Jurnal E-Komtek
Vol. 6, No. 1 (2022) pp. 27-35
https://jurnal.politeknik-kebumen.ac.id/index.php/E-KOMTEK
p-ISSN : 2580-3719 e-ISSN : 2622-3066

Sentiment Analysis on Twitter Using Maximum Entropy : a Case Study on Indosat Ooredoo

Gilang Ryan Fernandes, Ika Mei Lina
Department of Informatic Engineering, Universitas Indraprasta PGRI, Jakarta, Indonesia, 12530

gilang.fernandes@gmail.com

https://doi.org/10.37339/e-komtek.v6i1.911

Published by Politeknik Piksi Ganesha Indonesia

Artikel Info
Submitted:
21-05-2022
Revised:
02-06-2022
Accepted:
05-06-2022
Online first :
30-06-2022

Abstract

The result of the current technological developments makes increasingly tight telecommunication provider competition. Various opinions expressed by customers about telecommunication providers are found in social media. Twitter is widely used by the public to share information and socialize; also, to share opinions, express opinions of a product or service, and provide reviews on communication providers. Many reviews are provided by users on Twitter, making it hard to classify manually. Therefore, to make it easy to classify the tweets, an automation system is needed to determine that a comment is positive or negative. Maximum Entropy can be used for sentiment analysis of Indosat Ooredoo tweets. On these grounds, this study explains how to define tweets into positive and negative classes with applications created using Java. Based on the research and after testing, the Maximum Entropy obtained an accuracy value of 86.21% and an AUC value of 0.968.

Keywords: Twitter, Sentiment analysis, Classification, Data mining, Maximum entropy

Abstrak

Perkembangan teknologi saat ini mengakibatkan persaingan provider telekomunikasi semakin ketat. Berbagai opini yang dikemukakan oleh pelanggan tentang provider telekomunikasi dapat diketahui melalui media sosial. Twitter banyak digunakan masyarakat untuk berbagi informasi dan bersosialisasi, selain itu twitter juga dapat digunakan untuk berbagi pendapat, mengemukakan opini suatu produk, jasa dan memberikan ulasan pada provider komunikasi. Banyaknya ulasan yang diberikan oleh pengguna pada twitter, membuat kesulitan dalam melakukan pengklasifikasian secara manual. Oleh karena itu, untuk memberikan kemudahan dalam mengklasifikasikan tweets, diperlukan suatu sistem otomatisasi untuk menentukan komentar yang bernilai positif atau negatif. Maximum Entropy dapat digunakan untuk analisis sentimen terhadap tweets pada indosat ooredoo. Pada penelitian ini adapun cara untuk menentukan tweets ke dalam class positif dan negatif yaitu dengan aplikasi yang dibuat menggunakan Java. Berdasarkan penelitian dan setelah dilakukan pengujian, maka didapat Maximum Entropy dengan nilai akurasi 86.21% dan nilai AUC sebesar 0.968.

***Kata-kata kunci**: Twitter; Analisis sentimen, Klasifikas, Data mining, Maximum entropy*

1. Introduction

In the current technological developments, many telecommunication providers are emerging. They make the competition so tight that a decision should be given quickly and precisely. The right decision is strongly influenced by opinion analysis. This opinion is information written by users of the product. By utilizing the volume of data generated by product users, it can evaluate public opinion on interesting topics such as products, services, people, concepts, etc [1].

In general, the opinions felt by the community can be a satisfaction statement or even disappointment, specifically for Indosat telecommunications providers. The large numbers of public opinion when conducting sentiment analysis may become a hindrance if it is done by collecting data and determining positive and negative opinions manually. Therefore, one of the methods currently being developed is sentiment analysis. Methods of sentiment analysis is a development of text mining that is still growing and popular today. The method is used to determine an opinion in large numbers and divide them into two: positive and negative classes.

Sentiment analysis or opinion mining is a field of study that analyzes people's opinions, sentiments, evaluations, attitudes, judgments, and emotions towards the entity of each product, organization service, and attribute owned and expressed in text form [2]. Twitter has nearly 600 users and more than 250 million tweets per day; this is a golden opportunity for organizations to see their reputation and products by extracting and analyzing sentiments from tweets posted by the public about products and competitors [3].

Based on the background of the problem, then the problem is difficulties in getting information about customer opinions on Indosat products on Twitter due to the many records and unstructured data.

The purpose of this study is to obtain the accuracy of the application of the Maximum Entropy Method to produce fast and accurate information and help in analyzing the customer or community sentiment through tweets on Twitter.

This study discusses the opinion mining or sentiment analysis from the public to Indosat by utilizing data sources from Twitter, so it obtained information or knowledge that is hidden and not yet processed from a data and makes it valuable that can be utilized for various purposes.

2. Method

Research method is a technique for searching, obtaining, collecting, and recording data, either primary or secondary, that is used in preparing a scientific paper and then analyzing the factors related to the subject matter to get the correct data. Research is an activity that aims to create an original contribution to science. There are four commonly used research methods: Action Research, Experiment, Case Study, and Survey [4]. This research is an experimental study. Experimental research is an investigation of causal relationships using tests controlled by the researchers themselves [4].

a. Sample Selection Methods

Sampling method aims to retrieve data based on sources by using primary data, that is, data obtained from the original source or main source collected by data collection techniques in the field. For taking the data sample, this study employed a simple random sampling. Random sampling is the method of sampling that gives the same opportunity to be taken to every element of the population. So, every member of the population has the same chance to be chosen.

b. Method of collecting data

The data collected is a collection of tweets taken from Twitter based on the keyword @IndosatCare using Apache Nifi tools. The data were then classified into positive tweets and negative tweets. The data were as many as 500 tweets. Apache Nifi is developed by the Apache Software Foundation, a web-based system that enables the management and distribution of information. This system allows the user to build their own process flow that is used to obtain the desired information. Apache nifi is a data platform for automating the movement of data between different systems. Apache nifi provides real-time control that makes it easy to manage the movement of data between any source and destination **[5]**.

c. Maximum Entropy Algorithm

This Maximum Entropy method uses the weight vector to determine the classification of sentiments **[6]**. The Maximum Entropy algorithm is a training process first. The training aims to calculate the weight of words in each sentence in the processed tweets. Maximum Entropy combines contextual evidence to examine a particular linguistic class that occurs within a linguistic context. In the classification is done observation linguistic context $b \epsilon B$ and predicted linguistic class $\alpha \epsilon A$ with conditional probability distribution p, in which one $p(a|b)$ is the probability of a class with multiple contexts b, where p is **[7]**:

$$H(p) = -\sum_{x \epsilon \varepsilon} P(x) \log P(x)$$

X is a combination of a and b, x=(a,b), $\alpha \epsilon$ A, b ϵ B, and ε = AxB. Having obtained the probability of distribution p, the probability is used to determine the evidence. Evidence is represented by functions known as contextual predicates and features. If A={a1 … aq} is a set of classes to be predicted, and B represents a set of contexts to be observed, the contextual predicate function is:

$$cp : B \rightarrow \{true, false\}$$

which will return true or false depending on whether or not information is useful for some contexts b ϵ B. Contextual predicates are used in features, in which the function is fj: AxB → {0,1}. Given feature k number, then Ep fj = Ep fj. Where 1 ≤ j ≤ k. Epfj is the expectation model p from fj,

$$E_{\tilde{p}} f_j = \sum_{x \epsilon \varepsilon} \tilde{P}(x) f_j(x)$$

where P is the observed probability x of the S sample training. Then, the p model is consistent with the observed evidence and if it only meets the constrain k. The principle of the Maximum Entropy used is:

$$P = \{p \mid E_P f_j = E_{\tilde{p}} f_j, j = \{1..k\}\}$$

$$P^* = \underset{p \epsilon P}{\arg max}\, H(P)$$

d. Evaluation and Classification Method Validation Data Mining

Data Mining is a statistical, mathematical, artificial intelligence, and machine learning technique for extracting and identifying useful information contained in large databases [8]. K-fold cross-validation is one of the variations of the cross-validation testing technique. K-fold cross-validation is done by dividing the training and the test set. The whole point of validation for this type of data is randomly split into subsets. K-fold cross-validation repeat k times to divide a set of random samples into k subsets most free. Each repetition has left a subset of a subset of testing and other training.

K subsets of the selected one subset into test data and (k-1) into training data. This process is repeated as many as k, where each k subsets exist in the test data and the rest in the training data. With cross-validation variations such as random sub-sampling repeated validation is all the data will be used both to test data and for training data.

An illustration of the k-fold cross-validation can be seen in **Figure 1 [9]**.

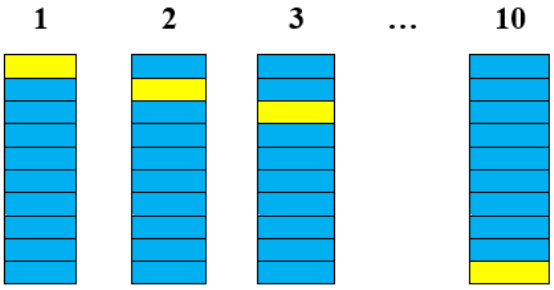

Figure 1. Illustration of K-Fold Cross Validation **[9]**

Information:

: Test Data

: Training Data

Tests on the classifier system should proceed as expected and can produce high accuracy values and produce the lowest possible error values. Then the test can be done by using a confusion matrix. The confusion matrix provides an assessment of classification performance by object correctly or incorrectly. For example, if there are classification cases for output data such as "yes" and "no"; "good" and "ugly"; and so on, each predicted class has four different possible positives (TP); and True negatives (TN) denotes classification accuracy. If the output prediction is positive, while the original value is negative, it is called false positive (FP); if the output prediction is negative, while the original value is positive, it is called false negative (FN), as illustrated in **Table 1**.

Table 1. Confussion Matrix for Two Class [10]

Classification	Predicted Class	
	Class = Yes	Class = No
Class = Yes	A(true positive - TP)	B(false negative - FN)
Class = No	C(false positive - FP)	D(true negative - TN)

The following is the equation in calculating the equation of confusion matrix model to calculate accuracy, recall, and precision:

1. Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$
2. Recall (pos) = $\frac{TP}{TP+FP}$
3. Precision (pos) = $\frac{TP}{TP+FP}$
4. Recall (neg) = $\frac{TN}{FN+TN}$
5. Precision (neg) = $\frac{TN}{FN+TN}$

3. Results and Discussion

The results of application development and performance measurement models are explained in this section. The application development is discussed in the test to show that the results of the application are as expected. The model performance measurement describes the results of a performance measurement model for the analysis of sentiment tweets Indosat ooredoo users by using Maximum entropy on the application developed.

a. Testing Model with 10-Fold Cross Validation

In this research, the model testing is done by using a 10-fold cross-validation technique, where this process divides data randomly into 10 parts. The testing process begins with the establishment of a data model in the first section. The model formed will be tested on the remaining 9 sections of the data. After that, the accuracy process is calculated by looking at how much data has been correctly classified.

b. Design and Implementation

This paper designed the application to test the model using dataset tweets Indosat ooredoo. Here is the explanation of the appearance of the application design and its implementation.

1) Dashboard

This dashboard view is the initial appearance of the application program created in this study. Initial view (dashboard) is used to enter (upload) dataset formatted JSON and to display all data that has been processed. Dashboard is presented in **Figure 2**.

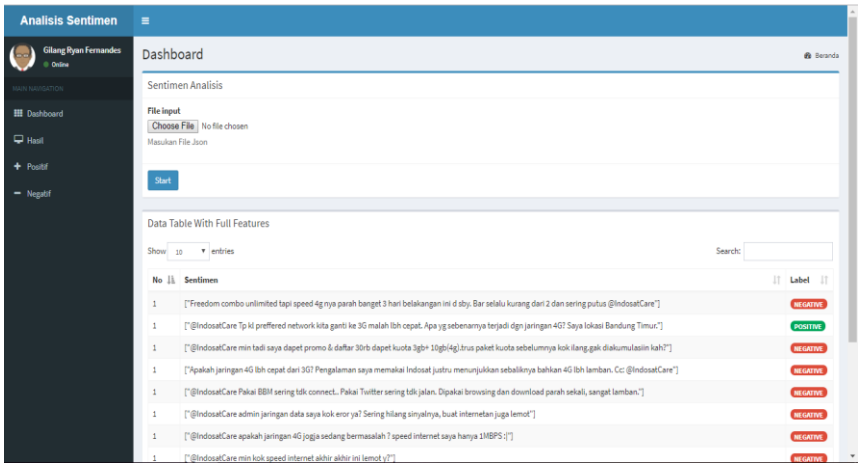

Figure 2. Dashboard

2) Results Views

The result view is the view to see the total number of tweets, diagrams of data that has already in process, and the percentage of the positive and negative sentiments. Results view is presented in **Figure 3**.

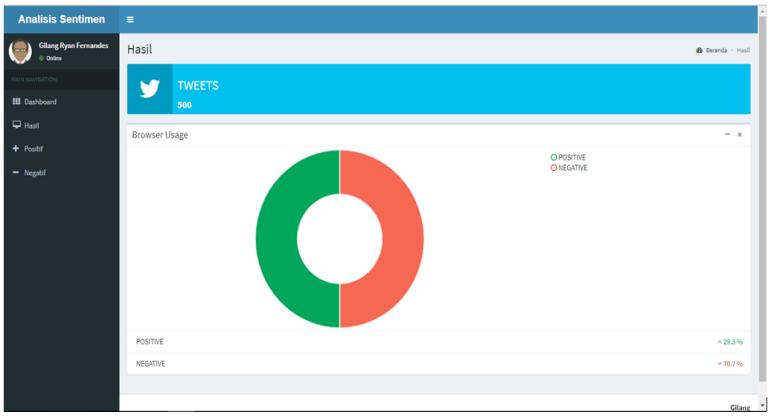

Figure 3. Results Views

3) Positive Views

This positive view shows the results of positive tweets that have been processed and displayed in tabular form. The positive sentiment contains tweets that are kind and praiseworthy. Positive views is presented in **Figure 4**.

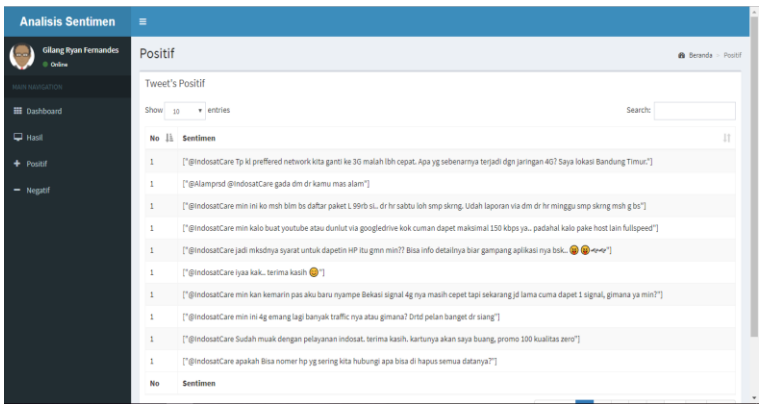

Figure 4. Positive Views

Here are some of the Positive tweets is presented in **Table 2**.

Table 2. Result of Positive Tweets

No	Text	Sentiment
1	"@IndosatCare iyaa kak.. terima kasih"	positive
2	"@IndosatCare min untuk wilayah Bekasi signal 4g nya masih cepet dan lancar"	positive
3	"@IndosatCare min ini 4g emang cepat, buat youtube lancar"	positive
4	"@IndosatCare pelayananya bagus, terima kasih"	positive
5	"@IndosatCare sanagt puas ganti kartu di geray indosat, pelayananya ramah dan sopan"	positive

4) Negative Views

This negative display shows the results of negative tweets that have been processed and displayed in tabular form. Negative sentiment contains tweets that express disappointment and criticism. Negative views is presented in **Figure 5**.

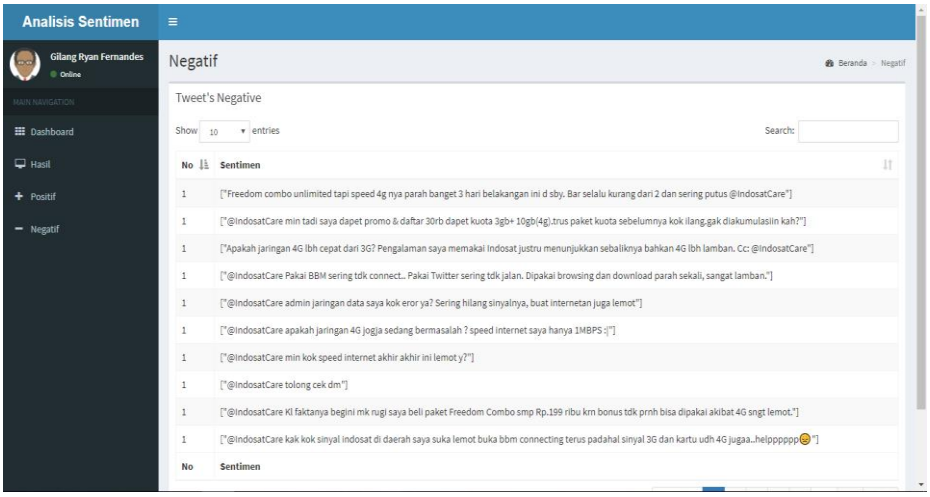

Figure 5. Negative Views

Here are some of the Negative tweets is presented in **Table 3**.

Table 3. Result of Negative tweets

No	Text	Sentimen
1	*"Freedom combo unlimited tapi speed 4g nya parah banget 3 hari belakangan ini d sby. Bar selalu kurang dari 2 dan sering putus @IndosatCare"*	negative
2	*"@IndosatCare min tadi saya dapet promo, daftar 30rb dapet kuota 3gb+ 10gb(4g). trus paket kuota sebelumnya kok ilang. gak diakumulasiin kah?"*	negative
3	*"Apakah jaringan 4G lbh cepat dari 3G? Pengalaman saya memakai Indosat justru menunjukkan sebaliknya bahkan 4G lbh lamban. Cc: @IndosatCare"*	negative
4	*"@IndosatCare Pakai Twitter sering tdk jalan. Dipakai browsing dan download parah sekali, sangat lamban."*	negative
5	*"@IndosatCare admin jaringan data saya kok eror ya? Sering hilang sinyalnya, buat internetan juga lemot"*	negative

4. Conclusion

The conclusions that are generated based on this research are Maximum Entropy method is one of the methods that can be used in doing sentiment analysis on Twitter in the form of tweets with the keyword @IndosatCare and can classify tweets into positive class and negative class. The stages in the process performed in the sentiment analysis of the initial data processing (text preprocessing), incuding tokenizing, stopwords removal, and stemming.

After that, the Maximum Entropy classification method was applied to classify the tweets into positive class and negative class. Based on the results of the research and after testing, the Maximum Entropy can run well with an accuracy of 86.21% and the value of AUC of 0.968.

References

[1] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, and T. Varvarigou, "Comparing Methods for Twitter Sentiment Analysis," *Proc. Int. Conf. Knowl. Discov. Inf. Retr.*, 2014.

[2] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.

[3] F. H. Khan, S. Bashir, and U. Qamar, *TOM: Twitter opinion mining framework using hybrid classification scheme*. Islamabad: Decision Support Systems, National University of Sciences and Technology (NUST), 2014.

[4] C. Dawson, *Projects in Computing and Information Systems. London: Addison Wesley*. 2009.

[5] "Secure government applications Apache Hadoop | Cloudera." https://www.cloudera.com/products/open-source/apache-hadoop/apache-nifi.html (accessed May 24, 2022).

[6] A. Go, Alec, R. Bhayani, and L. Huang, *Twitter Sentiment Classification using Distant Supervision*. 2009.

[7] S. Ester Irawati, "Komparasi Algoritma Untuk Analisa Sentimen Review Produk Pada Twitter," *Sekol. Tinggi Tek. Surabaya*, 2015.

[8] E. Turban, J. E. Aronson, and T. Peng, *Decision Support Systems and Intelligent Systems*. Yogyakarta: Andi Offset, 2005.

[9] R. Payam, T. Lei, and L. Huan, *Cross-validation. In: Encyclopedia of Database systems*. 2009.

[10] Adelheid and Andrea, *Cara Curang Menambah Follower Twitter*. Jakarta: Media Kita, 2013.