ID 407

# Comparison of Machine Learning Classifiers for Sentiment Analysis in Hotel Reviews

PLU Kaushalya# and WU Wickramaarachchi

*Department of Computing, Rajarata University of Sri Lanka, Mihintale, Sri Lanka*

#udanikaushalya95@gmail.com

**Abstract**—Sentiment analysis or opinion mining refers to the process of identifying people's sentiments, opinions, attitudes and emotions behind a written text. In recent years, sentiment analysis studies have become an active research area under natural language processing. Understanding the opinion behind the user-generated text can be applied to various applications. When it comes to the hotel sector and travel planning, user reviews and comments are quite useful. Therefore, guest reviews are becoming a prominent factor, which influence people's booking decisions. In addition, knowing about these comments is important for quality control of the hotel management too, because it may be worth checking out some stats over time. The fundamental objective of this research is to compare several machine learning classifiers and find out the best classifiers to develop a sentiment analysis model for the hotel reviews, to tackle customers' sentiment. Under this research, a comparative analysis was established among Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Logistic Regression (LR), Stochastic Gradient Descent Classifier (SGD), Linear Support Vector Classifier (SVC), Random Forest Classifier and Multi-layer Perceptron Classifier (MLP) classifiers. Moreover, two feature extraction techniques called Count Vectorizer and Term Frequency Inverse Document (TF-IDF)) are also compared to find out the best approach to perform the feature extraction. The result from this research shows that the highest results were obtained in Logistic Regression with TF-IDF method (Accuracy 87.39%) and SGD algorithms with TF-IDF (Accuracy 87.71%), while the lowest accuracy was obtained for Bernoulli NB classifier with Count Vectorizer (Accuracy 64.67%). Every time when using Count Vectorizer as the feature extraction method, the accuracies decreased, than when the TF-IDF method was used.

***Keywords:*** ***sentiment analysis, machine learning classifiers, feature extraction techniques***

## I. INTRODUCTION

With the rapid development of the internet and smart mobile devices, E-Commerce and social media have already penetrated everyone's daily life. Owing to that, Users are now able to freely express their opinion about a variety of products and services, by generating reviews, comments, and reports. According to recent estimates, more than 2.5 quintillion (1018) bytes of data are generated daily, while by 2020 1.7 MB of data will be created for each person on earth per second.

The tourism industry leads a crucial role in the economical side of each country around the world. Therefore, travel planning, hotel, and restaurant websites are frequently used by tourists before making their selection. Since these websites contain the bulk of opinions from previous customers, selecting a hotel or a restaurant from thousands of opinions is an overwhelming task. As a result, nowadays people tend to rely on online reviews, to take advantage of the experiences of others. User comments have become a critical and influential source for decision-making. User comments are not just useful for other users but also the business owners who can improve the quality and innovation of their business services and devise innovative marketing strategies. When it comes to the hotel field these users' opinions or reviews are very much useful to tourists to select the best place to stay as well as hotel managers to assess the quality of the hotels.

Sentiment analysis, which is the most popular decedent application of Natural Language Processing (NLP) has been successfully used to analyses user's generated reviews. NLP gives machines the ability to read, understand, and derive meaning from human languages. When it comes to sentiment analysis, it adopts a text analysis technique of NLP to identify the attitude of the text writer. As a result, this technique is mostly used in the business field to identify customer sentiment toward products, brands, or services in online conversations and feedback to identify their opinion and feelings.

The properly constructed sentiment analysis systems can eliminate the need for surveys and change the way traditional market research is conducted. It is because, without consuming so much time, the constructed sentiment analysis system can be used to summarize and identify the sentiment of all feedback comments. Various methods can be used as model sentiment analysis, including Neural Network, Support Vector Machine, Naïve Bayes, Decision Tree, and others. Among these methods, a proper accurate method that is compatible with the data set and the given case scenario should be selected to train the sentiment classification model.

So, this research aimed at developing a Sentiment Analysis model to identify customer satisfaction over the reviews on the hotel industry and conducting a comparative study on different machine learning algorithms and feature extraction techniques. The paper is organized as follows, Related Works in Section II, Methodology and Experimental Design in section III, Results in section IV, Discussion and Conclusion in section V, Future Works in section VI and finally the References.

## II. RELATED WORKS

### *A. Pre-Processing and feature extraction techniques*

As the initial step in Sentiment Analysis, all the reviews in the dataset must be labelled. Labelling the text data, whether it is positive, negative, or neutral also plays a significant role in the sentiment analysis domain. For that most researchers used labelling based on intuition and labelling based on sentiment polarity. Since labelling based on intuition differs from the user perception, labelling based on sentiment polarity has shown effective results rather than human intervention labelling ( Zvarevashe et al., 2018).

Secondly, the labelled dataset must be cleaned, for that most researchers have used the following phases. Removing emoji, numbers and remain only the text with letters, White space removal, tokenization, removal of stop words, stemming, and lemmatization (Ghosh et al.,2017; Symeonidis et al., 2018; Kasper et al., 2011).

As the last last step under the pre-processing stage, feature extraction needs to be performed. For that purpose, in one of research ((Ghosh et al.,2017), different feature's weights for a feature set were discussed, Feature Presence (if the feature appears on the text –1 if not –0), Feature Frequency (Number of times feature occurs in the document), Term Frequency Inverse Document (Evaluate how important a word is to a document). With referring to feature extraction techniques, rather than using only the frequency of the words, the usage of the TF-IDF method resulted to be more effective since it considered the importance of the word as well as its frequency (Shi and Li,2011). Another research has used a technique called CountVectorizer (Tripathy et al.,2015) for converting features into a numerical representation. There, it transforms the review into a token count matrix.TF-IDF transformation, mostly used for machine learning classifiers such as Logistic Regression, Bernoulli Naïve Bayes, and Linear SVC. For Neural Networks, the feature-learning method called word embedding was used.

### B. Machine Learning Classifiers

After conducting the pre-processing and feature extraction of the given text dataset, then the sentiment classification model should be built. In consideration of that, most of the existing studies, which have been conducted under the sentiment analysis domain, used machine learning classifiers. The most used machine classifiers were Naive Bayes, SVM, Logistic Regression, and Decision Tree. Among these classifiers, some of them used just only one from them and some of them have used two or three out of them to have a comparison of each other (Farisi et al.,2019; Yordanova et al.,2017; Bhargav et al.,2019; Nohn et al.,201; Patel et al.,2020).

As a summary, most of the researchers have used Naive Bayes, SVM (Support Vector Machine) and logistic regression, Stochastic Gradient Decedent, and Decision tree machine learning classifiers. Among them, Naive Bayes, SVM classifiers were performed as out fliers by giving more accurate prediction models. When comparing different classifier's results, in sentiment analysis other than the accuracy it also evaluates their result based on precision and recall value measurements too.

## III. METHODOLOGY AND EXPERIMENTAL DESIGN

The proposed system workflow shows in the following Figure 1. The main phases of the system workflow as follows, Collection of Hotel Reviews, Review Preprocessing, Feature Extraction, Model training with Machine Learning algorithms, Selecting the best machine learning algorithm which gives the highest accuracy to train the final sentiment analysis model.

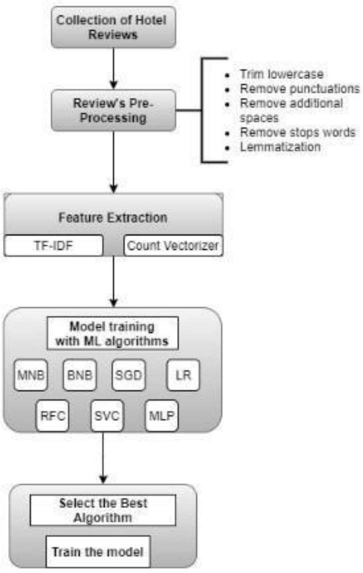


Figure 1. System Workflow

### A. Dataset

The Data Set was collected from the Kaggle website (Sharma,2017). And it included around 40000 hotel reviews under 5 columns. Only the review and the satisfaction column were considered throughout this research. In the initial dataset there were 26521 positive reviews and 12411 negative reviews. Therefore, we had to balance it before applying it to train the prediction model.

*B. Pre-processing and Feature extraction* the following Figure 2 depicts the main steps performed under the pre-processing stage. In sentiment analysis, the main intention of doing pre-processing is to remove all the unnecessary wordings and symbols from the reviews and to remain only the words that are important to identify the sentiment of a review. Therefore, under this research in the basic pre-processing section, two functions were used.

Within the first function initially, the whole review texts were converted into lowercase letters, and then the square bracket, numbers, and punctuation were removed. And that pre-processed reviews were again sent through a second function where it removed additional spaces, newline characters, and only remained the words with English letters. Basic pre-processing functions were performed using Regular Expressions. For that built-in package in python called "re" was used.

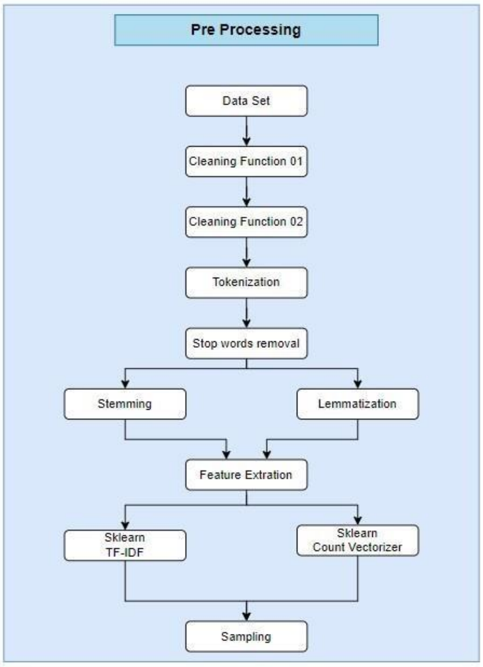


Figure 2. Pre-Processing Workflow

After doing basic pre-processing functions then to easier to filter out unnecessary tokens, the tokenization was performed. Before transforming text into vectors using nltk.word_tokenize the text was converted into tokens. In this case, the reviews were tokenized into words.

Thereafter the stop words were removed from the token set using nltk.corpus.stopwords.words method. The Stop words were removed because they do not add much information in the overall modelling procedure. When it comes to sentiment analysis by removing stops from the given text, it can focus more on words that add some value to the meaning of the text.

Afterward either stemming or lemmatization was performed to achieve the root forms of inflected words. In this study, both stemming, and lemmatization were tried out to see which method is more appropriate for the sentiment analysis. Stemming is different from lemmatization in the approach since it uses to produce root forms of words and the word produced. Lemmatization always gives the dictionary meaning word while converting into root form.

In this research, the WordNet lexicon database was used for lemmatization purposes.

Afterward, then the next step is feature extraction. The main intention of feature extraction was before giving the input to the training model it must be converted into numbers. Since machines do not understand the meaning of the text, it needed to be transformed in a way that machines can interpret it. Therefore, the reviews were transformed in the form of vectors. This helps to increase the accuracy of learned models by extracting features from the input data. This phase of the general framework reduces the dimensionality of data by removing redundant data.

To do that in this research two techniques were tried out. One is Word Frequency Indexing using Sklearn CountVectorizer, Word Frequency Indexing using Sklearn Term Frequency Inverse Document. Both CounterVectorizer and TF-IDF methods were used before training the model with different machine learning algorithms. When it comes to CountVectorizer it considers the count of each word in reviews. In the TF-IDF method, it considers the value count of words with the importance of each word within the whole word Corpus.

### *C. Model Training with Machine Learning Algorithms*

The dataset which was used was not a balanced one therefore before training the model with machine learning algorithms, the dataset was balanced using a method called oversampling.

And with the oversampling method, generates synthetic data that tries to randomly generate a sample of the attributes from observations in the minority class. To do that we have used a common technique called SMOTE (Synthetic Minority Over-sampling Technique). And it generated a result dataset with (Positive Reviews, 21234), (Negative Reviews, 21234), which was a balanced one.

When it comes to the training process 80% of the dataset was allocated for training purposes and 20% of the data set was allocated for testing purposes. After doing the sampling then the models were trained using different machine learning algorithms.

SGD Classifier and Logistic Regression were imported from sklearn.linear_model library. MultinomialNB and BernoulliNB were imported from sklearn.naive_bayes Library. Linear SVC was imported from sklearn.SVM, MLP Classifier was imported from sklearn.neural_network and Random Forest Classifier was imported from sklearn. Ensemble. Using these different algorithms different prediction models were built. While training the model the four parameters were recorded to select the best machine learning algorithm. They were Accuracy, Precision, Recall Score, F1 Score, and finally the confusion matrix. These statistics were calculated using sklearn.metrics module.

Following equations show how the calculations were performed in each parameter which we have discussed earlier.

Accuracy simply means the ratio of correctly predicted observations.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{False Negative} + \text{True Negative}}$$

Precision means the ratio of correctly predicted positive observations to the total predicted positive observations.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+ False Positive}}$$

Recall is the ratio of correctly predicted positive observations to all observations in actual positive cases.

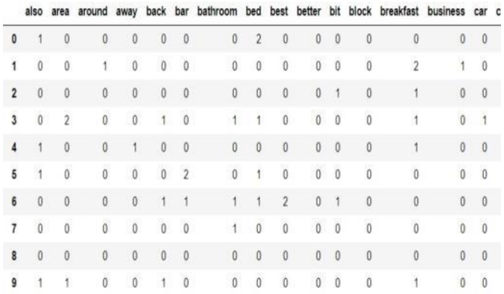
$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+ False Negative}}$$

F1 score means the average of precision and recall. Since here we are dealing with an uneven class distribution, the F1 score also needs to consider, when selecting the best machine learning classifier.

$$\text{F1 Score}= \frac{2^*(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

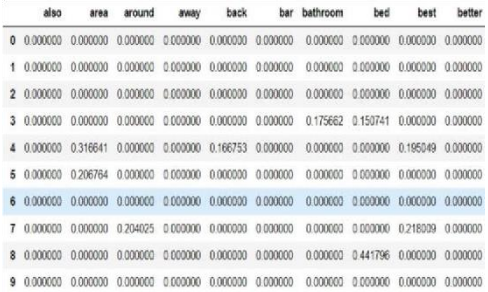
## IV. RESULTS

The following Table I shows the result obtained after using different classifiers. The result shows that the accuracy was higher when using the TF-IDF method as the feature extraction technique rather than using the Countervectorizer method. It is because in TF-IDF method considered the overall document weightage of words. The meaning of it is, it takes into consideration how often the word appears in the document plus how often the word appears across all documents in the data set. But in CountVectorizer method only consider the frequency of the word count within each review. Following Figure 3 shows the matrix of CountVectorizer and Figure 4 shows the matrix of the TF-IDF method.



| | also | area | around | away | back | bar | bathroom | bed | best | better | bit | block | breakfast | business | car | c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | |
| 3 | 0 | 2 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | |
| 5 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 6 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 9 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | |

Figure 3. Count Vectorizer Matrix



| | also | area | around | away | back | bar | bathroom | bed | best | better |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 1 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 2 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 3 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.175662 | 0.150741 | 0.000000 | 0.000000 |
| 4 | 0.000000 | 0.316641 | 0.000000 | 0.000000 | 0.166753 | 0.000000 | 0.000000 | 0.000000 | 0.195049 | 0.000000 |
| 5 | 0.000000 | 0.206764 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 6 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 7 | 0.000000 | 0.000000 | 0.204025 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.218009 | 0.000000 |
| 8 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.441796 | 0.000000 | 0.000000 |
| 9 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |

Figure 4. TF-IDF Matrix

When it comes to machine learning algorithms, the highest accuracy result and correct confusion matrix were given when using the SGD classifier and Logistic Regression with the TF-IDF method. The lowest accuracy was given when using the BernoulliNB with both the CountVectorizer method and TF-IDF method.

Table 1. Result Table

| | **CountVectorizer** | **Term Frequency Inverse Document** |
|---|---|---|
| **Multinomial_N B** | Accuracy:86.68 %<br>Precision Score:90.82<br>Recall Score:89.51<br>F1 Score:90.16<br>Confusion Matrix:<br>[[1999 480]<br>[557 4751]] | Accuracy:85.44 %<br>Precision Score:90.64<br>Recall Score:87.63<br>F1 Score:89.11<br>Confusion Matrix:<br>[[2015 479]<br>[655 4638]] |
| **Bernoulli_NB** | Accuracy:64.67 %<br>Precision Score:75.44<br>Recall Score:71.36<br>F1 Score:73.35<br>Confusion Matrix:<br>[[1251 1232]<br>[1519 3785]] | Accuracy:74.07 %<br>Precision Score:78.02<br>Recall Score:86.64<br>F1 Score:82.11<br>Confusion Matrix:<br>1136 1305]<br>[ 714 4632]] |
| **Logistic Regression** | Accuracy:84.51 %<br>Precision Score:89.84<br>Recall Score:87.06<br>F1 Score:88.43 | Accuracy:87.39 %<br>Precision Score:92.62<br>Recall Score:89.00<br>F1 Score:90.77 |

| | | |
|---|---|---|
| | Confusion Matrix:<br>[[1972 521]<br>[685 4609 ]] | Confusion Matrix:<br>[[2140 374]<br>[580 4693]] |
| **SGD Classifier** | Accuracy:85.40 %<br>Precision Score:90.42<br>Recall Score:88.03<br>F1 Score:89.21<br>Confusion Matrix:<br>[[1950 498]<br>[639 4700]] | Accuracy:87.71 %<br>Precision Score:92.64<br>Recall Score:89.40<br>F1 Score:90.99<br>Confusion Matrix:<br>[[2177 372]<br>[555 4683]] |
| **Linear SVC** | Accuracy:82.05 %<br>Precision Score:87.99<br>Recall Score:85.16<br>F1 Score:86.56 | Accuracy:85.96 %<br>Precision Score:90.79<br>Recall Score:88.38<br>F1 Score:89.57 |
| | Confusion Matrix:<br>[[1889 614]<br>[784 4500]] | Confusion Matrix:<br>[[2001 476]<br>[617 4693]] |
| **Random Forest Classifier** | Accuracy:80.44 %<br>Precision Score:79.29<br>Recall Score:96.10<br>F1 Score:86.89<br>Confusion Matrix:<br>[[1217 1318]<br>[205 5047]] | Accuracy:83.34 %<br>Precision Score:86.50<br>Recall Score:89.60<br>F1 Score:88.02<br>Confusion Matrix:<br>[[1724 744]<br>[553 4766 ]] |
| **MLP Classifier** | Accuracy:84.51 %<br>Precision Score:88.57<br>Recall Score:88.71<br>F1 Score:88.64<br>Confusion Matrix:<br>[[1832 612]<br>[603 4740 ]] | Accuracy:85.90 %<br>Precision Score:88.28<br>Recall Score:87.36<br>F1 Score:87.82<br>Confusion Matrix:<br>[[1811 621]<br>[677 4678]] |

As per the results, the highest accuracy (87.71%) was obtained by the SGD Classifier when using the TF-IDF method for feature extraction. Not only for the accuracy, for the other three parameter also had the highest scores for the SGD classifier with the TF-IDF.

## V. DISCUSSTION AND CONCLUSION

According to the result obtained from this research, it is shown that in feature extraction, when the TF-IDF method is used, the accuracies of all classification algorithms were higher than the Count Vectorization method. Since the TF-IDF method considers the overall document weightage of the words, it gave higher accuracy for the sentiment analysis.

Among these, all classifiers Logistic Regression and SGD Classifier were outperformed rather than other classifiers. And the Bernoulli_NB classifier was the worst classifier to be used for the sentiment analysis.

In conclusion, to enhance the accuracy of the sentiment identification model initially the reviews had to be preprocessed to remove unnecessary words, symbols and to remain only the important words. And for the feature extraction, it is good to use the TF-IDF method rather than using other methods. For sentiment analysis, the Bernoulli_NB classifier was not good enough and the Logistic Regression and SGD Classifier performs best and gave better results.

## VI. FUTURE WORKS

In this research, several most used classifiers were tried out to find the best classifier among them. So, in the future, other classifiers also can be tried out to find the best one. Furthermore, other feature extraction methods also can be tried out to find the best approach. The proposed system works only for the English language; therefore, this can be enhanced to identify any language text's sentiment. Therefore, it could be useful for identifying the sentiment of reviews that are written in Sinhala and that could be very much useful in the Sri Lankan context. Here for convenience, we have removed the emoji at the very beginning during the pre-processing. But the emoji's also play a significant role to identify the sentiment of a text, so that feature also can be added to identify the sentiment of the emoji as well. And in the future, it is expected to build a useful application for the hotel domain using this proposed sentiment analysis model.

## REFERENCES

Sharma, A.2017.Hotel Review. *Kaggle*. [Online]. [Accessed 19 June 2020]. Available from: https://www.kaggle.com /anu0012/hotel-review/

Bhargav PS, Reddy GN, Chand RVR, et al., (2019) Sentiment analysis for hotel rating using machine learning algorithms, *Int. Journal of Innovative Technology and Exploring Engineering*,8,1225-1228. DOMO, "Data Never Sleeps 6.0" in The Business Cloud, 2020. [Online]. Available: https://www.domo.com/learn/data-neversleeps-6.

Farisi A,Sibaroni Y,Faraby S,(2019) Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier, *Journal of Phys.: Conf. Series*,1192.

Ghosh M, Sanyal G, (2017) Preprocessing and Feature Selection Approach for Efficient Sentiment Analysis on Product Reviews, *Advances in Intelligent Systems and Computing*,515, v-vii.

Kasper W,Vela M,(2011) Sentiment Analysis for Hotel Reviews, *Proceedings of the Computational Linguistics-Applications Conf.*,231527,45-52.

Nohh N,Megat N,Zainuddin M,(2019) Sentiment Analysis towards Hotel Reviews, *Open International Journal of Informatics (OIJI)*, 7,1-19.

Patel A,Chheda B, Jain B, et al., Sentiment Analysis of Customers Opinions on Hotel Stays using Voted Classifier, International *Journal of Engineering Research & Technology (IJERT), V*9,827-833.

Shi H,Li X,(2011) A sentiment Analysis model for hotel reviews based on supervised learning, *Proceedings - International Conference on Machine Learning and Cybernetics*,3,950-954.

Symeonidis S,Effrosynidis D,(2018) A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis, *Expert Systems with Applications*,110,298310.

Tripathy A, Agrawal A, Rath S, (2015) Classification of Sentimental Reviews Using Machine Learning Techniques, *Procedia Computer Science*, 57, 821-829.

Yordanova S, Kabakchieva D, (2017) Sentiment Classification of Hotel Reviews in social media with Decision Tree Learning, *International Journal of Computer Applications*,158,05,1-7.

Zvarevashe K, Olugbara O ,(2018) A Framework for Sentiment Analysis with Opinion Mining of Hotel Reviews, *2018 Conference on Information Communications Technology and Society, ICTAS 2018 - Proceedings*,1-4.

## AUTHOR BIOGRAPHIES

Udani kaushalya, A Final Year Undergraduate at the Faculty of Applied Sciences, Rajarata University of Sri Lanka. Following BSc. (Hons) degree in Information Technology. Currently working as Quality Assurance Intern at Echonlabs (Pvt) Ltd. Her research areas are Natural Language Processing, Speech Recognition, Artificial Intelligence, and Data Science.

Wiraj Udara Wickramaarachchi is a lecturer at Department of Computing, Rajarata University of Sri Lanka. Currently reading Doctoral Degree at Wuhan University of Technology, in China. His research interests are Biometrics, Information Security, Privacy protection, Image processing and Natural Language processing.