PROC. INTERNAT. CONF. SCI. ENGIN.
Volume 4, February 2021
Page 216-219

ISSN 1504607797
E-ISSN 1505707533

Sentiment Prediction Accuracy of Amazon Fine Food Review using TF-IDF and LightGBM models

Tanzilal Mustaqim[1], Aprilia Dewi Ardiyanti[2]
[1]Computer Science Department, Faculty of Mathematics and Natural Sciences, State University of Semarang.
Jl. Kolonel HR Hadijanto, Sekaran, Gn. Pati, Semarang, Central Java 50229, Indonesia. Tel. (024)8508032, Fax. (024)8508032.
[2] Departement of Physics, Faculty of Mathematics and Natural Sciences, State University of Malang.
Jl. Semarang No.5, Sumbersari, Lowokwaru, Kota Malang, East Java 65145, Indonesia. Tel. (0341)587958.
[1]Email: derekln14@gmail.com

Abstract. Changes in the pattern of society in meeting their needs develop as the times progress from conventional to digital. This makes service providers need to change business work patterns towards digitizing buying and selling transactions. Service providers serve consumer needs digitally and maintain optimal service patterns. One of the efforts to maintain optimal service is through community response to services, both positive and negative. The community response can be analyzed using sentiment analysis. This study focuses on the analysis of the accuracy of sentiment predictions on the Amazon fine food review dataset, which was taken as many as 20,000 data samples. The analysis was carried out in various stages, namely dataset collection, data preprocessing, TF-IDF, and LightGBM. The test results used TF-IDF and LightGBM with TF-IDF parameter settings of 1 to 2 grams and LightGBM parameter settings with a max_depth of 50. Num_leaves used were 40 and the learning rate was 0.1 on the Amazon Review dataset which took 20,000 samples. The analysis carried out resulted in a predictive level of sentiment accuracy above 90%, reaching 93.2%.

Keywords: Amazon Review, LightGBM, TF-IDF, Sentiment Analysis.

Running title: Sentiment Prediction Accuracy of Amazon Review.

INTRODUCTION

Changes in the habits of conventional societies towards digital societies have been felt and spread throughout the world (Demiran & Spohrer 2014). The pattern of fulfilling needs such as shopping by visiting stores offline has decreased drastically in this era. This change is triggered by the shift of service providers from conventional to digital on the grounds of adapting to changes in people's habitual climate. Staple goods service providers prefer to switch to a digital platform because it can expand the reach of the customer market which of course increases profits.

People prefer digital platforms because they can save resources to meet needs such as time and energy. The easy ordering process also increases public interest in digital services. This process of change makes the company strive to maintain excellent service to make customers sustainable customers and reducing the disappointment they experience (Agarwal & Pradeep 2013).

The process carried out by the company is to analyze historical data on the results of service activities to customers. One process that can be done through sentiment analysis. Sentiment analysis is the process of analyzing the polarity of public sentiment on comments and messages sent to companies for services or products provided. The results of sentiment analysis are very important because companies can find out service performance and things that need to be improved to increase profits or things that need to be reduced to minimize losses (Saif *et al*. 2016).One company that provides digital needs is Amazon. Amazon is a marketplace company based in the United States that provides a variety of needs ranging from food products to household products and other necessities (Al Amrani *et al*. 2018). The dataset in this study refers to historical data from the results of Amazon services to consumers from 1999 to 2012 by taking a sample of 20,000 research data.

Sentiment analysis on the Amazon fine food review dataset has previously been researched by Veera et al in 2020 using various machine learning algorithms such as 68.66% Decision Tree, 60.7% Neural Network, 77.61% Support Vector Machine, Random Forest as much as 70.65%, K-Nearest Neighbors 81.09% and Gradient Boosting Classifier 77.61% (Veera *et al*. 2020). Veera et al's research produced K-Nearest Neighbors with the highest accuracy. Sentiment analysis research with the same dataset was also conducted by Anees *et al* in 2019 using various machine learning algorithms such as SVM, Logistic regression, and Naïve Bayes. The highest result from Anees' research was the use of machine learning logistic regression at 78.11% (Anees *et al*. 2019s).

Two previous studies have shown that sentiment predictions have not been generated with accuracy above 90%. The gap analysis conducted by the authors in this study is to increase the accuracy of sentiment predictions to reach 90% or more. The method used to improve accuracy is to use the TF-IDF weighting feature and the LightGBM machine learning algorithm. TF-IDF stands for Term Frequency - Inverse Document Frequency which allows us to count every word in a document. The TF-IDF model takes into

account the frequency of appearance of each word in a document and calculates the weight as a preprocessing stage before being processed by machine learning (Al-Khalifa *et al*. 2020). LightGBM is tree-based machine learning included in the gradient enhancement framework. The advantages of using LightGBM are that it has a higher processing speed and higher efficiency and results in higher accuracy than other machine learning (Bi *et al*. 2020). The combination of the TF-IDF and LightGBM models is combined with parameter settings on LightGBM to improve sentiment prediction accuracy on Amazon's fine food review dataset.

MATERIALS AND METHODS

Study area

The process of research using data analysis tools provided publicly by Google, namely Google Colab. The computing specs consist of 12.72 GB RAM and 107.77 GB storage. The research dataset, namely the Amazon fine food review, was obtained publicly by Kaggle.

Procedures

Research

Procedure

The research procedure is shown in more detail in flowchart. Figure 1 shows the research sequence from dataset collection, data pre-processing, TF-IDF and LightGBM.

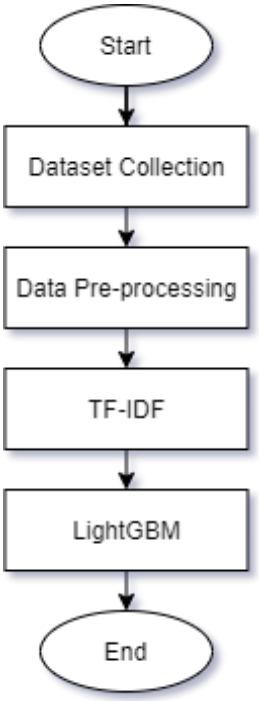

Figure 1. Flowchart Research.

Dataset Collection

The research dataset used is Amazon fine food review, amounting to 568,454 reviews and the period taken is from October 1999 to October 2012. The sample used in this study amounted to 20,000 with the aim of resource efficiency and research model testing.

Preprocessing Data

Data preprocessing is the stage of normalizing data from several things that have the potential to become noise and make the final research inaccurate. Some of the things that are done in this process are:

a. Case folding.

Case folding is to make the types of all alphabetic characters of the data uniform. In this study, all characters were made into a lowercase. Another purpose of case folding is to make the process of memory-efficient because the computer cannot distinguish between the letters "review" and "review" even though they have the same meaning in context.

b. Remove the Html tag.

The Amazon review dataset has several Html tags that need to be removed. The Html tag contained in the dataset is the result of thorough scraping from reviews available on the Amazon website.

c. Remove duplication of data

Data duplication is often a scourge in the data analysis process. This is because duplication makes the assessment inaccurate and makes the commutation process more complicated.

d. Deletes characters other than alphabetic letters.

Character words that contain numbers such as "review67" make the results of data analysis become noise. One of the words that contain numbers is caused by a typo at the time of writing and this needs to be normalized to make the quality of the data analysis process optimal.

e. Removes stopwords.

The communication process carried out by humans often uses repeated words. Examples are "and", "because", "for" and many more. The words that are often used are called stopwords. Stopwords need to be cleaned in the data pre-processing stage to make data slimmer and improve the quality of data analysis results.

f. Stemming.

Stemming is the process of cutting affixed words into their original form. The computer can not recognize the word differences in a straightforward manner because the computer cannot see the context of the meaning of the words being analyzed. It is necessary to change words with affixes such as "eating" to "eat" by removing the affix "ing".

TF-IDF (Term Frequency - Inverse Document Frequency)

TF-IDF is a method of weighting words in data

documents whose work processes are related to the number of words in the document being analyzed. The more words that appear, the weighting will change according to the number of words and the number of documents. TF-IDF is used as a method for converting feature data into vectors before being analyzed by machine learning (Al-Khalifa *et al*. 2020).

LightGBM

LightGBM is a machine learning algorithm whose work process is based on a tree algorithm and is included in the gradient boosting trees framework. LightGBM is a type of machine learning algorithm developed by the Microsoft company. This algorithm has advantages in the area of increasing accuracy and saving more memory usage (Bi *et al*. 2020).

RESULTS AND DISCUSSION

The testing process of this research uses TF-IDF and the LightGBM machine learning algorithm to test the accuracy of sentiment predictions generated from the data provided by the Amazon review dataset. The TF-IDF adjustment process uses a number of grams of 1 to 2 grams. The usage of the number of grams is determined based on the dominant number of word combinations used from 2 continuous words. Even though there is a continuous use of 3 or more words, only 2 grams are used to make the computation process more efficient.

The parameter setting process used by the LightGBM machine learning algorithm is max_depth of 50, num_leaves of 40 and learning_rate of 0.1. The process of determining parameters is obtained from gridsearch results in the python scikit-learn library with the rest of the other parameters set by default.

In this research, analysis of the results of the accuracy of sentiment prediction is carried out, comparison with previous studies and showing the feature importance of the analysis results. In accordance with the research stages described in the method chapter, the dataset that has been obtained is cleaned at the data pre-processing stage and analyzed at the TF-IDF and LightGBM stages. The results of data analysis are shown in Figure 2 and Figure 3.

Figure 2. Feature Importance.

The results of the feature importance analysis are shown in Figure 2. Feature importance shows the frequency of occurrence of words contained in the dataset. The larger the word size that appears, indicates the number of dominant words compared to the smaller word size. The order of feature importance in this research is love, good, great, best, and so on which are dominated by positive words.

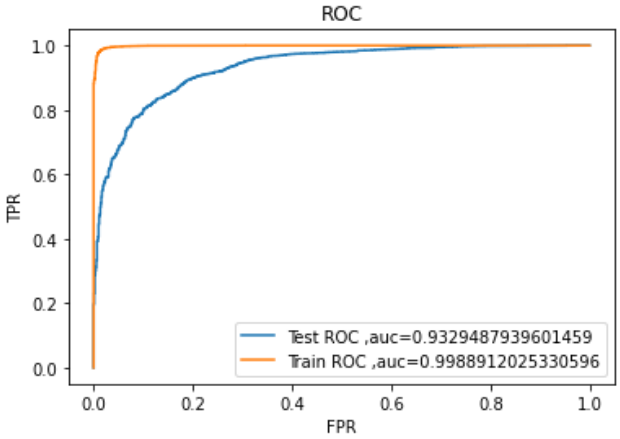

Figure 3. Sentiment Prediction Accuracy Rate

Figure 3 shows that the use of TF-IDF and LightGBM produces an accuracy rate of 93.2%. The results of sentiment prediction accuracy exceed those of the results conducted by Harika *et al* (2020) using the K-Nearest Neighbors algorithm with an accuracy of 81.09% and research conducted by Anees et al using logistic regression algorithm with results of 78.11%. Amazon review dataset analysis using TF-IDF and the LightGBM machine learning algorithm is proven to outperform previous studies and are following the research objectives.

CONCLUSIONS

The test results used TF-IDF and LightGBM with TF-IDF parameter settings of 1 to 2 grams and LightGBM parameter settings with a max_depth of 50. Num_leaves used were 40 and the learning_rate was 0.1 on the Amazon Review dataset which took 20,000 samples. The analysis carried out resulted in a high level of sentiment accuracy prediction above 90%,

reaching 93.2%.

REFERENCES

Agarwal, Raina & Pradeep, Y. 2013. Bridging the gap between traditional and online shopping methods for Indian customers trough digital interactive experience. Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics. India: 22-25 August 2013.

Al-khalifa, Shaima, Aljarah, Ibrahim, & Abushariah, Mohammad A M. 2020. Hate Speech Classification in Arabic Tweets. Journal of Theoretical and Applied Information Technology 98: 1816–1831.

Al Amrani, Yassine, Lazaar, Mohamed, & El Kadirp, Kamal Eddine. 2018. Sentiment Analysis using supervised classification algorithms. Procedia Computer Science 127: 511–520.

Anees, Aiman Abdullah, Prakash Gupta, Harsh, Dalvi, Aditya Prashant, Gopinath, Suhas, & Mohan, Biju R. 2019. Performance Analysis of Multiple Classifiers using different Term Weighting Schemes for Sentiment Analysis. International Conference on Intelligent Computing and Control Systems. India: 15-17 May 2019.

Bi, Ye, Wang, Shuo, & Fan, Zhongrui. 2020. A Multimodal Late Fusion Model for E-commerce Product Classification. ArXiv Preprint ArXiv 2008: 1-4.

Chandra Pandey, Avinash, Singh Rajpoot, Dharmveer, & Saraswat, Mukesh. 2017. Twitter sentiment analysis using hybrid cuckoo search method. Information Processing and Management 53: 764–779.

Demirkan, Haluk, & Spohrer, Jim. 2014. Developing a framework to improve virtual shopping in digital malls with intelligent self-service systems. Journal of Retailing and Consumer Services 21: 860–868.

Saif, Hassan, He, Yulan, Fernandez, Miriam, & Alani, Harith. 2016. Contextual semantics for sentiment analysis of Twitter. Information Processing and Management 52: 5–19.

Veera, K Mani, Ratna, Venkata, Anusha, M Sai, Tejaswini, S, & Swamy, B Tirupati. 2020. IX (V): 112–120.