Analysis on applying machine learning and classification approaches to predict the fraudulent reviews on the Yelp dataset

[1]Dr.S.Vijaya Ragavan, [2]M.Sangeetha, [3]D.Shalini, [4]Shaik Ishaq
[1]Professor, [2,3]Assistant Professor, [4]UG Student, [1,2,3,4]Department of Computer Science Engineering, Visvesvaraya College of Engineering & Technology, Hyderabad, India.

Abstract
The goal of our study, which is summarized in this article, is to create a machine learning model that can determine if reviews in Yelp's dataset are authentic or not. To determine which machine learning categorization approach would produce the best results, we specifically applied and contrasted them. To make it easier to understand why some approaches are preferable to others in specific situations, brief explanations are provided for each of the categorization strategies. The SVM classification algorithm produced the best result, with an F-1 score of 0.91 in the forecast.
Key Words: Classification, regression, svm, naviebayes, logestic regression, supervised learning

INTRODUCTION

Nowadays, with the proliferation of internet information, consumers frequently read reviews before making a purchase at a restaurant, a hotel, or any establishment they require. Yelp is a crowd-sourced review site and business directory that is frequently used by users to publish reviews about the companies they have dealt with. According to statistics, there will have been over 177 million reviews on the Yelp website by the end of 2018. Both consumers and companies profit from it. A company owner benefits from free promotion from consumers who provide a helpful and favorable evaluation of their establishment. Unfortunately, the issue occurs when some careless business owners attempt to increase their profits in their market by hiring people to create some fake reviews about their business on Yelp website.

Yelp realizes this potential threat will create misleading information for their users. To overcome this problem, Yelp has already provided reviews policy for business owners. Other than that, Yelp has also implemented a recommended software system that aims to automatically filter all reviews have been determined to be problematic. In order to keep their content helpful and reliable, Yelp tries not to highlight reviews written by users that they do not know much about or reviews that may be biased because they were solicited from family, friends, or favoured customers. The reviews are evaluated based on quality, reliability, and user activity. Currently, about 75 percent of all reviews on Yelp website is recommended.

However, no system or method can be truly foolproof. In an attempt to improve the accuracy of identifying fake reviews, machine learning can be very useful. In particular, machine learningclassification techniques can learning from data and then be applied to separate truthful reviews from fake ones.

The rest of this paper is organized as follows. Section II reviews relevant literature that sets the scene and forms the foundation of our research.In particular, it surveys four popular machine learning classification approaches. Section III explains our method. Section IV presents preliminary results of our method. Finally, Section V concludes the paper.

SYSTEM DESIGN

UML stands for Unified Modeling Language and isan acronym that identifies the same. In essence,UML is a way to create models and documentationfor software. One of the most common businessprocess modeling techniques is now in use.Diagrammatic depictions of software componentsare at the core of it. “A picture is worth a thousand words, like the saying goes. Using visualrepresentations helps we better comprehend possibleerrors or problems in business processes or software.

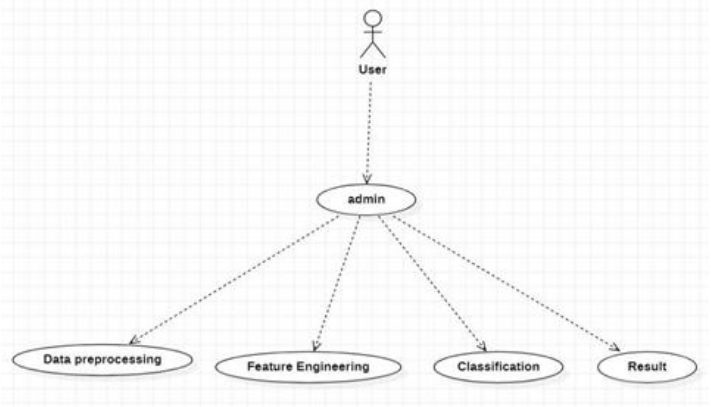
Because of the confusing nature of software design and documentation, UML was invented. For software systems, there were a variety of techniques in the 1990s. A more unified way to visually represent those systems arose, and as a result, three software engineers at Rational Software developed the UML during 1994-1996. It was accepted as the standard in 1997, and it has since received only afew minor updates, remaining as the standard.

GOALS: The following are the primary design goals of UML: A consistent, user-friendly, descriptive language that

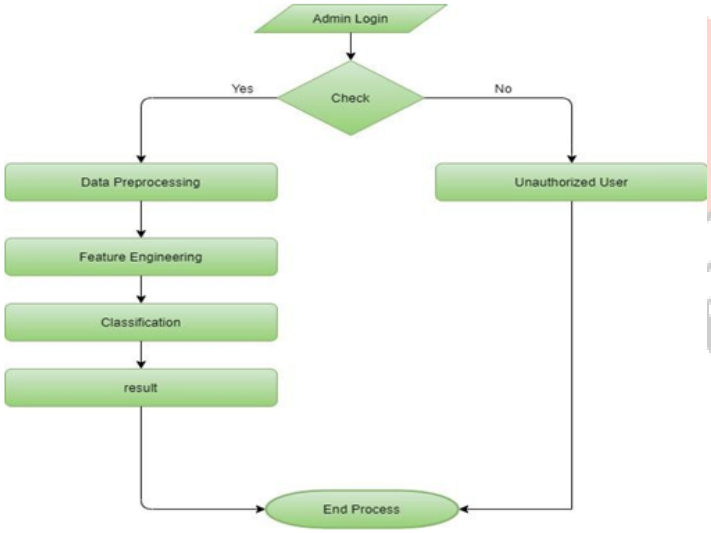
people can use to buildmodels and share them. Provide mechanisms to extend and specialize the core concepts. Operate freely regardless of the language or process. This formal modeling language understanding has a basis in how it is structured. Boost the development of OOtoolmakers.

USE CASEDIAGRAM

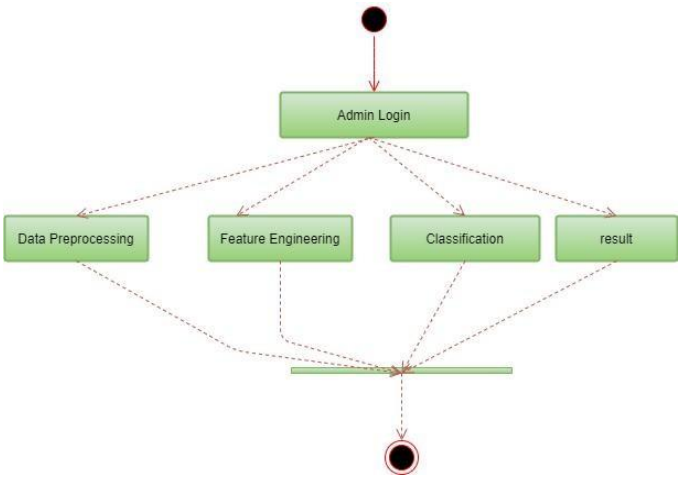
In Unified Modeling language (UML) terms, a use case diagram is a type of behavioral diagram that starts with a use case analysis. Its goal is to describe the functionality of a system in terms of actors, goals, and dependencies using a visual representation. The use case diagram serves two purposes: It reveals which actor is the primary user of the system, and which system features they rely on. There are ways to illustrate the actors' roles in the system.



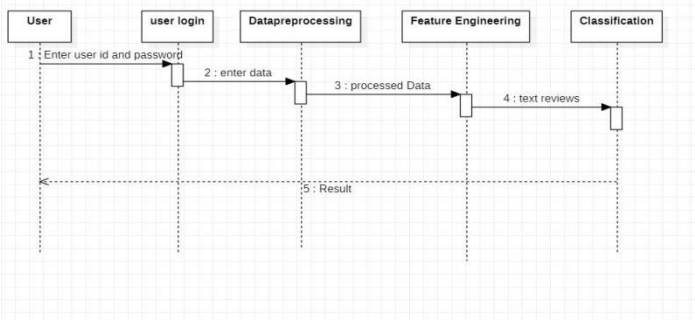
ATA FLOW DIAGRAM:A data-flowdiagram is a way of representing a flow of datathrough a process or a system. The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow — there are no decision rules and no loops.



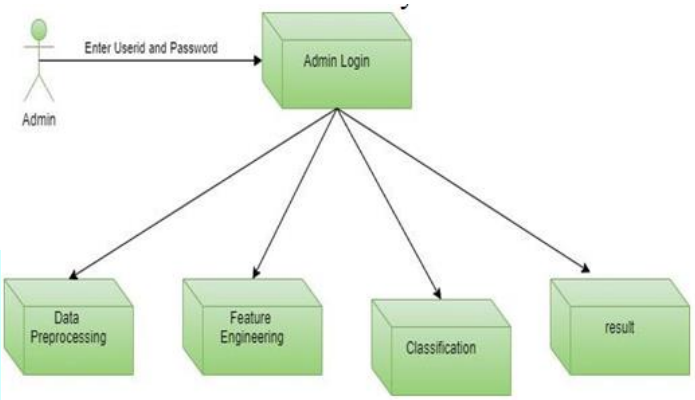
Activity Diagram: In UML, anactivity diagram provides a view of the behavior of a system by describing the sequence of actions in a process.



Sequence Diagram: A sequence diagram isa Unified Modeling Language (UML) diagram that illustrates the sequence of messages between objects in an interaction. A sequence diagram consists of a group of objects that are represented by lifelines, and the messages that they exchange over time during the interaction**.**



Component Diagram: Component diagrams are used to visualize the organization and relationships among components in a system. These diagrams are also used to make executable systems.



OUTPUTSCREENSHOTS

Figure 1: User Login Details

Within this screen, we can see that we have a home page, and we have to login in for the further process.In above screen showing the details of the user for the login process. Once the details are entered by theuser, we have to click LOGIN button.

Figure 2: New User Registration Details

In above screen showing the details of the user for the login process. Once the details are entered by the user, we have to click LOGIN button. We can create a new account

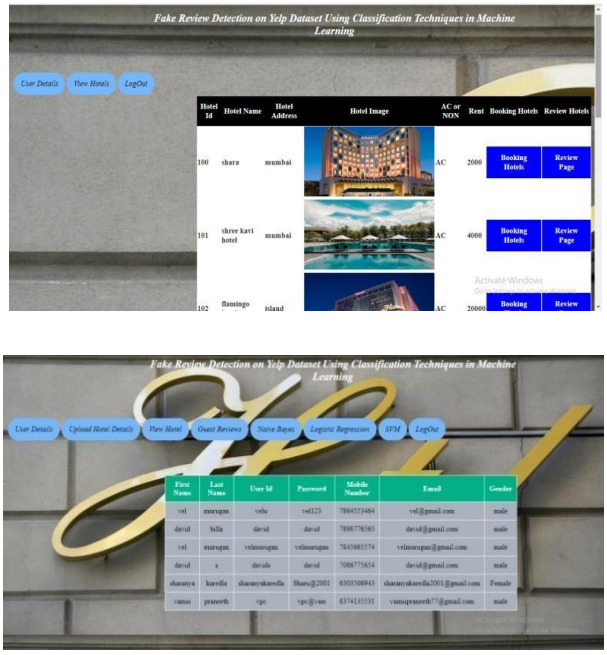


Fig. 3. In above screen showing the Data of hotels

Within the screen, wehave a data of hotels and alsowe have booking hotels and reviewing the hotels.

Fig.4. In above screen uploading reviews

After selecting the review button it redirect to abovepage within the page we can give reviews

Fig.5: Admin page login

In admin page login admin can login to the websiteand preview the hotels, reviews

By login to the Admin page he can see user detailson yelp dataset

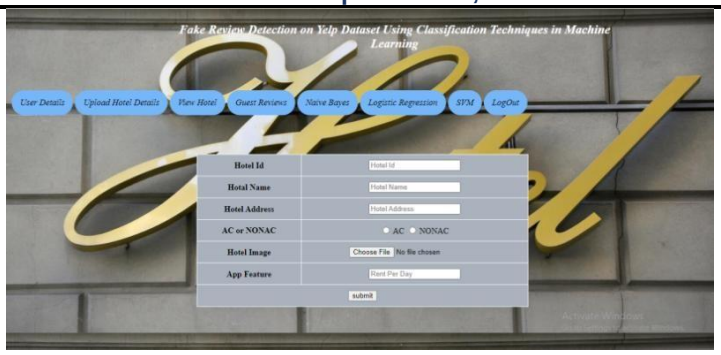


Fig.7: Admin uploading hotel details

In this page admin can upload new hotel details orhe can modify the exiting hotel details

Fig.8: review details on yelp dataset

By login to the admin profile he can access thereviews given to the hotels

Accuracy	Precision	Recall	F1-Score
accuracy: 0.814	0.1397815452091768	0.1357704302732498	0.13595309149831145

Fig.9. Within the screen showing the Accuracy of logistic regression

After adding the data set we analyse the data by using the logistic regression Algorithm. The above screen is showing the data by using the algorithm. The data represents accuracy, precision, recall, F1- score

Accuracy	Precision	Recall	f1-score
accuracy: 0.598	0.045998445998446	0.07692307692307693	0.057570747836234566

Fig.10 above screen showing the accuracy resultof the naivebayes

Within the screen showing the accuracy of Naive bayes. The data represents accuracy, precision, recall, F1- score

Accuracy	Precision	Recall	f1-score
accuracy: 0.851	0.10631313131313132	0.125	0.11490174672489084

Fig.12 above screen showing the accuracy resultof the SVM.

Within the screen showing the accuracy of SVM.The SVM accuracy results displays accuracy, precision, Recall, F1-score.

CONCLUSION

This study examined four well-known machine learning classification techniques for identifying phony Yelp reviews. Reviews that are given ratings like "helpful," "cool," and "funny" are only gained through unfiltered reviews, which mean that as soon as Yelp filters a review, it is buried and cannot be rated by anyone else. The experiment's findings demonstrated an extremely high prediction score when applying SVM. The dataset's limitations prevent us from

implementing a number of features, including the user trust factor based on user friendship and the user profile (join date, photo, etc.). Because imbalanced datasets produce subpar results in our experiment, imbalance on the dataset must be addressed. While running the experiment, we found that Logistic Regression took the longest time to train the model, and Gaussian Naïve Bayes gave the lowest score on average. In our opinion, we cannot say that reviews got filtered by YELP recommendation system is 100% fake, because there are still other factors that may lead machine learning into false prediction. Other technique that are potentially reliable and can be used for filtering review is using verified buyer method as some crowd sourceweb have been used.

REFERENCES

1. Recommended Reviews | Support Centre | Yelp." [Online]. Available: https://www.yelp-support.com/Recommended_Reviews?l=en_US. [Accessed: 07-Aug-2019].
2. H. I. Bülbül and Ö. Ünsal, "Comparison of classification techniques used in machine learning as applied on vocational guidance data,"Proc. 10th Int. Conf. Mach. Learn. Appl. ICMLA 2011, vol. 2, pp. 298–301, 2011.
3. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve Bayes and logistic regression," 2017 Int. Conf. Comput. Commun. Informatics, ICCCI 2017, 2017.
4. J. R. Brzezinski and G. J. Knafl, "Logistic regression modeling for context- based classification," pp. 755–759, 2008.
5. DoúarandD.Ünal," Comparison of Data Mining Classification Algorithms Determining the Default Risk," Sci. Program., vol. 2019.
6. S.Edition, A First Course in Machine Learning, Second Edition.2018.