http://ejournal.seaninstitute.or.id/index.php/InfoSains
Jurnal Info Sains : Informatika dan Sains, Volume 13, No 02 2023
E-ISSN.2797-7889, P-ISSN.2089-3329

Multi-Language Sentiment Analysis Using Machine Learning

[1]Parasian D.P Silitonga, [2]Limrot Imran Purba, [3]Irene Sri Morina, [4]Alex Rikki
[1]Fakultas lmu Komputer , Universitas Katolik Santo Thomas, [2]Fakultas lmu Komputer , Universitas Katolik Santo Thomas, [3]Sistem Informasi Rumah Sakit, RSUP Haji Adam Malik, [4]Fakultas lmu Komputer , Universitas Katolik Santo Thomas,
Email: parasianirene@gmail.com[1], limrotpurba7@gmail.com[2], morina_ginting@yahoo.com[3], alexrikisinaga@gmail.com[4]

Keywords

Sentiment Analysis, Multi Language, Machine Learning, Lake Toba Tourism

Abstract. Sentiment analysis is the interpretation and classification of user emotions (positive, negative, neutral) about a subject in text data using text analysis. Multilingual sentiment analysis is the process of assessing sentiment in more than one language. The tricky thing about being multilingual is that the emotions and behavior of our consumers are heavily influenced by culture and language. Therefore, for organizations with an international customer or user base, sentiment analysis is highly recommended to perform analysis not only in one language but in many languages. This is because the accuracy of the assessment will be better if it is done in more than one language. There are several methods that can be used to perform sentiment analysis, one of which is machine learning. Machine Learning is used as a tool to produce robots that are able to classify types of sentiment in textual data. This research was conducted to produce a machine learning model that can be used to measure the level of popularity of research objects based on comments written in two languages, namely Indonesian and English. The research was conducted based on comments on Twitter about the Lake Toba tourist attraction. Based on the results of testing the Naïve Bayes model on data testing, it shows that sentiment with positive predictions is 1,474 records or 54.03% and sentiment with negative predictions is 1,254 or 45.96%, with an accuracy rate of the method used at 97.1%.

1. INTRODUCTION

Sentiment analysis is the interpretation and classification of users' emotions (positive, negative, neutral) about a subject in text data using text analysis [1]. With the help of sentiment analysis, unstructured information can be converted into more structured data which can then be used to explain people's opinions regarding products, brands, services, politics, or other topics. Companies, governments, and other fields then use this data to carry out marketing analysis, product feedback, and community services [2].

Multilingual sentiment analysis is the process of assessing sentiment in more than one language. The tricky thing about being multilingual is that the emotions and behavior of our consumers are heavily influenced by culture and language. Therefore, for organizations with an international customer or user base, sentiment analysis is highly recommended to perform analysis not only in one language but in many languages. This is because the accuracy of the assessment will be better if it is done in more than one language [3].

There are several methods that can be used to perform sentiment analysis, one of which is machine learning. Machine Learning is used as a tool to produce robots that are able to classify types of sentiment in textual data [4]. Machine learning is a branch of artificial intelligence that has the ability to access existing data on its own command. Machine learning is able to study existing data and perform certain tasks and is able to learn given algorithms and statistical models.

The tourism industry is a business providing services for tourism traffic with the aim of seeking profits in the fields of accommodation, culture, restaurants, recreation and entertainment, travel, tour guides, souvenirs, and currency trading [5]. Lake Toba is one of the main tourist attractions in North Sumatra Province, which is located in Samosir Regency. The largest volcanic lake in Southeast Asia has beautiful natural views. Lake Toba is also one of the geological tourism areas that has been recognized by UNESCO and is now one of the five Super Priority Destinations (DSP) launched by the government [6].

Based on this description, this research was conducted to produce a machine learning model that can be used to measure the level of popularity of research objects based on comments written in two languages, namely Indonesian and English. The research was conducted based on comments on Twitter about the Lake Toba tourist attraction.

2. METHOD

2.1. Sentiment Analysis

Sentiment analysis is the interpretation and classification of users' emotions (positive, negative, neutral) about a subject in text data using text analysis [7]. With the help of sentiment analysis, unstructured information can be converted into more structured data which can then be used to explain people's opinions regarding products, brands, services, politics, or other topics. Companies, governments and other fields then use this data to carry out marketing analysis, product feedback and community services [8].

Sentiment analysis, often also referred to as opinion mining, is a field that analyzes people's opinions, sentiments, evaluations, assessments, attitudes and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes [9]. There are many terms that are often mentioned even though in principle there are several differences in them, such as sentiment analysis, opinion mining, opinion extraction, subjectivity analysis, emotion analysis and so on. But all of these terms are basically in the field of sentiment analysis or opinion mining.

Currently there is a lot of information in text form on the internet in the format of forums, blogs, social media, and sites containing reviews. The text information is divided into 2 groups, namely facts and opinions. Facts are objective information about an object or situation, while opinion is subjective information that can be in the form of feelings, appreciation or assessment of an object. The positive or negative value of an opinion can be used to measure the size of the opinion giver's support for a problem topic [10].

By utilizing data found on the internet, companies, governments and other fields then use it to produce marketing analysis, product reviews, product feedback and community services. Subjective language processing which is manifested as sentiments, beliefs and judgments is a developing field in natural language processing [11]. This development cannot be separated from the increasing significance of informal information sources, such as blogs and twitter, user review sites and the rapid growth of online social networks. To produce the required analysis, sentiment analysis defines several elements of a text, including objects, attributes, opinion givers, opinion orientation, and strength of opinion [12].

2.2. TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) weighting is a process for transforming data from textual data into numeric data for each word or feature to be weighted. TF-IDF is a statistical measure used to evaluate how important a word is in a document [13]. TF is the frequency of occurrence of a word in each given document indicating how important that word is in each of those documents. DF is the frequency of documents that contain the word indicating how common the word is. IDF is the inverse of the DF value. The result of word weighting using TF-IDF is the multiplication result of TF multiplied by IDF. The word weight is greater if it appears frequently in a document and is smaller if it appears in many documents [14].

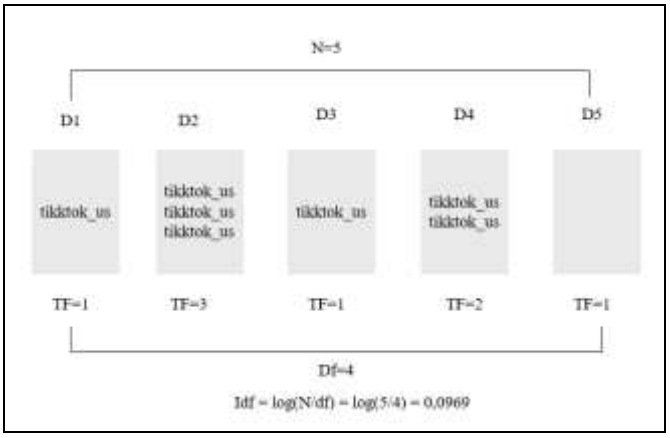

Figure 1. Illustration of the TF-IDF Algorithm [15]

2.3. Naive Bayes

Naïve Bayes is a simple probability classification method that applies Bayes' theorem with a high degree of independence. The use of the Naive Bayes method in this study is based on the large number of datasets used, so it requires a method that has fast performance in classification and high accuracy [16]. Naive Bayes is one of the algorithms used for text classification and is a machine learning method that uses probability and statistical calculations proposed by Thomas Bayes. The algorithm is used to predict future probabilities based on past experience. The advantage of using naive Bayes is that this method only requires a small amount of training data to determine the parameter estimates needed in the classification process [17].

The naive Bayes method takes two stages in the text classification process, namely the training stage and the testing stage. The training process is used for sentiment analysis models which aim to guide classification with testing data or different data. The comparison calculation between the terms in the testing data and each existing class can be made using equation 1 [18].

$$P(w_i|C) = \frac{count(w_i,c)+1}{count(C)+|V|} \quad \text{.....(1)}$$

Where,

C = class category tested
d = documents
wi = the i word
w(i, c) = number of words wi in C
Count(c) = words in class C
|V| = number of words

3. RESULTS AND DISCUSSION

3.1. Data Collecting

The data source used in this research is crawled text data from Twitter social media using the Twitter search attribute for those who have connected the Twitter API to get an 'Access Token', Select Attributes to select data that is crawled, only comment text, text that is already crawled it is named the name 'LakeToba.csv crawl data'. Data collection was conducted in English and Indonesian. The implementation of the data collection process model is presented in Figure 2.

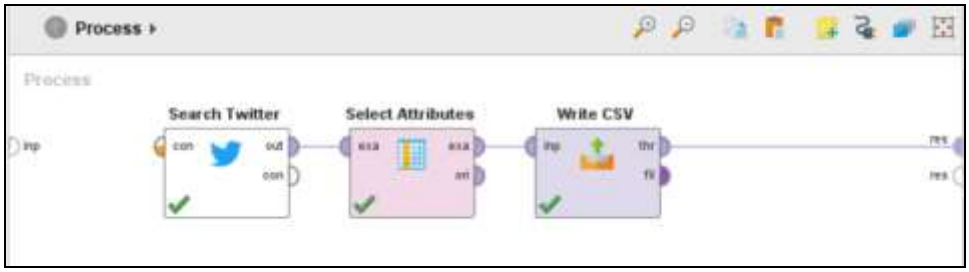

Figure 2. Data Crawling

3.2. Data Translation

The data translation process is a process for translating the data obtained into English. The data translation process is carried out using Google Translate. The translation process is carried out to map data based on polarity in English. The results of the data translation are shown in Figure 3 and Figure 4.

Figure 3. Data Translation

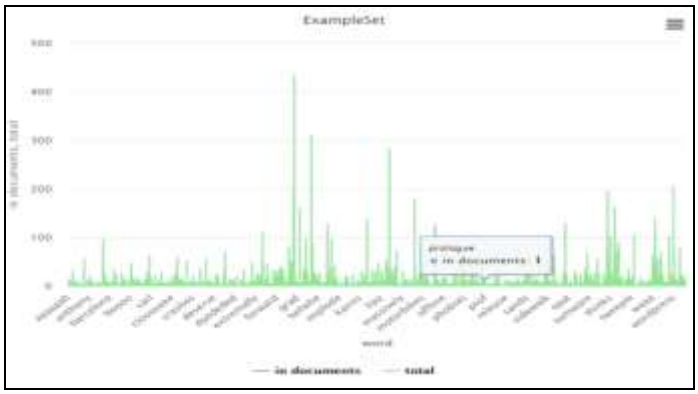

Figure 4. Data Visualitation

3.3. Model Desain

Sentiment analysis in this study was carried out using the Naïve Baye method. The naive Bayes method takes two stages in the text classification process, namely the training stage and the testing stage. The training process is used for sentiment analysis models which aim to guide classification with testing data or different data. If given sample data as in Table 1.

Table 1. Data Sampel

No	Sentiment	Text	Data Categories
1	Positive	the most fun traveling of this year	Training
2	Negative	entirely predictable and no good	Training
3	?	predictable with no fun	Testing

In Table 1, it is known that there are 2 training data with 2 classes, namely "Positive", "Negative. The following is the probability calculation for the "Positive" class. Probability is symbolized as p. Calculating the prior probability of the positive and negative classes.

p(Positif) = ½ = 0,5
p(Negatif) = ½ = 0,5

Based on the results of data pre-processing, words that do not appear in the training dataset are discarded, such as the word "with". Based on the training data, there are 3 words left that are predicted in the testing data, namely "predictable "no" and "fun". Based on these 3 words, the results of the Naïve Bayes calculation are as follows :

$$P(\text{"predictible"}|negative) = \frac{1+1}{5+12} = 0.118$$
$$P(\text{"no"}|negative) = \frac{1+1}{5+12} = 0.118$$
$$P(\text{"fun"}|negative) = \frac{0+1}{5+12} = 0.059$$
$$P(\text{"predictible"}|positive) = \frac{0+1}{7+12} = 0.053$$
$$P(\text{"no"}|positive) = \frac{0+1}{7+12} = 0.053$$
$$P(\text{"fun"}|positive) = \frac{1+1}{7+12} = 0.105$$

The process of calculating the posterior probability using the equation from the Naïve Bayes formula is as follows :

$$P(Negative|Training) = 0{,}5 * 0{,}118 * 0{,}118 * 0{,}059 = 0{,}000410758$$
$$P(Positive|Training) = 0{,}5 * 0{,}053 * 0{,}053 * 0{,}105 = 0{,}0001474725$$

Based on the Naïve Bayes calculation process, it is concluded that the training data is in the negative class. Below are presented the results of modeling data that is labeled positive and negative using the Naïve Bayes algorithm in Figure 5.

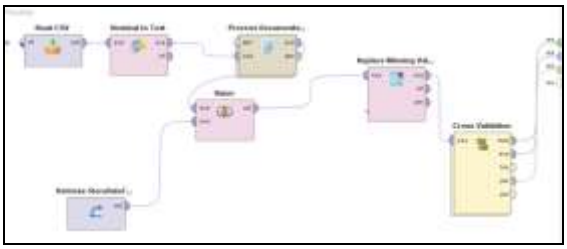

Figure 5. Naïve Bayes Model Design

Based on the results of pre-processing data processing, after sorting based on the highest occurrence of data from each document, we obtained the words with the highest occurrence of 2 words as in Figure 6.

Figure 6. List of Highest Occurring Words

Based on the results of testing the Naïve Bayes model on testing data of 2,859 records, it shows that sentiment with positive predictions was 1,474 records and sentiment with negative predictions was 1,254, and there were 131 records that had no predictions. Sentiment prediction results are presented in Figure 7 and Figure 8.

Figure 7. Naïve Bayes Sentiment Analysis Prediction Results

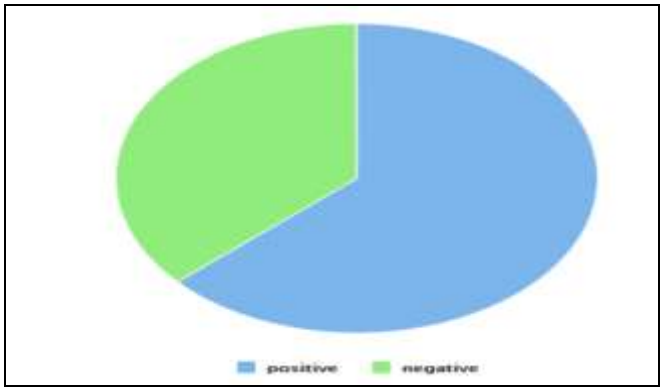

Figure 8. Graph of Comparison of Positive Sentiment and Negative Sentiment

The accuracy of the sentiment analysis model built in this study was measured using cross validation, and resulted in an accuracy rate of 97.1%. The true positive prediction rate is 96.39% and the true negative accuracy rate is 100%. The results of the model accuracy process are presented in Figure 9 and Figure 10.

accuracy: 97.72% +/- 3.23% (micro average: 97.71%)

	true positive	true negative	class precision
pred. positive	160	0	100.00%
pred. negative	6	96	94.12%
class recall	96.39%	100.00%	

Figure 9. Accuracy of the Naïve Bayes Model

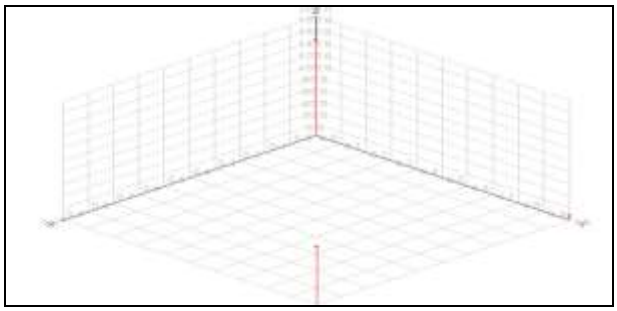

Figure 10. Confusion Matrix Accuracy Level

4. CONCLUSION

Based on the results of testing the Naïve Bayes model on data testing, it shows that sentiment with positive predictions is 1,474 records or 54.03% and sentiment with negative predictions is 1,254 or 45.96%. Thus, in general, based on the results of the sentiment analysis, Lake Toba tourism is positive with an accuracy rate of 97.1%.

REFERENCES

[1] M. A. Sghaier and M. Zrigui, "Sentiment analysis for Arabic e-commerce websites," 2016. doi: 10.1109/ICEMIS.2016.7745323.

[2] S. Bandari and V. V Bulusu, "Survey on Ontology-Based Sentiment Analysis of Customer Reviews for Products and Services," *Advances in Intelligent Systems and Computing*, vol. 1079. Computer Science and Engineering, JNTU Hyderabad, Hyderabad, India, pp. 91–101, 2020. doi: 10.1007/978-981-15-1097-7_8.

[3] "Multilingual Sentiment Analysis: How to Do It | Crisol." https://www.crisoltranslations.com/our-blog/multilingual-sentiment-analysis/ (accessed Oct. 02, 2022).

[4] A. Reyes-Menendez, J. R. Saura, and C. Alvarez-Alonso, "Understanding #worldenvironmentday user opinions in twitter: A topic-based sentiment analysis approach," *Int. J. Environ. Res. Public Health*, vol. 15, no. 11, 2018, doi: 10.3390/ijerph15112537.

[5] J. R. Saura, A. Reyes-Menendez, and P. Palos-Sanchez, "A feeling analysis in Twitter with machine learning: Capturing sentiment from #BlackFriday offers ," *Espacios*, vol. 39, no. 42, 2018, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055057928&partnerID=40&md5=830513edad8b62fe8adf0de560d53a07

[6] "9 Tempat Wisata Samosir, dari Perbukitan hingga Desa Halaman all - Kompas.com." https://travel.kompas.com/read/2021/08/05/211000827/9-tempat-wisata-samosir-dari-perbukitan-hingga-desa?page=all (accessed Oct. 02, 2022).

[7] K. Gaurav and P. Kumar, "Consumer satisfaction rating system using sentiment analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10595 LNCS. Department of Computer Science and Engineering, National Institute of Technology Patna, Patna, India, pp. 400–411, 2017. doi: 10.1007/978-3-319-68557-1_35.

[8] S. Uma Maheswari and S. S. Dhenakaran, "Sentiment analysis on social media big data with multiple tweet words," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 10, pp. 3429–3434, 2019, doi: 10.35940/ijitee.J9684.0881019.

[9] C. C. Yang and Y. C. Wong, "Mining consumer opinions from the Web," in *WEBIST 2008 - 4th International Conference on Web Information Systems and Technologies, Proceedings*, 2008, vol. 2, pp. 187–192. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-58049163023&partnerID=40&md5=4e53a7a66104dd890951718dbd99926a

[10] M. H. Krishna, K. Rahamathulla, and A. Akbar, "A feature based approach for sentiment analysis using SVM and coreference resolution," in *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2017*, 2017, pp. 397–399. doi: 10.1109/ICICCT.2017.7975227.

[11] P. D. Silitonga and R. Damanik, "Perbandingan Algoritma k-Nearest Neighbors (k-NN) dan Support Vector Machines (SVM) untuk Klasifikasi Pengenalan Citra Wajah," *J. ICT Inf. Commun. Technol.*, vol. 20, no. 1, pp. 186–191, 2021, doi: 10.36054/jict-ikmi.v20i1.354.

[12] V. P. Lijo and H. Seetha, "A distributed approach for tweet polarity detection," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 10 Special Issue, pp. 891–900, 2019, doi: 10.5373/JARDCS/V11SP10/20192884.

[13] B. Thanasopon, N. Sumret, J. Buranapanitkij, and P. Netisopakul, "Extraction and evaluation of popular online trends: A case of Pantip.com," in *2017 9th International Conference on Information Technology and Electrical Engineering, ICITEE 2017*, 2017, vol. 2018-Janua, pp. 1–5. doi: 10.1109/ICITEED.2017.8250454.

[14] H. Zhang, Y. Qin, and X. Lv, "Conditional random field in the application of the product feature extraction," in *PIC 2016 - Proceedings of the 2016 IEEE International Conference on Progress in Informatics and Computing*, 2017, pp. 128–132. doi: 10.1109/PIC.2016.7949480.

[15] M. Qasem, R. Thulasiram, and P. Thulasiram, "Twitter sentiment classification using machine learning techniques for stock markets," in *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015*, 2015, pp. 834–840. doi: 10.1109/ICACCI.2015.7275714.

[16] S. Das and A. K. Kolya, "Sense GST: Text mining & sentiment analysis of GST tweets by Naive Bayes algorithm," in *Proceedings - 2017 3rd IEEE International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 2017*, 2017, vol. 2017-Decem, pp. 239–244. doi: 10.1109/ICRCICN.2017.8234513.

[17] R. Sanda, Z. K. A. Baizal, and F. Nhita, "Opinion mining feature-level using Naive Bayes and feature extraction based analysis dependencies," in *AIP Conference Proceedings*, 2015, vol. 1692. doi: 10.1063/1.4936448.

[18] K. L. Santhosh Kumar, J. Desai, and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," 2017. doi: 10.1109/ICCIC.2016.7919584.