ISSN 2622-626X

Naive bayes algorithm performance for smartphone sentiment analysis in social media

Monalisa Fatmawati Sarifah [a,1,*], Iqbal [b,2], Ilham Mubarog [c,3]

[a, b, c] Dept.of Informatics Engineering MTI AMIKOM University Yogyakarta, Indonesia
[1] monafsaa@gmail.com *; [2] Iqbal.iq@students.amikom.ac.id ; [3] Ilhammubarog@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received 2018-11-10
Revised 2018-11-29
Accepted 2018-12-30

Keywords
Smartphone
Classification
Comment
Naïve Bayes

ABSTRACT

Indonesia with a population of 250 million is a large market, Millennials tend to be more adaptive to the development of communication technology [1]. There are lot of opportunities that are used by various groups, one of which is the need to use smartphones that can make it easier for people to exchange information [2]. The shift in sales of smartphone brands in Indonesia is influenced by massive advertising carried out by smartphone vendors (smartphone capitalists) to consumers [3]. The enthusiasm of the community in welcoming this platform is so great, lot of comment about smartphone brand stated by public is an interesting thing to be processed to be information. Utilization of that information requires analytical techniques so that the produced information can help many parties. The method used in this study is Naïve Bayes classification method which is a learning technique for data mining algorithms that uses probability and statistical methods [4]. This method is used to classify comments given by the community to smartphone brands. The comments given in this application will later be classified into positive, negative, and neutral comments. The purpose of this study was to find out how much positive, negative and neutral comments the community gave to smartphone brands, so that later it would facilitate the smartphone brand in providing policies or development in the future.

1. Introduction

The development of smartphone in Indonesia, is a pioneer for the growth of mobile applications in Indonesia, Minister of Research and Technology said the number of smartphone users in Indonesia now reaches around 25% of the total population or around 65 million people [5]. This rate is a very interesting number for market opportunities, so that it is used by many people, some people certainly do give positive comments on the smartphone brand, some people comment negatively, therefore a technique and method is needed to know the percentage classification of comments given by the community towards smartphone brands [6]. Classifications are divided into negative, positive and neutral comments. The percentage of negative, positive, and neutral comments is needed by various groups such as society, government, even the company itself [7]. This percentage can be needed to make a policy or plan for future development for the company.

Statistical data mentions the number of smartphone users worldwide from 2014 to 2020. About 15% of all smartphone sold to consumers choose iOS as their operating system. The leading smartphone vendors are Samsung and Apple, with around 20% to 25% and 15% of their respective shares, followed by Huawei, OPPO and VIVO [8]. Sentiment analysis itself or also commonly called opinion mining is one part of text mining. This field studies about people's opinions, sentiments, evaluations, behaviors and emotions towards an entity such as products, services, organizations, individuals, problems, topics, events and attributes [9].

The study conducted by Ginting (2018) the study aims to determine the success of these customers. The data is analyzed and calculated using the naive bayes algorithm. Naive bayes algorithm is used to classify data in a particular class, then the pattern can be used to estimate debtors who want to join, so the company can make a decision to accept or reject the debtors selection [10].

2. Method

2.1. Sentiment Analysis

According to Medhat et al (2014), sentiment analysis is a field that takes focus in text-based research, is opinion mining with a study about ways to prevent problems from public opinion, attitudes and emotions of an entity, where the entity can represent individuals [11]. Sentiment analysis is the process of understanding, extracting and processing textual data automatically to obtain sentiment information contained in an opinion sentence. Sentiment analysis is carried out to see opinions or inclinations of opinions on a subject or object by someone, whether they tend to have positive or negative opinions [12].

2.2. Naïve Bayes Algorithm

Naïve Bayes is a classification with probability and statistical methods proposed by British scientist Thomas Bayes, it is used to predicting opportunities in future based on previous experience so that it is known as the Bayes theorem [13]. The theorem is combined with "naive" where it is assumed that the conditions between attributes are mutually independent [14]. In a dataset, each row / document I is assumed to be a vector of attribute values <x1, x2, ..., x3> where each value becomes a review of the attributes of Xi (iЄ [1, n])). Each line has a ci class label Є {c1, c2, ..., ck} as the value of the class C variable, so the classification can be calculated the probability value p (C = ci | X = xj), because at Naïve Bayes each attribute is assumed to be free , then the equation obtained is as follows: Opportunity p (C = ci | X = xj) shows the opportunity for the attribute Xi with value xi given class c, where in Naïve Bayes, class C is of qualitative type while attribute Xi can be qualitative or quantitative.

When the Xi attribute type is quantitative, the probability of p (X = xi | C = cj) will be very small that the opportunity equation cannot be relied on for quantitative type attribute problems. So to handle quantitative attributes, there are several approaches that can be used such as normal distributions (Gaussian):

$$\hat{f} = N(X_i; \mu_c, \sigma_c) = \frac{1}{\sqrt{2\pi}\sigma_c} e^{-\frac{(X_i-\mu_c)^2}{2\sigma_c^2}},$$

There is another mechanism for dealing with quantitative attributes (numerical), named discretization. Discretization process occurs during data preparation process or when data is preprocessing, where X numeric attributes are changed to nominal attributes X*. Naive Bayes classification performance will be better when numerical attributes are discretized than assumed by the distribution approach as above (Dougherty). Numerical values will be mapped to values in the form of intervals which still pay attention to the class of each numeric value that is mapped, Naive Bayes' calculations depicted as follows [15]:

$$p(I=i_j|C=c_i) = \frac{p(I=ij)\ p(C=ci|I=ij)}{p(C=ci|)}$$

Explanation: p(I=ij|C=ci) : interval opportunities i-j for ci class
p(C=ci|I=ij) : ci class opportunities for i-j interval
p(I=ij) : probability interval of -j on all formed interval
p(C=ci) : chance of an -i class for all classes in the dataset

3. Research Methodology

The method used in this study is Naïve Bayes Classifier. The steps taken in this study are as follows:

3.1 Data Collecting

Tweets data taken in 1 month period, specifically in November to December 2018 collected from Twitter social media. Choosen tweets are using Indonesian, taken randomly from normal user or online twitters online.

3.2 Labeling

This study is label manually by three Twitter user. 1000 tweets data will be given which will be classified into three category, neutral, positive, negative. Every documents will be classified.

3.3 Data Prepocessing

Cleaning process and filtering process [16]. In the cleaning process, attributes that are less influential on the classification process are reduced. The data entered at this stage is still in the form of raw data, so the results of this process are in the form of quality documents that will simplify the classification process. The last stage of data preprocessing is the filtering process. Filtering aims to delete words that are less important for process classification.

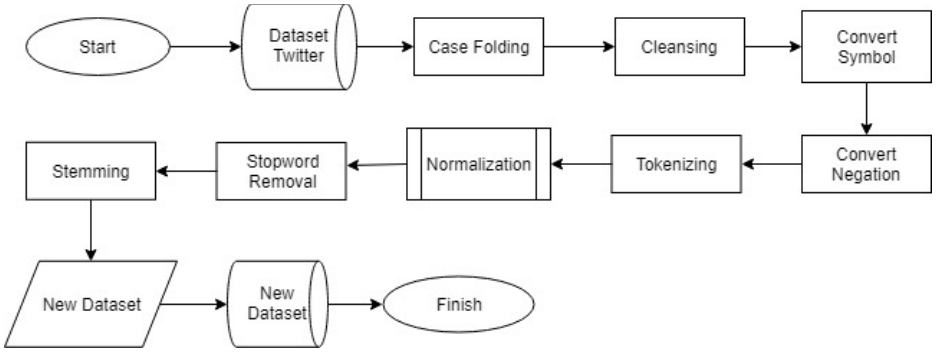

Fig 1. Flow Prepocessing

3.4 Cleansing

In the study of Feitosa et al. (2015) cleaning was carried out to arrange words that were not approved at all [17]. The research that will be carried out will also do data cleaning. The word removal phase has no effect on the results of the sentiment classification. The tweet document component has various attributes that do not affect sentiment. Examples of nonessential attributes are ('#'), links that start ('http', 'bit..ly') and symbol characters (~! @ # $% ^ & * () _ + {}: "<> ?) Attributes that have no effect will be removed then replaced with character spaces.

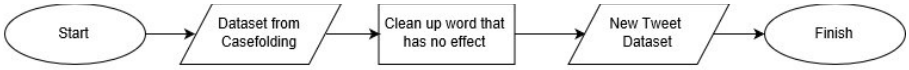

Fig 2. Flow Cleaning

3.5 Convert

In the study of Al-khurayji et al (2017) convert using for negation words like "no", "no", "don't" [18]. The convert stage affects the sentiment of a document, because the symbol will describe a person's feelings when sad or happy. Negation words like "no", "no", "don't" will also change the meaning of sentiment.

3.6 Tokenizing

Words reduction based on each word that composes them into single pieces. The word document in question is a word separated by spaces. So the outcome of this process is a single word that is entered into the database for weighting purposes.

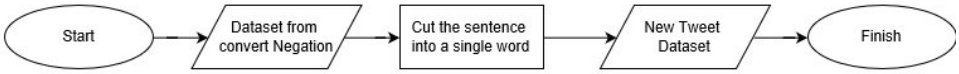

Fig 3. Flow Tokenizing

3.7 Normalization and Stopword Removal

Words then convert based on good spelling, so that it will reduce document sentiment. Convertion taken from short word, standardization, and foreign. Clearing up words which is not appropriate to topic, if words do not affect sentiment classification accuracy.

3.8 Term

The data of tweets that have been deleted and filtered, is carried out the conversion process into vector shapes. The process of converting tweet data into vector form. After that the weighting of each word (term) aims to get the word value successfully extracted.

3.9 Classification

The classification process is done using an application and the method used is Naïve Bayes Classifier. Testing data using cross validation [19]. Divided into two processes in the form of a training process and process testing.

3.10 Evaluation

The classification process is done using an application [20]. The classification method used in this study is naïve bayes classifier. Evacuation is done by seeing the results of accuracy, precision and memory.

4. Results and Discussion

In the testing section, the data used from reviews of smartphone brands. The sample negative and positive review data will be reviewed in table 1.

Table 1. Comparison of Accuracy

Positif	Negatif	Netral
191	72	566

Table 2. Sample Data

Sentimen	Review
Negative	Dengan harga kisaran segitu saya bisa mendapatkan hp dengan spek yang bagus. hp ini hanya mahal harga dengan kualitas di bawah ratarata
Possitive	warning infinity display ram 4gb dari handphone ini bisa menyebabkan foto lebih jelas. bermain game pun lancar pokoknya saya senang dengan produk ini
Netral	kece abis dong bisa dapet ini ada hadiah buat kamu

4

The classification results obtained show of the 830 tweets of data collected there were 566 neutral tweets, 72 negative comments, and 191 positive comments on brand smartphone in Indonesia. That classification chart. From table 1, show the accuracy of the percentage and can be illustrated in the graph approved by Figure 4.

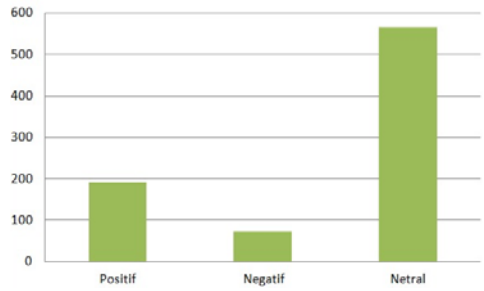

Fig 4. Graph of The Precision

5. Conclusion

The dataset in this study was obtained from Indonesian language tweets with the topic of smartphone brands on social media. The data is taken randomly both from online media users on Twitter. After getting the tweet data, the data is labeled. In this study the labeling process was done manually. With 830 tweets that will be classified as neutral, positive and negative. The next data is through the classification process using the naïve bayes classifier method. The accuracy of the Bayes naive classifier is accurate.

Suggestions from this study can be denied by using a more complete dictionary of keywords to replace unnecessary words. So words that are processed in pre-preparation are important. Can also train with more data with more training data too. For more research, it is expected that more data will be used and from this much data, more accurate results will be obtained.

Acknowledgment

First of all praise and gratitude to God Almighty, because of His blessing the author can complete this research. The author wishes to thank these insights for the reviews and suggestions from reviewers in the previous version of this paper. Their comments have substantially helped to the paper improvement.

References

[1] Poushter, J. (2016). Smartphone ownership and internet usage continues to climb in emerging economies. *Pew Research Center*, *22*, 1-44.

[2] Steven, S. (2018). Consumer Dependence On Smart Phones: The Effect Of Social Needs, Social Influence And Convenience In Surabaya. *Calyptra*, *7*(1), 839-857.

[3] The shift in sales of smartphone brands in Indonesia is influenced by massive advertising carried out by smartphone vendors (smartphone capitalists) to consumers.

[4] Jiang, L., Li, C., Wang, S., & Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence*, *52*, 26-39.

[5] Wulandari, N., & Sari, R. K. (2016). Linking Experiential Value To Loyalty In Smartphone Industry. *Studies And Scientific Researches. Economics Edition*, (24).

[6] Aggrawal, N., Ahluwalia, A., Khurana, P., & Arora, A. (2017). Brand analysis framework for online marketing: ranking web pages and analyzing popularity of brands on social media. *Social Network Analysis and Mining*, *7*(1), 21.

[7] Esparza, G. G., Díaz, A. P., Canul-Reich, J., De-Luna, C. A., & Ponce, J. (2016). Proposal of a Sentiment Analysis Model in Tweets for improvement of the teaching-learning process in the classroom using a corpus of subjectivity. *International Journal of Combinatorial Optimization Problems and Informatics*, *7*(2), 22-34.

[8] Assemi, B., Safi, H., Mesbah, M., & Ferreira, L. (2016). Developing and validating a statistical model for travel mode identification on smartphones. *IEEE Transactions on Intelligent Transportation Systems*, *17*(7), 1920-1931.

[9] Kumar, P., & Vardhan, M. (2018). Aspect-Based Sentiment Analysis of Tweets Using Independent Component Analysis (ICA) and Probabilistic Latent. *Advances in Data and Information Sciences: Proceedings of ICDIS 2017*, *2*, 3.

[10] Ginting, S. L. B., Adler, J., Ginting, Y. R., & Kurniadi, A. H. (2018, August). The Development of Bank Application for Debtors Selection by Using Naïve Bayes Classifier Technique. In *IOP Conference Series: Materials Science and Engineering* (Vol. 407, No. 1, p. 012177). IOP Publishing.

[11] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113.

[12] Park, M. S. (2014). Configurable Accelerators for Visual and Text Analytics.

[13] Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

[14] Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (icml-03)* (pp. 616-623).

[15] Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, *4*(4), 816-820.

[16] Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

[17] Feitosa, S. A., Patel, D., Borges, A. L., Alshehri, E. Z., Bottino, M. A., Özcan, M., ... & Bottino, M. C. (2015). Effect of cleansing methods on saliva-contaminated Zirconia—An evaluation of resin bond durability. Operative dentistry, 40(2), 163-171.

[18] Al-khurayji, R., & Sameh, A. (2017). An Effective Arabic Text Classification Approach Based on Kernel Naive Bayes Classifier. International Journal of Artificial Intelligence Applications, 01-10.

[19] Wu, J., Pan, S., Zhu, X., Cai, Z., Zhang, P., & Zhang, C. (2015). Self-adaptive attribute weighting for Naive Bayes classification. Expert Systems with Applications, 42(3), 1487-1502.

[20] Feng, W., Sun, J., Zhang, L., Cao, C., & Yang, Q. (2016, December). A support vector machine based naive Bayes algorithm for spam filtering. In Performance Computing and Communications Conference (IPCCC), 2016 IEEE 35th International (pp. 1-8). IEEE.