International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Recommender System- Analyzing products by mining Data Streams

Siddhi Divekar[1], Gunashree Attarde[2], Adesh Chavan[3], Ankita Dahiphale[4], Prof. Rahul Patil[5]

[1,2,3,4]*Students, Dept. of computer engineering, Pimpri Chinchwad College of engineering, Pune, Maharashtra, India.*
[5]*Professor, Dept. of computer engineering, Pimpri Chinchwad College of engineering, Pune, Maharashtra, India.*

Abstract - *Due to the spread of covid many people lost their jobs so to earn their lives they started with small occupations. But these occupations are still unknown and are not able to earn profits. So as a helping hand to these people we have come up with an ecommerce website which will help them earn profits and get real review from the customers which will help them improve in their sectors. For earning the profits, we are about to build a recommendation system by analysing the best sales of a product using the Boyer Moore Voting Algorithm. The analyses of the product will also be shown using data visualization using the Power BI Software. We will be using the various algorithms like the SVM, LinearSVC, Naïve Bayes, etc for detecting whether the provided review is real or. Fake*

***Key Words*: Data stream mining, Power BI, Recommendation, Review, SVM, Naïve Bayes and Ecommerce.**

1. INTRODUCTION

Due to the pandemic situation many people started their own small-scale business. We are providing an e-commerce platform for these small-scale entrepreneurs which would help them to sell their products and get product Reviews from the customers. The reviews Recommendation will be based on two types:

1. The reviews from the customers will be in the streaming form which will be then converted into data visualization and further will help in the product recommendation system.

2. The supplementary occupations from which they can also buy supplementary products with the actual product purchased will help in the occupation recommendation system.

3. We will also take care of the review submitted are not fake by applying the various false review algorithms

2. BACKGROUND

[1] Many of our regular activities have been affected by the Internet's fast expansion. Ecommerce is one of the fastest- growing areas. Customers can post evaluations about e- commerce services in general. These reviews might be utilized as a source of data. Companies, for example, can use it to develop goods or services, while potential customers can use it to determine whether to buy or use a product. Unfortunately, some people have tried to generate false reviews in order to boost the popularity of the product or to discredit it. The goal of this study is to use the language and rating properties of a review to detect fraudulent product reviews. In summary, the suggested system (ICF++) would assess the honesty of a review, the trustworthiness of the reviewers, and the product's dependability. Text mining and opinion mining techniques will be used to determine a review's honesty value. The results of the experiment demonstrate that the suggested system has a higher accuracythan the iterative computation framework (ICF) method's outcome.

[2] Fake review detection has gotten a lot of attention in recent years. Both the business and research communities are paying attention to this issue. For Detecting reviews that represent actual user experiences and opinions Fake reviews are a significant issue. The benefits of supervised learning are numerous. One of the primary methods to resolving the issue Obtaining branded bogus training reviews, on the other hand, is challenging. because it is extremely difficult, if not impossible, to properly identify fakes manual examinations Various forms of data have been utilized in previous studies. Training reviews that aren't entirely true. The faux false evaluations created with the Amazon Mechanical Turk (AMT) crowdsourcing tool are maybe the most intriguing. Using simply word n-gram characteristics, reported an accuracy of 89.6% using AMT created bogus reviews. This level of precision is both shocking and promising. The AMT produced reviews, albeit false, are not actual bogus reviews on an e- commerce website. The Turkers are unlikely to be in the same psychological condition as the authors of actual bogus reviews who have enterprises to promote or downgrade other products while producing such evaluations. This notion is supported by our research. Following that, it's reasonable tocompare fake review detection accuracies on pseudo-AMT data with real-life data to determine if various states of mindmay lead to different writings and, as a result, different classification accuracies. We undertake a complete set ofclassification tests using just n-gram features for actual review data, using all filtered and non-fake reviews from Yelp.com. Although the accuracy of false review identification on Yelp's

real-life data is just 68.8%, this accuracy suggests that n-gram characteristics are definitely useful. The information theoreticmeasure KL-divergence and its asymmetric attribute are then used to offer a novel and principled technique for determining the precise difference between the two types of review data. This exposes some fascinating psycholinguistic phenomena concerning false reviewers, both forced and natural. We offer a new set of behavioral characteristics about reviewers and their reviews for learning to enhance classification on real-life Yelp review data, which substantially improves theclassification result on real-life opinion spam data.

[3] Power BI has completely changed the business data visualisation, intelligence, and analytics worlds. Power BI is a web-based application that enables users to search for data, convert it, visualise it, and share the reports and dashboards they create with other users in the same or different departments/organizations, as well as the general public. As of February 2017, Power BI was used by over 200,000 businesses in 205 countries. Power BI has emerged as a viable competitor for use as a business intelligence tool in small and medium businesses, thanks to a free version that includes sufficient features and capabilities. Power BI's Quick Insights feature (Michael Hart, 2017) is a new tool built on a growing collection of powerful logical algorithms. After upload dataset to PowerBI, a single click may activate this function, which generates a number of reports based on the data's analysis without the need for human interaction. This also aids in reducing human mistakes in computations, statistical procedures, which may result to research that isn't verified. PowerBI is simple to use as a platform for Research Data Analysis, visualizations and accepting even Excel files as input. The goal of this article is to demonstrate how quickly Power BI can turn a dataset of research data into a collection of reports and dashboards that can be simply shared.

[4] The ability to store, gather, and manipulate data has greatly increased as technology has advanced. Data analysis has grown more crucial as the amount of information and its complexity grows at a rapid pace. The purpose of this article is to suggest to the user goods that are more likely to be purchased. This paper initially discusses several recommendation approaches and research on recommendation systems, before proposing a better strategy for a successful recommendation system and explaining the outcomes of that approach. On a transactional dataset, is combinations of the k-means clustering method and the apriori algorithm is used to provide a better recommendation list.

[5] The quantity and influence of online reviews grows because of the growth in the significance of internet worldwide. Comments, reviews and feedback about services are very important for the items and service providers because they influence the consumers and frequently are the most convenient method for the customer to decide if they can buy a particular product or not. Reviews can have a positive as well as negative impact. And hence, trusting reviews blindly is not advisable because they involves risk both for the customers and sellers. Some selling organizations sometimes offer incentives to people who post positive reviews and feedbacks for their particular services on the other hand others may pay to some people to write negative reviews for their competitor product service providers. Thus, providing a bad influence over the consumers and deflecting their decision of buying a product or not. Such false reviews are called as spam reviews and are very common in online E-Commerce systems. Moreover, consumers must also be careful while going through the reviews and selecting an particular product or service to make the decision based on reviews. In this article, we explain how the suggested system aids in the detection and removal of false reviews, with a focus on data mining techniques utilizing the "J48 Algorithm," as well as the system's performance

[6] User input in the form of app ratings and reviews is becoming increasingly common in app stores. Researchers and, more recently, tool providers have provided analytics and data mining solutions to developers ana analysts for eg, to assist release choices. Positive feedback, according to research, boosts app downloads and revenue, and therefore it's success. As a result, a market for pho bogus, incentivized app evaluations arose, with yet-to-be-determined ramifications for developers, app users and owners.This study investigates false reviews, their sources, characteristics, and the degree to which they may be identified automatically. To understand their tactics and services, we ran disguised questionnaires with 43 bogus review providers and analyzed their review rules. We discovered substantial discrepancies between the matching applications, reviewers, rating distribution, and frequency by comparing 60 thousands bogus reviews with 62 millions review from the App Store. This prompted the creation of a simple classifier that can automatically detect fraudulent app store reviews. Our classifier has a recall of 91 percent and an AUC/ROC value of 98 percent on a labelled and unbalanced dataset with one- tenth of false reviews, as documented in other areas. Our findings are discussed, as well as their implications for software engineering, app consumers, and app store owners.

[7] The importance of internet evaluations on businesses has risen dramatically in recent years, and they are now critical in determining business performance in a wide range of industries, from restaurants to hotels to e-

commerce. Unfortunately, some individuals utilize unethical methods to boost their internet image, such as creating false reviews of their own companies or competitors. Fake review detection has already been studied in a variety of sectors, including product and company evaluations in restaurants and hotels. Despite its economic importance, however, the consumer electronics industry has yet to be properly investigated. This paper presents a feature framework for identifying fraudulent reviews in the consumer electronics area, which has been tested. The four part contribution is as follows- a) creating a database with four different cities for consumer electronics domain in order to classify the fake reviews. b) identify a feature framework for detection of false reviews. c) on the proposed framework development of classification method. d) analyse the output for each cities. The Ada Boost classifier has been proved to be the best by statistical methods according to the Friedman test, with an F-score of 82 percent on the classification job.

[8] In this field of study, two types of datasets are typically used: pseudo-fake and real-life evaluations. When compared to pseudo fake reviews, literature shows that classification models perform poorly in real-world datasets. Following our analysis we discovered that behavioral and contextual factors are crucial for detecting fraudulent reviews. In particular, we utilized an important behavioral aspect of reviewers known as "reviewer deviation." Our research focuses on the relationship between reviewer deviance and other environmental and behavioral factors. The relevance of a certain feature set for a classification algorithm to detect fraudulent reviews was empirically demonstrated. We rated features in a chosen feature set, and reviewer deviation came in eighth. We scaled the dataset to test the feasibility of the selected feature set and found that scaling the dataset can increase both recall and accuracy. A contextual feature in our chosen feature set captures text similarity between a reviewer's reviews. For calculating text similarity of reviews, we used the NNC, LTC, and BM25 term weighting methods. BM25 outperformed other word weighting schemes, according to our findings.

3. PROPOSED SYSTEM

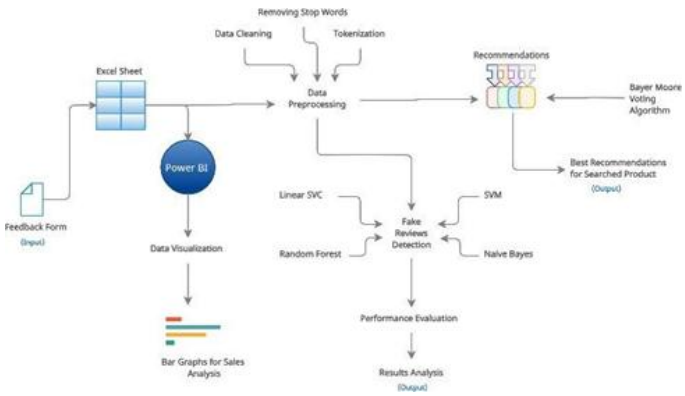

Fig-3.1-System Diagram

3.1 WORKING OF THE SYSTEM

The main motive of our system is to get recommendation and to identify whether a review is true or fake.

So, to achieve the first motive that is the recommendation we will be generating a goggle form which will take feedback from the customers related to the purchased products. The data in the form will then be converted into an excel sheet which will be an input to the Power Bi software which will give us a clear data visualization of the products sales. Further this Product sales data will be given as an input to the streaming algorithm (Boyer Moore voting Streaming algorithm) after the data preprocessing which will help us in analyzing the best sales ofa product which can help the small-scale entrepreneurs to analyze their profits and loss.

The next motive is to let the small-scale entrepreneurs know whether the review provided through the google feedback form are true or fake. We will be using various machine learning algorithms like the naïve bayes, SVM, random forest which will us to classify whether the review is true or fake.

Parameters on which the review will be classified are:

Time span of the review

Technical terms in the review

Ratings

Verify the Purchase

Inspecting the user profile

Customer Jacking

3.2 STREAMING ALGORITHMS

Boyer Moore voting Streaming algorithm:

The Boyer-Moore voting method is one of the most often used optimum algorithms for determining the majority element among elements with more than N/ 2 occurrences.

Using this algorithm, we will get the best sales product for the recommendation purpose as the o/p.

Time Complexity = O(N) Space Complexity = O(1)

3.3 ML Algorithms for Fake Review detection Gaussian Naïve Bayes:

The gaussian Naïve Bayes is a type of the Naïve Bayes algorithm which acts in accordance with the Gaussian normal distribution. It also contributes to the continuous data.

LinearSVC:

This classifier divides data into groups by offering the best suited hyper plane.

SVM:

Various investigations have revealed If you employ SVC's default kernel, the Radial Basis Function (RBF) kernel, you're likely using a nonlinear decision boundary, which will greatly outperform a linear decision boundary in the case of the dataset.

Random Forest: This approach, which is supplied by the sklearn package, has also been used for classification by building numerous decision trees set randomly on a sample of training data.

After applying all of these classifiers, the accuracies of each are compared, and their accuracy for detecting false reviews is evaluated.

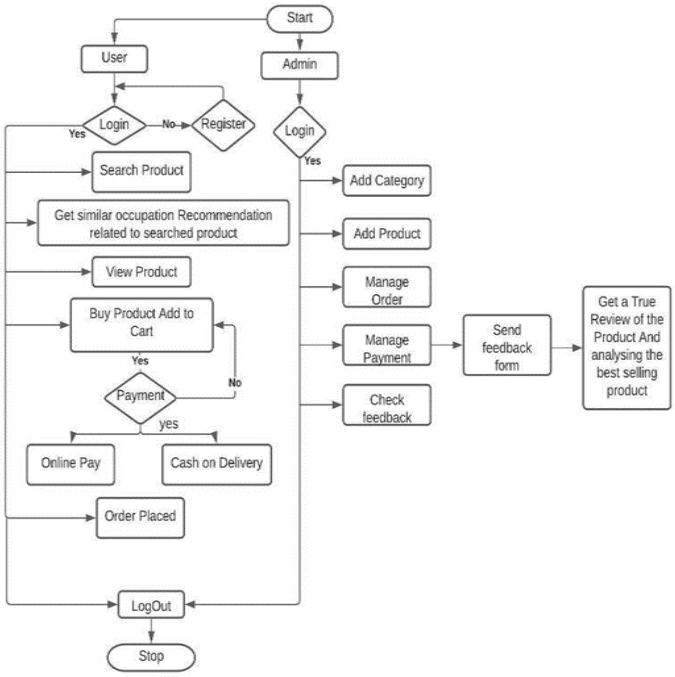

Fig-3.2-Flow Diagram

3.4 SDLC MODEL:

We will be using ITERATIVE MODEL. Because the iterative methodology starts with a modest implementation of a limited set of software requirements and repeatedly improves the evolving versions until the entire system is built and ready for deployment. The Iterative and Incremental model is depicted in the figure below.

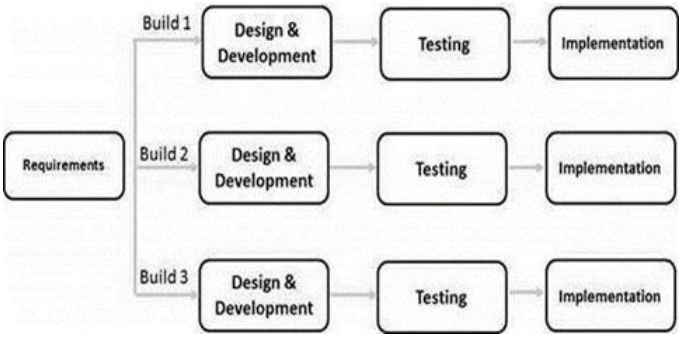

Fig-3.3-Model

3.5 UML

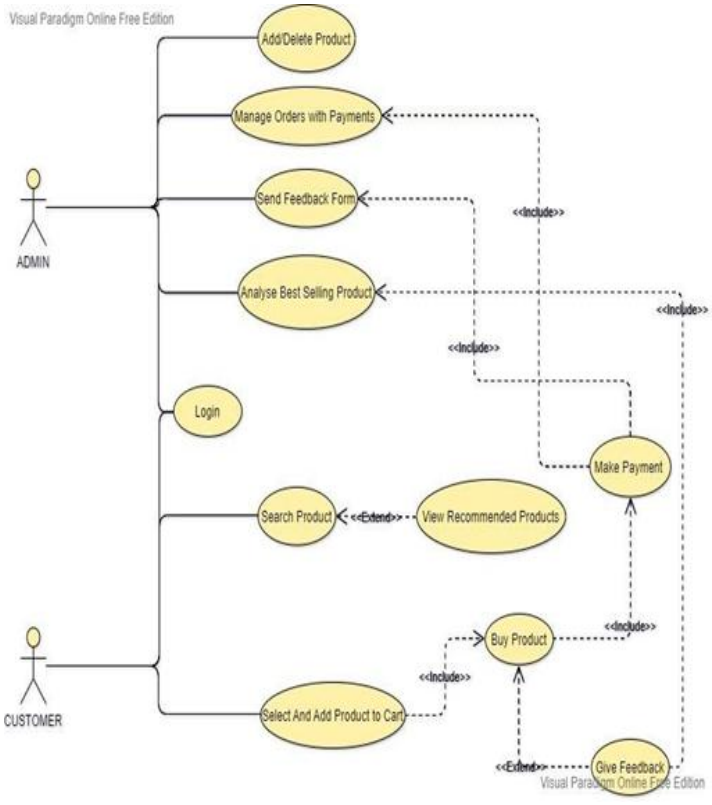

Fig-3.4-Use case Diagram

4, DEMO OF POWER BI STREAMING

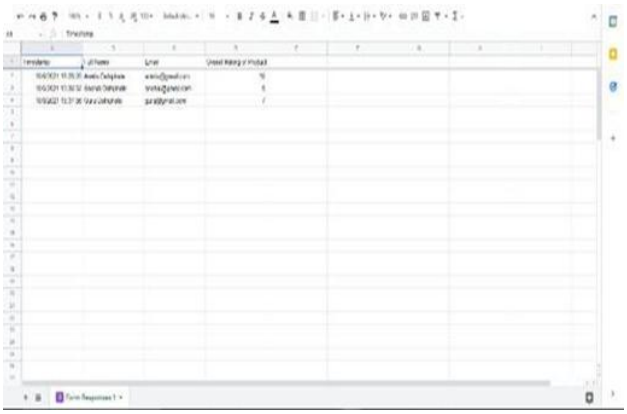

Fig-4.1-Streaming data from the form

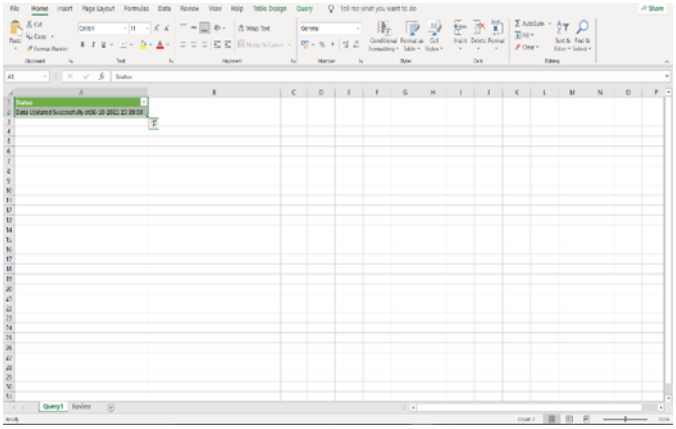

Fig-4.2- Successfully updated timestamp of streaming data

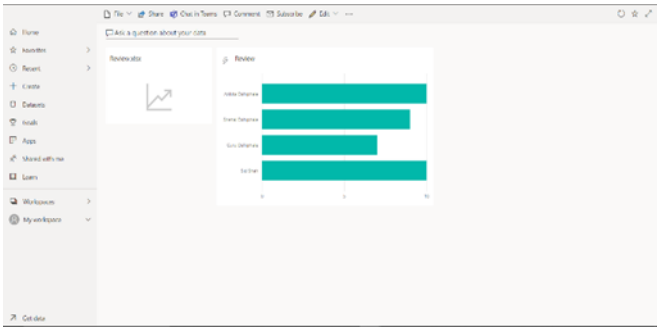

Fig-4.3-Data visualization of the product sales

6. CONCLUSION

We presented an overview of our ecommerce website whichwill help the people earn profits for the similar occupation recommendation of the searched product and also get a true review of their sales so that these reviews help them to improvise in their field. In future scope the website can also be used for the marketing the advertisement of the products to earn more profits.

7. ACKNOWLEDGEMENT

We express our heartfelt gratitude to Prof. Rahul Patil, our Project Guide, for his encouragement and support throughout our Project, particularly for the helpful ideas made during the Project and for laying the groundwork for our work's accomplishment.

We'd also want to express our sincere gratitude to Prof. Dr. S. V. Shinde, our Research & Innovation coordinator, and Prof. S. R. Vispute, our Project Coordinator, for their help, real support, and guidance from the beginning of the seminar until the end. We'd like to express our gratitude to Prof. Dr. K. Rajeswari, Head of the Computer Engineering Department, for her unflinching support during the seminar.

REFERENCES

[1]https://www.researchgate.net/publication/303499094_Fake_Review_Detection_From_a_Product_Review_Using_Modified_Method_of_Iterative_Computation_Framework

[2]http://www2.cs.uh.edu/~arjun/papers/UIC-CS-TR-yelp-spam.pdf

[3]http://ir.inflibnet.ac.in:8080/ir/bitstream/1944/2116/1/2

[4] Application of Data Mining to E-Commerce Recommendation Systems

[5]https://www.ijsr.net/archive/v7i10/ART20191163.pdf

[6]https://ir.inflibnet.ac.in/bitstream/1944/2116/1/24.pdf

[7]https://www.youtube.com/watch?v=AGrl-H87pRU