Plagiarism Checker X Originality Report

Similarity Found: 20%

Date: Sunday, May 31, 2020
Statistics: 534 words Plagiarized / 2658 Total words
Remarks: Medium Plagiarism Detected - Your Document needs Selective Improvement.

The 6th International Conference on Cyber and IT Service Management (CITSM 2018) Inna Parapat Hotel – Medan, August 7-9, 2018 Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies Siti Ernawati STMIK Nusa Mandiri Jakarta rna2103@gmail.com Eka Rini Yulia STMIK Nusa Mandiri Jakarta ekariniyulia@gmail.com Frieyadie STMIK Nusa Mandiri Jakarta frieyadie@nusamandiri.ac.id Samudi STMIK Nusa Mandiri Jakarta samudi.net@gmail.com Abstract- Opinion rivalry that occurs in social media have an important role in increasing the potential customers to the company or agency.

The review is a rich and useful resource for marketing, social and others for excavations and mining opinions such as views, moods, and behavior. The reviews describe perceptions of something, such as review of a product, review of airline services, reviews of restaurant and others. The analysis of sentiment is an ongoing field of text-based research.

The analysis of sentiment or opinion mining is the study of ways to solve problems of public opinion, attitudes, and emotions of an entity, in which the entity may represent individuals, events or topics. Sentiment analysis is an important tool for analyzing opinions in social media. This measurement begins with pre-processing consisting of tokenizing, stopwords removal and stemming.

This study uses naïve Bayes algorithm and genetic algorithms as applied feature selection. Selection features aim to classify text for the review of online fashion companies. This measurement results in the classification of text in form of positive text and negative text.

Measurements are based on the accuracy of naïve Bayes before addition of genetic algorithms and after addition of genetic algorithms as feature selection. Validation using 10 fold cross-validation. For measurement accuracy using confusion matrix and ROC curve. The purpose of the study is to calculate the increased accuracy of naïve Bayes algorithm if using genetic algorithms for feature selection.

The results showed that the genetic algorithm was able to improve the accuracy. KeyWord—sentiment analisys, fashion online companies, text classification, genetic algorithm, naïve bayes I. INTRODUCTION The review is a rich and usef ul resource for marketing, social and others for excavations and mining opinions such as views, moods, and behavior.

The reviews describe perceptions of something, such as review of the product, review of airline services, reviews of restaurant and others. A review can describe the views, attitudes or nature of someone about something. Review available on the internet we can use to be processed in order to produce a knowledge and useful information.

The available reviews are a very useful resource in various fields, such as marketing, social and others[1]. Some of the studies that have been conducted on the review include, Analysis sentiment for restaurant reviews using naïve Bayes algorithm [2]. The main problem in text classification is the high dimension of the feature space, this is often the case with text that has tens of thousands of features.

Most of these features are irrelevant and not used for text classification can even reduce accuracy and a hi gh number of features can slow down the classification process or even make some classifiers inapplicable [3]. Many consumers are expressin g their experiences through social media such as facebook, twitter or other media sites.

An online sales company review is a channel that connects consumers to another, they can express opinions about the company in which they have made a purchase transaction. The purpose of the study is to calculate the increased accuracy of naïve Bayes if using genetic algorithms for feature selection. II. LITERATURE A. Sentiment Analysis Detection of text sentiment has attracted much attention and has grown rapidly in recent years, due to the increased availability of online reviews in digital form. The review is a rich and useful resource for marketing, social and others for excavations and mining opinions such as views, moods, and behavior.

For example, whethe r a review of a positive or negative product, how the mood among bloggers at that time, how the public reflects on political affairs [1]. The analysis of

sentiment is an ongoing field of text-based research. The analysis of sentiment or opinion mining is the study of ways to solve problems of public opinion, attitudes, and emotions of an entity, in which the entity may represent individuals, events or topics [4]. Consumer reviews affect whether or not an online sales company is good. Internet becomes an important part of the life.

Now, not only from family and friends but also from foreigners located all over the world who may have used The 6th International Conference on Cyber and IT Service Management (CITSM 2018) Inna Parapat Hotel – Medan, August 7-9, 2018 products, shop online on sites, visit places or destinations and see movies can pour their opinions online. B. Pre-Processing If the data has been structu red data and a numeric value, the data can be presented as a source of data that can be processed further.

The processes performed in pre-processing are: 1) Tokenization Tokenize used to separate wo rds or letters of punctuation marks and symbols. 2) Stopwords Removal Remove words that are considered unnecessary in the processing of data, for example if, the, of, or, etc. 3) Steaming The process of converting a wo rd into a basic word.

This method of converting word forms into basic words adjusts the language structure used in the steaming process. C. Naïve Bayes Naïve Bayes is an algorithm that is often used in text categorization. The basic idea is to combine the probability of words and categories to estimate the probability of the category of a document [5].

Naïve Bayes is an approach that leads to Bayes theorem, combining previous knowledge with new knowledge. So this is one of the simplified classification algorithms but has high accuracy. [6] Bayesian Classification is based on the Bayes theorem that has similar classification capabilities to the decision tree and neural network.

Bayesian Classification is proven to have high accuracy and speed when applied to databases with large data. D. Genetic Algorithm The Genetic Algorithm is one of the optimization algorithms, which was created to mimic some of the processes observed in natural evolution. The Genetic Algorithm is also a strong stochastic algorithm based on the principles of natural and natural genetic selection that is quite successfully applied in machine learning and optimization problems. [7] The success of the Genetic Algorithm is highly dependent on two factors, population diversity, and selective pressure.

There is a strong influence between these two factors. An increase in selection pressure can increase the number of chromosomes directly copied from the previous generation. In contrast, an increase in population diversity can decrease the proportion of inherited chromosomes and lose the opportunity for them to evolve according to offspring.

[8] E. Validation and Evaluation Confusion matrix provides the decisions obtained in the transfers and testing, the Conf usion matrix provides an assessment of the classification performance by object correctly or falsely [9]. Confusion matrix contains the actual information and predictions on the classification system.

Equation model of Confusion Matrix: 1) Accuracy value is the proportion of the number of correct predictions. ?(????????????? ????????? ???????????????? 1 2) Sensitivity was used to compare the proportion of TP to positive tuples. ???????????? ?? ????????? 2 3) Specificity was used to compare the proportion of TN to negative tuples.

????????????? ?? ???????? 3 4) PPV (positive predictive value) is the proportion of cases with a positive diagnosis. ???? ?? ????????? 4 5) NPV (negative predictive valu e) is the proportion of cases with a negative diagnosis. ???? ?? ???????? 5 The ROC curve shows accuracy and compares the classification visually. The ROC curve expresses the confusion matrix.

ROC is a two-dimensional graph with false positive as horizontal and true positive lines as vertical lines. Guidelines for classifying the accuracy using the AUC: [9]. 0.90 - 1.00 = Excellent Classification; 0.80 - 0.90 = Good Classification; 0.70 - 0.80 = Fair Classification; 0.60 - 0.70 = Poor Classification; 0.50 - 0.60 = Failure. III. METHOD The following are the steps of the research method: 1.

Data Collection This study uses data taken from websites that provide online reviews. Many of the reviews available from the site include customer reviews of online fashion companies. The data us ed in this study as many as 200 data consisting of 100 positive reviews and 100 negative reviews. 2.

Initial Data Processing The next stage is the initial data processing. Dataset used as many as 200 data, 100 positive reviews and 100 negative reviews are used as data training. This dataset in the preprocessing stage must go through three processes.

The 6th International Conference on Cyber and IT Service Management (CITSM 2018) Inna Parapat Hotel – Medan, August 7-9, 2018 The processes are tokenization , stop word removal, and stemming. 3. Proposed Model The model that researchers propose is to use the feature selection method of the genetic algorithm. Genetic algorithms are used so that the accuracy of using naïve Bayes may increase. The picture below illustrates the model proposed in this study. Fig. 1.

The Model Proposed 4. Experiments and Testing Models a. Setting up datasets for

experiments b. Input reviews that have not been previously classified c. if the text has been inputted all then do pre-process d. Design the Naïve Bayes algorithm architecture and do the training and testing and record the accuracy and AUC. e.

Perform testing with 10 fold cross-validation and look for the value of feature selection. f. Designing the naïve Bayes algorithm architecture, the feature selection algorithm is the genetic algorithm and performs. g. Training and testing and record the accuracy and AUC. h. Perform parameter optimization on genetic algorithm to find out the highest accuracy and AUC. 5.

Evaluation and Validation of Results The final stage will evaluate the previously tested data by evaluating the comparison results of the whole experiment between using naïve Bayes algorithm with naïve Bayes algorithm and genetic algorithm. The higher value of accuracy, indicating the proposed model is the best. IV. RESEARCH RESULTS A. Model by Classification Method Using Naïve Bayes In the study show 10 data from a total of 200 data.

5 words related to the sentiment and most often appears that recommend, disappoint, horrible, good and great. Validation used 10-fold cross validation for model testing, where each section will be randomly generated. Principle 10- fold cross validation is 1: 9, 1 part becomes data testing and other data into training data, so that 10 part is the chance to be data testing.

TABLE I Accuracy results using Naïve Bayes algorithm Accuracy: 68.50% +/- 4.50% (mikro: 68.50%) true negatif true positif class precision pred.negatif 78 41 65.55% pred.positif 22 59 72.84% class recall 78.00% 59.00% Fig. 2. Graph Area Under Curve (AUC) using Naïve Bayes Algorithm B. Model with Classification Method Using Naïve Bayes and Selection of Genetic Algorithm Features The optimal parameters in Genetic Algorithm were obtained with population size 50, the number of generation 30, p crossover 0.8 and p mutation 0.08 [10]. To get the highest accuracy results required parameters that require adjustment.

Here are the parameters that are adjusted. The 6th International Conference on Cyber and IT Service Management (CITSM 2018) Inna Parapat Hotel – Medan, August 7-9, 2018 TABLE 2 Experiment Plan Maximum Number of Generation Population Size P Crossover P Mutation Accuracy AUC 30-100 5-50 0.5-1.0 0.5-1.0

? ? TABLE 3 Experimental Results Maximum Number of Generation Population size P Crossover P Mutation Accuracy AUC 40 45 0.5 0.5 87.50% 0.819 40 45 0.5 0.6 87.50% 0.819 40 45 0.5 0.7 87.50% 0.819 40 45 0.5 0.8 87.50% 0.819 40 45 0.5 0.9 87.50% 0.819

40 45 0.5 1.0 87.50% 0.819 The final adjustment of p mutation parameter starting from 0.5-1.0 and there was no change in the value of accuracy and AUC.

Therefore, the adjustment of p mutation is taken from the default value of 0.5. From the experimental process that has been done, it can be concluded that to obtain the highest accuracy and AUC value, the optimal parameters for the maximum number of generation values are 40, population size 45, p crossover 0.5 and p mutation 0.5.

TABLE 4 Confusion Matrix Naïve Bayes Algorithm after added selection features of Genetic Algorithm Accuracy: 87.50% +/- 7.50% (mikro: 87.50%) true negatif true positif class precision pred.negatif 81 6 93.10% pred.positif 19 94 83.19% class recall 81,00% 94,00% Accuracy value of confusion matrix naïve bayes algorithm after the addition of genetic algorithm are: ?(????????????? ??????????? ????????????????????? 6 ?(?????????????? ???????? ????`?????????????????? ? ??? ??? ?????G?? ? 7?? ? ? ?????? 7 Fig. 3.

Graph Area Under Curve (AUC) Naïve Bayes Algorithm after addition selection feature of Genetic Algorithm TABLE 5 Naïve Bayes Algorithm model before and after using feature selection Naïve Bayes Algorithm Naïve Bayes Algorithm + Genetic Algorithm Successful classification of positive reviews 59 94 Successful prediction of negative reviews 78 81 Accuracy 68.50% 87.50% AUC 0.515 0.819 Based on results of the above evaluation is known that naïve Bayes algorithm after addition of genetic algorithm feature selection can increase the accuracy value for online fashion company review. Figure 4 is a graph showing the accuracy of the naïve Bayes algorithm and the naïve Bayes algorithm and the genetic algorithm.

Figure 5 is a graph showing a value of AUC. Fig. 4. Naïve Bayes Accuracy Charts before and after using feature selection 0 10 20 30 40 50 60 70 80 90 100 NB NB+GA Accuracy Algorithm Naïve Bayes Algorithm Accuracy Charts Before and After Using Feature Selection The 6th International Conference on Cyber and IT Service Management (CITSM 2018) Inna Parapat Hotel – Medan, August 7-9, 2018 Fig. 5.

AUC Graph Value before and after using feature selection CONCLUSION Based on the data processing that has been done, merging the naïve Bayes algorithm with selection features of genetic algorithm can improve the accuracy. Online fashion company review can be classified well into positive and negative reviews. Accuracy naïve Bayes algorithm before using feature selection of 68.50% and AUC 0.515. While accuracy after using genetic algorithm feature selection of 87.50% and AUC 0.819. Accuracy increased in the amount of 19.00% and the accuracy of testing using AUC included in Good Classification category. REFERENCES [1] H.

Tang, S. Tan, and X. Cheng, “Expert Systems with Applications A survey on sentiment detection of reviews,” Expert Syst. Appl. , vol. 36, no. 7, pp. 10760–10773, 2009. [2] H. Kang, S. J. Yoo, and D. Han, “Senti -lexicon and improved Na??ve Bayes algorithms for sentiment analysis of restaurant reviews,” Expert Syst. Appl., vol. 39, no. 5, pp. 6000–6010, 2012.

[3] J. Chen, H. Huang, S. Tian, and Y. Qu, “Feature select ion for text classification with Na??ve Bayes,” Expert Syst. Appl., vol. 36, no. 3 PART 1, pp. 5432–5435, 2009. [4] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Eng. J. , vol. 5, no. 4, pp. 1093–1113, 2014. [5] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, “Sentiment classification of Internet restaurant reviews written in Cantonese,” Expert Syst. Appl., vol.

38, no. 6, pp. 7674–7682, 2011. [6] S. F. Rodiyansyah and E. Winarko, “Klasifikasi Posting Twitt er Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification,” FMIPA UGM, vol. 6, no. 1, pp. 91 –100, 2012. [7] P. Guo, X. Wang, and Y. Han, “The Enhanced Genetic Algorithms for the Optimization Design,” no. Bmei, pp. 2990–2994, 2010. [8] W. Song, C. H. Li, and S. C.

Park, “Expert Systems with Applications Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures,” Expert Syst. Appl., vol. 36, no. 5, pp. 9095–9104, 2009. [9] F. Gorunescu, Data mining: concepts and techniques. 2011. [10] S. Günal, “Hybrid feature selection for text classification,” Turkish J. Electr. Eng. Comput. Sci. , vol. 20, no. SUPPL.2, pp. 1296 –1311, 2012.

0,000 0,100 0,200 0,300 0,400 0,500 0,600 0,700 0,800 0,900 NB NB+GA AUC Algorithm AUC Graph Value Before and After Using Feature Selection

INTERNET SOURCES:

<1% - http://frieyadie.web.id/abouts/
<1% - http://www.ijns.org/journal/index.php/ijns/issue/current
6% - https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8674286
3% - https://ieeexplore.ieee.org/document/8674286/
<1% - https://www.researchgate.net/publication/324701535_Cross-Validation
<1% - https://quizlet.com/ca/151294971/chapter-1-flash-cards/
<1% - https://pdfs.semanticscholar.org/6e51/8946c59c8c5d005054af319783b3eba128a9.pdf
<1% - https://www.researchgate.net/profile/Naveen_Amblee/publication/215655590_Harnessin

g_the_Influence_of_Social_Proof_in_Online_Shopping_The_Effect_of_Electronic_Word-of-Mouth_on_Sales_of_Digital_Microproducts/links/5451fe9d0cf24884d88728b0.pdf
<1% -
https://s3-us-west-2.amazonaws.com/visionresources/current_affairs/082a9-march-2020.pdf
<1% - https://www.sciencedirect.com/science/article/pii/S0167404820300213
<1% - http://www.www2008.org/papers/pdf/p715-dzhouA.pdf
1% - http://www.jatit.org/volumes/Vol96No13/12Vol96No13.pdf
<1% -
https://www.researchgate.net/profile/Amit_Ganatra2/publication/265068741_A_Comparative_Study_of_Training_Algorithms_for_Supervised_Machine_Learning/links/5780b65f08ae9485a43ba431.pdf
1% - https://link.springer.com/chapter/10.1007/978-3-642-25507-6_1
<1% - https://research.ijcaonline.org/ncipet/number14/ncipet1111.pdf
<1% - https://www.sciencedirect.com/science/article/pii/S0022283696908979
<1% -
https://www.researchgate.net/publication/271770810_Hybrid_Ensemble_Classification_of_Tree_Genera_Using_Airborne_LiDAR_Data
<1% - https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html
<1% -
http://applications.emro.who.int/imemrf/Emergency_journal/Emergency_journal_2015_3_3_87_88.pdf
1% - http://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/download/697/572
1% -
https://www.researchgate.net/publication/332077528_IT_Operation_Services_Impacts_of_Maturity_Levels_of_IT_Governance_on_Online_Stores_in_West_Kalimantan
<1% -
http://www.ijstr.org/final-print/feb2020/Performance-Analysis-Of-Ensemble-Feature-Selection-Method-Under-Svm-And-Bmnb-Classifiers-For-Sentiment-Analysis.pdf
<1% -
https://s2editor-guides.readthedocs.io/New_Tutorials/04_Data_Editor/075_Buttons/
<1% - http://www.cmar.csiro.au/e-print/open/cmar_rp020.pdf
<1% - http://madlib.apache.org/docs/latest/group__grp__validation.html
<1% - https://jneuroengrehab.biomedcentral.com/articles/10.1186/s12984-017-0309-z
<1% -
https://www.researchgate.net/publication/259697143_Impact_of_body-mass_factors_on_setup_displacement_in_patients_with_head_and_neck_cancer_treated_with_radiotherapy_using_daily_on-line_image_guidance
<1% -
https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=

FAQ
<1% - https://www.hindawi.com/journals/wcmc/2019/9507938/
<1% -
https://www.researchgate.net/publication/220587952_Adapting_Crossover_and_Mutation_Rates_in_Genetic_Algorithms
<1% - https://doi.acm.org/10.1145/1276958.1277395
<1% -
https://www.researchgate.net/publication/268451815_Opinion_mining_of_text_documents_written_in_Macedonian_language
<1% - https://dl.acm.org/doi/10.1016/j.knosys.2015.06.015
<1% -
https://www.researchgate.net/publication/321005822_Condition_Monitoring_of_Roller_Bearing_by_K-Star_Classifier_and_K-Nearest_Neighborhood_Classifier_Using_Sound_Signal
<1% - https://repository.bsi.ac.id/index.php/repo/viewitem/523
<1% - https://link.springer.com/article/10.1007/s10462-019-09794-5
<1% -
https://www.researchgate.net/publication/284733751_Hybrid_algorithm_based_on_Genetic_Algorithm_and_Tabu_Search_for_Reconfiguration_Problem_in_Smart_Grid_Networks_Using_R
1% - https://dl.acm.org/doi/10.1016/j.eswa.2014.10.023