Bit-Tech

Vol.4, No.2, Decermber 2021
Available online at: http://jurnal.kdi.or.id/index.php/bt

Twitter Opinion Mining Analysis of Web-Based Handphone Brand Using Naïve Bayes Classification Method

Suryadi Wijaya [1)*], Yo Ceng Giap [2)]

[1)2)]*Universitas Buddhi Dharma*
Jl.Imam Bonjol No. 41 Karawaci Ilir, Tangerang, Indonesia
[1)]suryadiw08@gmail.com
[2)]cenggiap@ubd.ac.id

Article history:

Received 5 December 2021;
Revised 15 December 2021;
Accepted 18 December 2021;
Available online 30 December 2021

Keywords: {use 4-6 keywords}

Social Media
Twitter
Handphone
Sentiment Analysis
Naïve Bayes Classification Method

Abstract

Social Media is now very commonly used for the benefit of society. People mostly use social media to convey information, give opinions, even for media to express themselves. One of the social media that is widely used to convey this information is Twitter. From the use of Twitter, a public opinion tweet emerged about a mobile phone product. The more that is posted on Twitter about cellphones, the more public opinion will arise about cellphone brands. From these opinions, a classification is needed that can distinguish Neutral, Negative, or Positive Opinions. Sentiment analysis or opinion mining is one part of text mining that can help with these problems. In connection with the above, an application is designed that can analyze sentiment analysis from Twitter using the Naïve Bayes classification method. The results of the application of the Naïve Bayes classification method will result in a classification of sentiments into neutral, negative, or positive opinions.

I. INTRODUCTION

Along with the times, social media is now very commonly used for the benefit of the community. The growth of social media continues to increase along with its ease of use anywhere, whether it is based on mobile applications or websites. One of the social media that is widely used to convey this information is Twitter. Twitter is a popular social media among the public, with a maximum use of 140 characters which makes people more creative in conveying information or opinions. Because of this, one of them is used by the official Mobile Review account (Handphone) to convey information about several mobile phone brands that are emerging today. From the use of twitter, public opinion tweets emerged about a mobile phone product.

According to the Big Indonesian Dictionary, opinion is an opinion, thought, or position. The more that is posted on twitter about cellphones, the public's opinion on cellphone brands will arise. From this opinion, a classification is needed that can distinguish the type of opinion in question. Sentiment analysis or opinion mining is one part of text mining that can help problems classifying opinions that are neutral, positive, or negative.

Research in the field of opinion mining has been carried out on the automotive car market. Based on related research, it can be concluded that opinion mining is carried out to see opinions that produce various accuracy values [1].

In this study, the Naïve Bayes classification method is used where the theory is a simple classification algorithm but has high accuracy. This study aims to classify neutral, positive, or negative sentiments regarding public opinion on cellphone brands circulating on Twitter social media.

II. RELATED WORKS/LITERATURE REVIEW (OPTIONAL)

Twitter

Twitter is a social networking site that is currently growing rapidly because users can interact with other users from their computers or mobile devices from anywhere and anytime. After its launch in July 2006, the number of Twitter users recorded was around 160 million users [2]. Various kinds of benefits can be obtained from tweets starting from

* Corresponding author

ISSN 2622-2728 (online) 2622-271X (print)
doi:

incident detection (event detection, one of which is natural disasters), prediction of market movements, election predictions to the spread of disease in an area. For example, to predict stock market movements, analysis is done by analyzing tweets containing positive and negative moods related to the stock market such as the Dow Jones, S&P 500, NASDAQ. To benefit from this abundance of tweets, of course, research and analysis of existing tweets is needed, one of which is data mining research that destroys data from tweets.

Sentiment Analysis

According to [3], sentiment analysis or often referred to as opinion mining is a computational study to identify and express opinions, sentiments, evaluations, attitudes, emotions, subjectivity, judgments, or views contained in an object text.

Data Mining

According to [4], Data Mining is the process of finding interesting patterns and knowledge from large amounts of data. Meanwhile, according to Linoff and Berry (2011: 7), Data Mining is a search and analysis of a very large amount of data and aims to find the meaning of patterns and rules. And according to Vercellis (2009:77). Data mining is an activity that describes an analytical process that occurs iteratively on a large database, with the aim of extracting accurate and potentially useful information and knowledge for knowledge workers related to decision making and problem solving [5].

Classification Algorithm

In helping classification work there are several classification algorithms that have been compiled by several research experts. Based on the training method, classification algorithms can be divided into two types, namely eager learners and lazy learners [6].

The algorithms included in the eager learner are designed to perform reading/training/learning on the training data in order to correctly map each input vector to its output class label. So that at the end of the training period, the model can map all the test data vectors to the output class label correctly [7]. After the training process is complete, the model is stored as memory, while the training data is not used. The prediction process will run fast but the training process is quite long. Included in this algorithm are Support Vector Machine, Decision Tree, Neural Network and Bayesian.

On the other hand, lazy learners do little training (or not at all, only store part or all of the training data, then use it in the prediction process. The prediction process becomes slower because they have to read all the training data. The algorithms that fall into this category are K -Nearest Neighbor, Linear Regression, etc.

According to [3], because sentiment analysis is classifying text, the most suitable algorithms are the Naïve Bayes algorithm and Support Vector Machine (SVM). The Naïve Bayes algorithm is a simple probabilistic-based prediction technique based on the application of the Bayes theorem with the assumption of strong or naive independence [6]. While the SVM algorithm is a technique the results are more promising and provide a better method than the others but are more complicated. From the above comparison, the Naïve Bayes algorithm is used with consideration of simplicity

Naïve Bayes

Naïve Bayes is a classification with probability and statistical methods proposed by the British scientist Thomas Bayes. According to [8], explaining Naïve Bayes for each decision class, calculates the probability on the condition that the decision class is true, given the object's information vector. This algorithm assumes that object attributes are independent. The probabilities involved in producing the final estimate are calculated as the sum of the frequencies from the "master" decision table. Meanwhile, according to [4], the process of The Naïve Bayesian Classifier, or Simple Bayesian Classifier, is as follows:

1. Simple Bayesian Classifier, is as follows:
2. Variable D becomes the training set of tuples and labels associated with the class. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x1, x2, …, xn), this represents the n measurements made on the tuple of n attributes, respectively, A1, A2, …, An.
3. Suppose there are classes m, C1, C2, …, Cm. Given a tuple, X, the classifier predicts that X that belongs to the group having the highest posterior probability, the condition is specified in X. That is, the nave Bayesian classifier predicts that the X tuple belongs to class Ci if and only if:

$$P(Ci|X) > P(Cj|X) \text{ for } 1 \leq j \leq m, j \neq i$$

Formula Classifier Naïve Bayesian (1)
Source : (Han dan Kamber, 2011:351)

So maximizing P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximal posteriori hypothesis. By Bayes' theorem:

$$\boldsymbol{P(Ci|X)} = \frac{\boldsymbol{P(X|Ci)P(Ci)}}{\mathbf{P(X)}}$$

Formula Classifier Naïve Bayesian (2) [4]

Information:

P(Ci|X) = The probability of the hypothesis Ci if given facts or records X (Posterior probability)
P(X|Ci) = look for the parameter value that gives the greatest possibility (likelihood)
P(Ci) = Prior probability of X (Prior probability)
P(X) = The number of probability tuples that appear

1. When P(X) is constant for all classes, only P(X|Ci) P(Ci) needs to be maximized. If the probability of the previous class is not known, it is generally assumed to be in the same class, namely P (C1) = P (C2) = … = P (Cm), therefore it will maximize P (X | Ci). Otherwise, it will maximize P(X | Ci) P(Ci). Note that the pre-class probability can be estimated by P(Ci) = | Ci, D | / | D | where | Ci, D | is the number of training tuples of class Ci in D.
2. Given that the dataset has many attributes, it will be very difficult to compute to calculate P(X|Ci). In order to reduce computations in evaluating P(X|Ci), a nave assumption of conditional class independence is made. Assuming that the values of the attributes are conditionally independent of each other, given the class label of the tuple (i.e. that there is no dependency relationship between the attributes) thus:

$$\boldsymbol{P(X|Ci)} = \prod_{\boldsymbol{k=1}}^{\boldsymbol{n}} \boldsymbol{P\,(Xk|Ci)}$$

$$= \boldsymbol{P\ (x1|Ci)\ x\ P(x2|Ci)x\ \ldots x\ P(Xn|Ci)}$$

Formula Classifier Naïve Bayesian (3) [4]

Then it can easily estimate the probabilities P(x1 | Ci), P(x2 | Ci), … , P(xn | Ci) from tuple training. Note that here xk refers to the value of attribute Ak for tuple X. For each attribute, see whether the attribute is categorical or continuous-valued. For example, to calculate P(X|Ci) consider the following:

a) If Ak is categorical, then P(Xk|Ci) is the number of tuples of class Ci in D having a value of Xk for attribute Ak, divided by |Ci,D|, the number of tuples of class Ci in D.
b) If Ak is continuous-valued, then it needs to do a bit more work, but the calculations are pretty simple. A continuous-valued attribute is usually assumed to have a Gaussian distribution with a mean and a standard deviation , defined by:

$$\boldsymbol{g(x, \mu, \sigma)} = \frac{\mathbf{1}}{\sqrt{\boldsymbol{2\pi\sigma}}}\boldsymbol{e} - \frac{\boldsymbol{(x-\mu)2}}{\boldsymbol{2\sigma 2}}$$

Formula Classifier Naïve Bayesian (4)
Source : (Han dan Kamber, 2011:351)

$$\mathbf{P}(\boldsymbol{x|Ci}) = \boldsymbol{g(xk, \mu Ci, \sigma Ci)}$$

Formula Classifier Naïve Bayesian (5) [4]

After that, calculate Ci and Ci, which are the mean (mean) and standard deviations of each k attribute value for the Ci class training tuple. Then use the two quantities in the equation, together with xk, to estimate P(xk | Ci)

3. To predict the class x label, P(X|Ci)P(Ci) is evaluated for each Ci class. The classifier predicts that the label class of tuple x is class Ci, if

P(X|Ci)P(Ci) > P(X|Cj)P(Cj) for 1 ≤ j ≤ m, j≠i

Formula Classifier Naïve Bayesian (6) [4]

In other words, the predicted class label is Ci where P(X | Ci) P(Ci) is maximum.
Bayesian classifier has minimal error rate compared to other classifications. However, in practice this is not always the case, due to inaccurate assumptions made for its use, such as class independent conditions, and the lack of available probability data. Bayesian classifiers are also useful in providing theoretical justification for other classifiers that do not explicitly use Bayes' theorem.

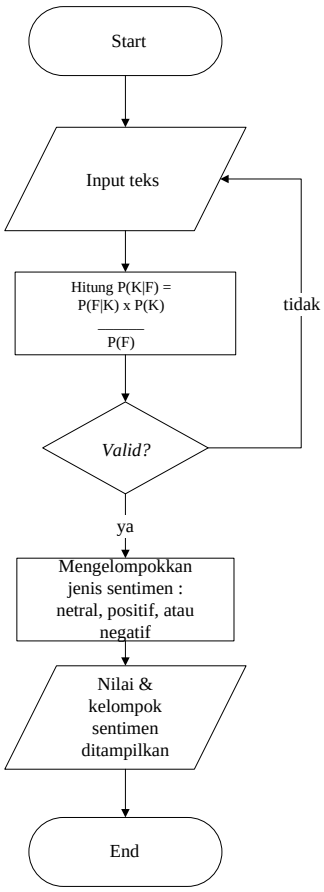

Fig. 1 Naive Bayes Application Calculation Flowchart

Application Programming Interface (API)

Application Programming Interface (API) is a technology to facilitate the exchange of information or data between two or more software applications. An API is a virtual interface between two software functions that work together, such as between a word processor and a spreadsheet. An API defines how programmers take advantage of a certain feature of a computer. APIs are available for windowing systems, file systems, database systems and network systems. The development of API technology begins with the creation of a simple subroutine that provides interoperability and system modifiability to support data exchange between multiple applications. The subroutine is only able to carry out simple mathematical calculations until a calculation library API is formed which is almost always present in every programming language. From a simple subroutine, ideas began to emerge on how the API should be developed, especially in line with the development of object-oriented programming paradigms which resulted in a collection of similar subroutines being collected into a wrapper class for these subroutines.

With the development of a software into a system consisting of several other software (subsystems) then the API is also developing to keep realizing its goal, namely as a bridge between software. The development of the API is realized by increasing the nature of the API which is able to support interoperability between software. API is not only tasked with exchanging data and information between subroutines in a software but also exchanging data and information between software. In this case, the API must have the ability to communicate between processes either through file intermediaries, sockets, or other IPC services.

In the development of a wider system, sometimes an API can become a middleware, which is a separate subsystem that has functions that are useful by other subsystems and to access these functions requires a separate connection to the middleware. Connection to middleware is generally done using sockets. There are several standard protocols for accessing the middleware. Examples of middleware access protocols include:

a) Remote Procedure Calls (RPC), a software user calls procedures or subroutines running on a remote middleware, procedure calls can be either synchronous or asynchronous.

b) Message Oriented Middleware (MOM), a system that pools data and information into a middleware, the data waiting to be processed by the subsystems contained in the entire system in that middleware.
c) Object Request Broker (ORB), this protocol allows a software to send and receive objects and request services on an object-based middleware.
d) Structured Query Language (SQL), protocol as well as language for reading and writing data stored in database middleware.

In the process of developing a software, both API in particular and other software requires rules that need to be understood and applied. A good API is an API that has the following properties:

1. Easy to learn
2. Easy to use, even without accompanying documentation
3. Hard to abuse
4. High performance in completing the task
5. Easy to develop further
6. Web

According to [9], the Web is a system with information presented in the form of text, images, sound, and others stored on an Internet Web server that is presented in the form of hypertext. The Web can be accessed by Web client software called a browser. Browsers read web pages stored on a web server via a protocol called HTTP (Hypertext Transfer Protocol).

According to [9], HTML is a markup language for disseminating information on the Web. When designing HTML, this idea was taken from the Standard Generalized Markup Language (SGML). Although HTML is not easily understood by most people, when it is published its users become clear. HTTP is a stateless communication protocol based on TCP which was originally used to retrieve HTML files from web servers when it was designed in 1991.

URL (Uniform Resource Locator). URL is composed of three parts:

a) Transfer Format
b) Hostname
c) Document file path
d) For example, the URL could be: http://university.com/index.html

III. RESULTS

As a result of the study, data was obtained from a questionnaire that had been filled out by 20 respondents and it was found that the application that had been made was easy to use and could be used to determine the classification of sentiments in a simple but fairly accurate manner, the questionnaire yielded a score of Strongly Agree as much as 80.7% to the application created.

IV. CONCLUSIONS

From the results of making Twitter Opinion Mining Analysis of Web-Based Mobile Brands Using the Naïve Bayes Classification Method, the following conclusions are obtained: 1). Based on the tweet data used as training data, the Naïve Bayes method was successful in classifying 15 data from the 20 data tested. So that the Naïve Bayes method is successful in classifying twitter sentiment with an accuracy percentage of 75%. 2). The Naïve Bayes Classification Method can be applied as a method for classifying sentiment analysis. 3). In the classification process it will be more accurate depending on the amount of training data.

REFERENCES

[1] D. Rustiana and N. Rahayu, "Analisis Sentimen Pasar Otomotif Mobil: Tweet Twitter Menggunakan Na"ive Bayes," *Simetris J. Tek. Mesin, Elektro Dan Ilmu Komput.*, vol. 8, no. 1, pp. 113–120, 2017.
[2] O. Chiang, "Twitter hits nearly 200M accounts, 110M tweets per day, focuses on global expansion," *Forbes Mag. Retrieved from http//blogs. forbes. com/oliverchiang/2011/01/19/twitter-hits-nearly-200m-users-110m-tweets-per-day-focuses-on-global-expansion*, 2011.
[3] L. Bing, *Sentiment Analysis and Opinion Mining :* Morgan & Claypool Publisher, 2012.
[4] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
[5] A. Hermawan, Y. Kurnia, Riki, B. Daniawan, and L. Septarina, "Design of community-based regional language dictionary platforms in Indonesia with the autocomplete method," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 5 Special Issue, pp. 1849–1857, 2019.
[6] E. Prasetyo, "Data mining konsep dan aplikasi menggunakan matlab," *Yogyakarta Andi*, 2012.

[7] Y. Kurnia, Y. Isharianto, Y. C. Giap, A. Hermawan, and Riki, "Study of application of data mining market basket analysis for knowing sales pattern (association of items) at the O! Fish restaurant using apriori algorithm," in *Journal of Physics: Conference Series*, 2019, vol. 1175, no. 1, doi: 10.1088/1742-6596/1175/1/012047.
[8] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science \& Business Media, 2008.
[9] J. Simamarta, *Rekayasa Perangkat Lunak*. Yogyakarta: Andi, 2010.