnal INFORMA Politeknik Indonusa Surakarta. 5(2) : 1-7.

[3] Mahendrajaya, R., Buntoro, G.A. & Setyawan, M.B. 2019. Analisis Sentimen Pengguna Gopay Menggunakan Metode *Lexicon Based* dan *Support Vector Machine*. Jurnal Teknik Universitas Muhammadiyah Ponorogo. 3(2) : 52-63.

[4] Herlambang, M.B. 2019. *Machine Learning*: *Support Vector Machines*. https://www.megabagus.id/machine-learning-support-vector-machines/. 03 Maret 2022.

[5] Ipmawati, J., Kusrini & Luthi, E.T. 2017. Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen. Indonesian Journal on Networking and Security. 6(1) : 28-36.

[6] Buntoro, G.A. 2017. Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal*. 2(1) : 32-41.

[7] Anggreany, M.S. 2020. *Confusion Matrix*. https://socs.binus.ac.id/2020/11/01/confusion-matrix/. 11 Oktober 2021

[8] Rofiqoh, U., Perdana, R.S. & Fauzi, M.A. 2017. Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonedia Pada Twitter Dengan Metode *Support Vector Machine* dan *Lexicon Based Features*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 1(12) : 1725-1732.

[9] Mahayani, I., Agushinta, D. & Supriyadi, M.E. 2020. Analisis Sentimen Twitter Terhadap Pembayaran *ShopeePayLater* Pada Aplikasi Belanja *Online* (Shopee) Menggunakan Metode *Lexicon Based* dan *Naive Bayes Classifier*. Jurnal Ilmiah KOMPUTASI. 19(4) : 545-558.

[10] Yunus, M. 2020. TF-IDF (*Term Frequency-Inverse Document Frequency*) : Representasi *Vector* Data *Text*. https://yunusmuhammad007.medium.com/tf-idf-term-frequency-inverse-document-frequency-representasi-vector-data-text-2a4eff56cda. 10 Maret 2022

[11] Ramadhan, D.A. & Setiawan, E.B. 2019. Analisis Sentimen Program Acara di SCTV Pada *Twitter* Menggunakan Metode *Naive Bayes* dan *Support Vector Machine*. e-Proceeding of Engineering. 6(2) : 9736-9743.

[12] Mardi, Y. 2016. Data Mining : Klasifikasi Menggunakan Algoritma C4.5. Jurnal Edik Informatika. 2(2) : 213-219.

[13] Statiswaty, Rusnia & Ransi, N. 2017. Analisis Sentimen Wisata Bahari Di Sulawesi Tenggara Memanfaatkan Media Sosial *Twitter* Dengan Menggunakan Metode *Lexicon-Based*. 3(2) : 161-168.

[14] Jumeilah, F.S. 2017. Penerapan *Support Vector Machine* (SVM) untuk Pengkategorian Penelitian. Jurnal Rekayasa Sistem dan Teknologi Informasi. 1(1) : 19-25.

[15] Syaifudin, Y.W. & Irawan, R.A. 2018. Implementasi Analisis *Clustering* Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode *K-Means*. Jurnal Informatika Polinema. 4(3) : 189-194.