E-ISSN: 2476-9606
Abstract Proceedings International Scholars Conference
Volume 7 Issue 1, October 2019, pp. 1719-1728

https://doi.org/10.35974/isc.v7i1.1003

Sentiment Analysis of Customer Reviews in Zomato Bangalore Restaurants Using Random Forest Classifier

Bern Jonathan[1], Jay Idoan Sihotang[2], Stanley Martin[3]
[1] Departement of Technology, Female Daily Network
[2,3]Department of Information Technology, Universitas Advent Indonesia
bern@femaledaily.com

ABSTRACT

Natural Language Processing is one part of Artificial Intelligence and Machine Learning to make an understanding of the interactions between computers and human (natural) languages. Sentiment analysis is one part of Natural Language Processing, that often used to analyze words based on the patterns of people in writing to find positive, negative, or neutral sentiments. Sentiment analysis is useful for knowing how users like something or not. Zomato is an application for rating restaurants. The rating has a review of the restaurant which can be used for sentiment analysis. Based on this, writers want to discuss the sentiment of the review to be predicted. The method used for preprocessing the review is to make all words lowercase, tokenization, remove numbers and punctuation, stop words, and lemmatization. Then after that, we create word to vector with the term frequency-inverse document frequency (TF-IDF). The data that we process are 150,000 reviews. After that make positive with reviews that have a rating of 3 and above, negative with reviews that have a rating of 3 and below, and neutral who have a rating of 3. The author uses Split Test, 80% Data Training and 20% Data Testing. The metrics used to determine random forest classifiers are precision, recall, and accuracy. The accuracy of this research is 92%. The precision of positive, negative, and neutral sentiment are 92%, 93%, 96%. The recall of positive, negative, and neutral sentiment are 99%, 89%, 73%. Average precision and recall are 93% and 87%. The 10 words that affect the results are: “bad”, “good”, “average”, “best”, “place”, “love”, “order”, “food”, “try”, and “nice”.

Keywords: Sentiment Analysis, Random Forest, Precision-Recall, Feature Selection

INTRODUCTION

Sharing on the internet is something we usually do. Giving a review is also a useful activity so that other people on the internet can find out something else and see opinions about things. The usual things reviewed by someone in the form of experiences, places, objects, and others. Give a review we usually use text to explain something that we experience with an item, place, or event that we normally experience.

Customer satisfaction is an opinion between expectation and reality obtained by consumers (Ilieska, 2013) Giving a review is also a useful activity so that other customer on the internet

can find out something else and see opinions about things and its satisfaction. Commonly, most people express their opinion through social media like Facebook and Twitter or review platform like Zomato, Google My Business, Yelp, etc. Customer reviews on online media like Zomato become important as it might increase the popularity of something.

Zomato is a site where someone can give a review of a restaurant, how the restaurant is and someone's opinion about the restaurant. Restaurant customer satisfaction can be analyzed by their review on Zomato. Sometimes, restaurants see the reviews in Zomato, but they didn't get if the reviews are positive or negative to their restaurants.

Review on Zomato is still in the form of text and can be classified with positive, negative, or neutral with their ratings. Zomato doesn't have an analysis of how users interact with the reviews and what words will indicate they like or not it. We need to extract the words in review and analysis it so we can know how users interact in Zomato and get customers satisfaction by their review.

In this paper, we purpose a method to analyze user's sentiment of Zomato Restaurants and focusing review in Bangalore for study case. We are using Random Forest Classifier to classify the sentiments of users based on their review. We also find words that affects the classifier model.

Theorical Based And Related Work

1. Sentiment Analysis

Sentiment analysis, is that field of study people's opinions, sentiments, evaluations, judgments, attitudes, and emotions towards entities such as products, services, tourism, movies, organizations, political issues, individuals, problems, events, topics, etc (Liu, 2012). Sometimes, organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies. With the explosive growth of social media on the Web, individuals and organizations are increasingly using the content in these media for decision making using its text (Liu, 2012:8).

Sentiment Analysis is part of research areas such as natural language processing, data mining and text mining (Farhadloo & Rolland, 2016), that often used to analyze words based on the patterns of people in writing to find positive, negative, or neutral sentiments. The goal to Sentiment Analysis is to know how people feel about something from their text.

2. Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges as. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost, but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression (Breiman, 2001).Detailed random forest algorithm flow is described.

Step1: Using bootstrap sampling to resample, randomly produce n training sets;

Step2: Use each training set, and generate the corresponding decision tree, in each leaf node before selecting attributes, from the M attribute randomly extracted m attributes as the current node of the set of attributes, and the best split in the M attribute to the node split;

Step3: Each tree is fully grown without pruning;

Step4: For test set samples, using each decision tree to test, get the corresponding category;

Step5: By voting method, the category with the most output categories in the decision tree as the test set sample belongs to (Zhou, Guo, Fu, & Liang, 2019). The concrete steps are shown in Figure 1.

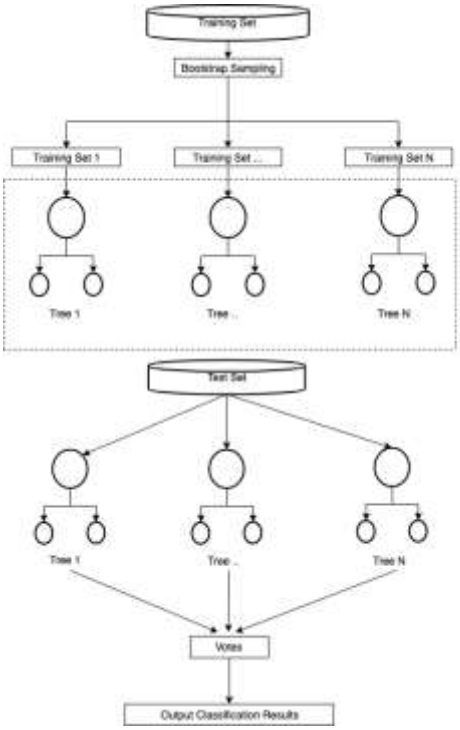

Figure 1. **Random Forest**

3. Related Work

Research that has conducted by Valonia Inge S. et al. shows us that they evaluate lecturer performance with sentiment analysis in Universitas Kristen Duta Wacana. They are using Support Vector Machine (SVM) to evaluate the performance. In that research, they are only using 307 documents that split by 103 positive, 163 negatives, and 41 neutral. And with their research, they got with term frequency of words and got 67.83% accuracy (Santoso, Virginia, & Lukito, 2017).

Thai sentiment analysis for customer's review conducted by Paitoon Porntrakoon, using Thailand's language. That research analysis about how sentiment analysis of customers in Thailand review products, price, and shipping using Sentiment Compensation Technique (SenseComp). They are using 2500 users review in Lazada website. The results of this research show that their proposed method outperforms sentiment to dimension (S2D) and dimension to sentiment (D2S) methods with the overall accuracy 93.60% (Porntrakoon & Moemeng, 2019).

Sentiment analysis of restaurant customer reviews on TripAdvisor using Naïve Bayes research conducted by Rachmawan also show us how sentiment analysis works in restaurant's reviews. Their research collects 337 data and 269 for training data and 68 for test data. The result from this research shows that these two methods get the customer response accurately and Naïve

Bayes method is more accurate than TextBlob sentiment analysis with a different accuracy of 2.9% (Laksono et al., 2019).

METHODS

This paper takes restaurant review data on Kaggle, specific restaurants in Bangalore, and will be analyzed using the Random Forest in Python Scikit-Learn and analyze it with accuracy and precision-recall. There are several steps of the research method.

1. Data Collection

Data that we got from Kaggle is a platform for predictive modeling and analytics competitions in which companies and researchers post data and statisticians and data miners compete to produce the best models for predicting and describing the data. Data that we collect specific is about reviews on Zomato Bangalore and collect 150.000 reviews to be analyzed.

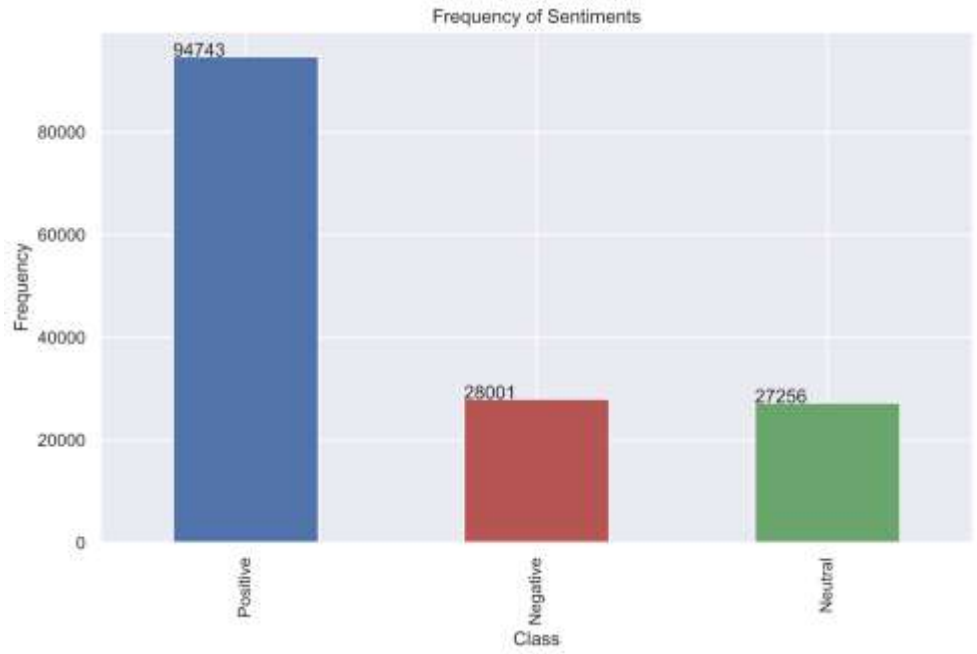

Figure 2. **Frequency of Sentiments**

From 150.000 reviews we split them by its ratings. Make positive with reviews that have a rating of 3 and above, negative with reviews that have a rating of 3 and below, and neutral who has a rating of 3. The results of this split, we got 94.743 positive reviews, 28.001 negative reviews, and 27.256 neutral reviews. There are imbalanced datasets but, from research by

Yusran, Juliana, and Bern imbalanced data set didn't affect significant accuracy (Samuel, Hutapea, & Jonathan, 2019). So in this research, we are not handling the imbalanced data set.

2. Workflow Process

The study was conducted and processed in Python 3.6 and with the Scikit-Learn library using the Random Forest method to implement the Sentiment Analysis program. The following figure is a block diagram of the stages of research.

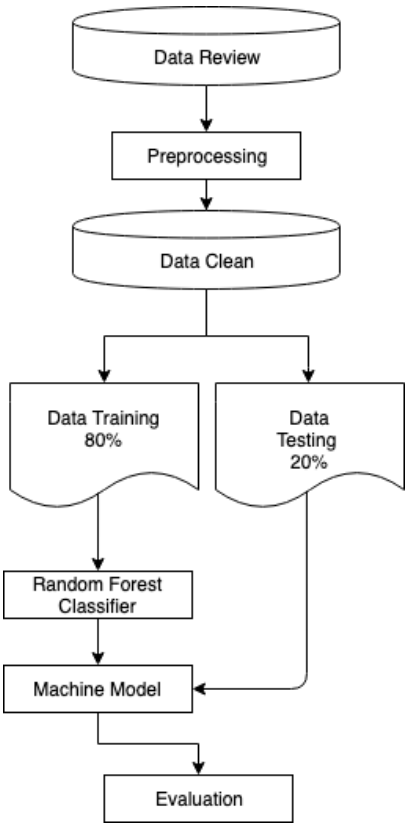

Figure 3. **Workflow Process**

From data review we are pre-processing text then if the data clean, we split it to 80% training data and 20% test data. Then the 80% data training data we train using Random Forest. After the machine model finish trained, we are testing it to data testing and evaluate the accuracy and precision-recall to see how the metrics of our machine model.

3. Text Preprocessing

Text preprocessing is the first stage of text mining. The purpose of text preprocessing is to prepare unstructured text documents into structured data that is ready to be used for processes then by eliminating noise, homogenize word forms and reduce word volume (Putraranti & Winarko, 2014). Stages of preprocessing text used in this study are lowercase, tokenization,

remove punctuation, stopwords removal, pos tags, lemmatized, and using Tf-Idf Vectorizer to vectorize text to number.

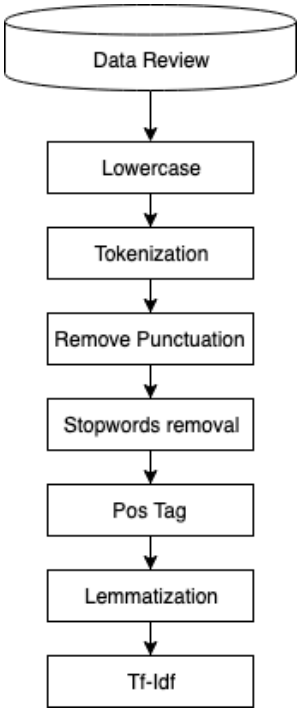

Figure 4. **Text Preprocessing**

RESULTS

From split data 80% of training data, we made a machine model to predict. The machine model predicts the other 20 % of data as data testing to see how our model work. We use Random Forest Classifier with tree estimators to 100 trees for make the machine model good enough to predict.

1. Accuracy

The results of this machine model are having 92,43667% accuracy. It is explained from 20% data, or 30.000 data to predict the machine correct around 27.740 data.

2. Precision-Recall

Precision is the level of accuracy between the information requested by the user and the answers provided by the system. Recall is the level of success of the system in rediscovering information.

Table 1. **Precision- Recall**

	Precision	Recall
Positive	92%	99%
Negative	93%	89%

Neutral	96%	73%
Average	**93%**	**87%**

3. The precisions of positive, negative, and neutral sentiment are 92%, 93%, 96%. The recalls of positive, negative, and neutral sentiment are 99%, 89%, 73%. Average precision and recall are 93% and 87%.Confusion Matrix

Confusion Matrix is a metric that shows the true and prediction errors of data from the results of an algorithm.

Table 2. **Confusion Matrix**

	Positive	Negative	Neutral
Positive	18758	126	49
Negative	488	4987	114
Neutral	1221	271	3986

From 30.000 data test, we split it the actual data are 18933 positive, 5589 negative, 5478 neutral. From the confusion matrix, we can explain the precision-recall from.

4. Feature Importance

Feature importance is part of random forest to extract any features that affect a machine model. From this machine model we got the 10 words that affect the results are: “bad”, “good”, “average”, “best”, “place”, “love”, “order”, “food”, “try”, and “nice”.

Table 3. **Feature Importance**

Word	**Importance**
word_bad	0.032656
word_good	0.019180
word_average	0.011238
word_best	0.010433
word_place	0.010175
word_love	0.009697
word_order	0.009500
word_food	0.008006
word_try	0.007604
word_nice	0.007451

DISCUSSION

Based on the results of the analysis and testing that has been carried out in this study, there are some conclusions that can be given:

1. The Accuracy of this machine model are 92%
2. The best precision is neutral with 96% accuracy and the good recall is positive with 99% accuracy.
3. The least percentage of recall is neutral. It is indicating the machine model is least to predict neutral than the others.
4. From the feature importance we got the "bad" word is more indicative to predict the sentiment of the people trying to say.

Suggestions given from this research and for the development of further research are as follows:

1. There are imbalanced data in positive, negative, and neutral data. We can use imbalanced dataset algorithms to improve the results.
2. We can use the word data of sentiments to find the sentiments rather than see their ratings.

REFERENCES

Breiman, L. (2001). *Random Forest*.

Farhadloo, M., & Rolland, E. (2016). Fundamentals of sentiment analysis and its applications. In *Studies in Computational Intelligence* (Vol. 639, hal. 1–24). https://doi.org/10.1007/978-3-319-30319-2_1

Ilieska, K. (2013). *Importance of Customer Satisfaction*. Diambil dari www.temjournal.com

Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S. (2019). Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes. *12th International Conference on Information & Communication Technology and System (ICTS) 2019*, 54–59.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Porntrakoon, P., & Moemeng, C. (2019). Thai sentiment analysis for consumer's review in multiple dimensions using sentiment compensation technique (SenSecomp). ECTI-CON 2018 - *15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 25–28. https://doi.org/10.1109/ECTICon.2018.8619892

Putraranti, N. D., & Winarko, E. (2014). Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *8*(1), 91–100. https://doi.org/10.22146/ijccs.3499

Samuel, Y. T., Hutapea, J. J., & Jonathan, B. (2019). Predicting the Timeliness of Student Graduation Using Decision Tree C4 . 5 Algorithm in Universitas Advent Indonesia. *12th International Conference on Information & Communication Technology and System (ICTS) 2019*, 281–285.

Santoso, V. I., Virginia, G., & Lukito, Y. (2017). PENERAPAN SENTIMENT ANALYSIS PADA HASIL EVALUASI DOSEN DENGAN METODE SUPPORT VECTOR MACHINE. *Jurnal Transformatika*, *14*(2), 72. https://doi.org/10.26623/transformatika.v14i2.439

Zhou, Y., Guo, J., Fu, L., & Liang, T. (2019). Research on Aero-Engine Maintenance Level Decision Based on Improved Artificial Fish-Swarm Optimization Random Forest Algorithm. *Proceedings - 2018 International Conference on Sensing, Diagnostics, Prognostics, and Control, SDPC 2018*, 606–610. https://doi.org/10.1109/SDPC.2018.8664905