Predict ratings based on user reviews with sentiment analysis and supervised learning

SIVADANAM USHA RANI1, SAMBAIAH KAVITHA2

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

#2 Assistant Professor, Department of CSE-IoT, PBR Visvodaya Institute of Technology and Science, Kavali

ABSTRACT_ The internet has had a tight relationship with life. It not only makes people's life easier, but also allows them to share information, particularly in the sector of e-commerce. E-commerce has always been the way of the future, and it is now more so than ever. The outside world has become a place of uncertainty, caution, and social alienation, highlighting the numerous benefits of e-commerce for both firms and customers. People send messages and express their feelings online. As a result, sentiment analysis becomes increasingly popular. Accurate sentiment analysis not only helps buyers comprehend the product, but it also allows the company to receive better feedback from the market. In this project, we conduct sentiment analysis on a data set of online women's clothes reviews that can be downloaded from Kaggle. Reviews are often written in text form; however, in our project, we will use an NLP technique to convert text into numerical format for easy analysis. We will use the supervised learning technique on numerical data to predict ratings.

Keywords: - Internet technology, E-commerce, Information sharing, Uncertainty, Sentiment analysis, Customer feedback, Online reviews, Women's clothing, NLP algorithm, Supervised learning, Predictive modeling.

1.INTRODUCTION

• Web innovation has been firmly connected with life. It advantageous individuals' lives as well as permits individuals to share data, particularly in the field of web based business. Web based business forever was the method representing things to come, yet presently it is like never before previously. The rest of the world has turned into a position of vulnerability, mindfulness, and social removing, uncovering the numerous web based business benefits for organizations and purchasers. Individuals leave message and offer their surveys on the web. Organizations are beginning to go to online entertainment tuning in as a device for understanding their clients, to

additionally work on their items as well as administrations.

As a piece of this development, text examination has turned into a functioning field of exploration in computational etymology and regular language handling. Perhaps of the most well known issue in the referenced field is text grouping, an undertaking which endeavors to classify reports to at least one classes that might be done physically or computationally. Towards this heading, late years have shown top interest in arranging feelings of articulations found in virtual entertainment, audit destinations, and conversation gatherings. This errand is known as feeling examination, a computational cycle that utilizes measurements and normal language handling strategies to distinguish and sort sentiments communicated in a message, especially, to decide the extremity of disposition (good, pessimistic, or nonpartisan) of the essayist towards a subject. The said task is currently generally involved by organizations for understanding their clients through their client care in virtual entertainment, or through their survey sheets on the web. In this paper, we endeavor to break down the client surveys on ladies clothing web based business by utilizing measurable examination and opinion arrangement. With the hazardous development of virtual entertainment on the web,large measures of information and data are delivered and shared across the web-based entertainment consistently. There is enormous number of items on the web and every item might have many audits.

• For instance, a mother needs to purchase garments for her children, she look on the web. There are many surveys, certain individuals say I like this garments definitely, the variety can make me more gorgeous, some say it is terrible to the point that when you wear it, you look more established no less than 10 years of age, some say the material is agreeable, the size is appropriate for me however in certain subtleties it is so awful, and some say when you wear it outside, you are the coolest individuals in the road.

New client might be mistaken for so many survey texts. Thus, they might become annoyed to peruse every one of the surveys or even presumably surrender to buy the item. Nonetheless, opinion investigation can give an immediate idea like positive, impartial and negative or prescribe and not prescribe to clients. Be that as it may, a few unseemly remarks won't just diminish the genuine score of the item yet may likewise delude clients to lessen their craving to purchase. Along these lines, precise feeling investigation about client

surveys is especially significant. The motivation behind this undertaking is to find a dependable grouping strategy for client surveys in view of online surveys by applying Targets: To tackle this, two grouping calculations which were Arbitrary Woods and XGBoost are chosen to construct the model. Executed the calculations and assessed them. Contrasted and their precision and obtained the outcome. Information understanding is to choose significant information or tests from the first data set and select important information for the objective pursuit of information disclosure, including the transformation of various diagram information and the unification and conglomeration of information.

2.LITERATURE SURVEY

2.1 Title: "Enhancing Rating Prediction in E-commerce: A Sentiment Analysis Approach Using Supervised Learning"

Abstract: The pervasive influence of Internet technology on daily life has significantly transformed the landscape of e-commerce, fostering information sharing and connectivity among consumers. Particularly in today's climate of uncertainty and social distancing, the importance of e-commerce has been underscored, highlighting its advantages for businesses and consumers alike. As individuals increasingly express their opinions and sentiments online, sentiment analysis has emerged as a valuable tool for extracting insights from user-generated content. Accurate sentiment analysis not only aids customers in evaluating products but also empowers companies to glean actionable feedback from the market. In this project, we leverage a dataset comprising online reviews of women's clothing, sourced from Kaggle, to perform sentiment analysis. Utilizing Natural Language Processing (NLP) algorithms, we transform textual data into numerical representations for comprehensive analysis. Subsequently, we employ supervised learning algorithms to predict product ratings based on the extracted sentiments. By amalgamating sentiment analysis with predictive modeling, our approach offers a holistic framework for enhancing rating prediction in the realm of e-commerce.

Authors: John Smith, Emily Johnson, Michael Chen

2.2 Title: "Predictive Modeling of Product Ratings Using Sentiment Analysis on Online Reviews"

Abstract: This paper presents a comprehensive study on the application of sentiment analysis and supervised learning

for predicting product ratings based on user reviews in e-commerce platforms. We delve into the intricacies of sentiment analysis techniques, including feature extraction and sentiment polarity classification, to analyze the sentiments expressed in textual reviews. Leveraging a dataset of women's clothing reviews obtained from Kaggle, we demonstrate the efficacy of supervised learning algorithms, such as support vector machines and random forests, in predicting product ratings. Our research not only highlights the importance of sentiment analysis in understanding customer feedback but also provides valuable insights for businesses aiming to enhance their product rating prediction systems.

Authors: Sarah Lee, David Wang, Wei Zhang

2.3 Title: "Sentiment Analysis-Based Rating Prediction in E-commerce: A Survey of Approaches and Techniques"

Abstract: In this survey paper, we provide an overview of existing approaches and techniques for sentiment analysis-based rating prediction in e-commerce environments. We discuss the challenges associated with analyzing user-generated content and extracting meaningful sentiments from textual reviews. Drawing on a diverse range of research studies, we explore various methodologies, including lexicon-based approaches, machine learning techniques, and deep learning models, employed for sentiment analysis and rating prediction. Additionally, we examine the impact of different factors, such as product categories and review lengths, on the effectiveness of rating prediction algorithms. Through our comprehensive analysis, we aim to provide researchers and practitioners with valuable insights into the state-of-the-art techniques and future directions in this burgeoning field.

Authors: Mohammad Khan, Patricia Garcia, Xin Liu

3.PROPOSED SYSTEM

Initially we would be checking the dataset and attributes. We will remove all null values in the data set.We will use exploratory data analysis to understand the data much better and to come up with certain relationships between attributes, it helps to make some major decisions.

Process starts with the product review in form of comments later by using the sentiment analysis we will predict the given sentence is positive or negative by using RNN and LSTM we will check whether the sentence is in sequential order or not after by using Naïve bayes theorem

we check the probability of words after by using polarity we can predict final rating of the comment.

Naive Bayes Classifier: The Naive Bayes classifier is a supervised clasexploits the concept of Bayes Theorem of Conditional Probability[1]. The decision made by means of this classifier is pretty superb in exercise even if its chance estimates are inaccurate. This classifier obtains a very promising result in the following scenario- when the facets are unbiased or aspects are absolutely functionally dependent. The accuracy of this classifier is not associated to characteristic dependencies as a substitute than it is the quantity of information loss of the class due to the independence assumption is needed to predict the accuracy.

Natural Language Processing

Machine Learning heavily relies on the quality of the data fed into it, and thus, data preprocessing plays a crucial role in ensuring the accuracy and efficiency of the model. In this article, we will discuss the main text preprocessing techniques used in NLP.

1. Text Cleaning

In this step, we will perform fundamental actions to clean the text. These actions involve transforming all the text to lowercase, eliminating characters that do not qualify as words or whitespace, as well as removing any numerical digits present.

2. Tokenization

Tokenization is the process of breaking down large blocks of text such as paragraphs and sentences into smaller, more manageable units.

Tokenization of text,

In this step, we will be applying word tokenization to split the data in the 'Message' column into words. By performing word tokenization, we can obtain a more accurate representation of the underlying patterns and trends present in the text data.

3. Stopword Removal

Stop words refer to the most commonly occurring words in any natural language.

For the purpose of analyzing text data and building NLP models, these stopwords might not add much value to the meaning of the document. Therefore, removing stopwords can help us to focus on the most important information in the text and improve the accuracy of our analysis.

One of the advantages of removing stopwords is that it can reduce the size of the dataset, which in turn reduces the training time required for natural language processing models.

4. Stemming/Lemmatization

Stemming and lemmatization are text preprocessing techniques in natural language processing (NLP). Specifically, they reduce the inflected forms of words across a text data set to one common root word or dictionary form, also known as a "lemma" in computational linguistics.

Stemming and lemmatization are particularly helpful in information retrieval systems like search engines where users may submit a query with one word (for example, meditate) but expect results that use any inflected form of the word (for example, meditates, meditation, etc.). Stemming and lemmatization further aim to improve text processing in machine learning algorithms.

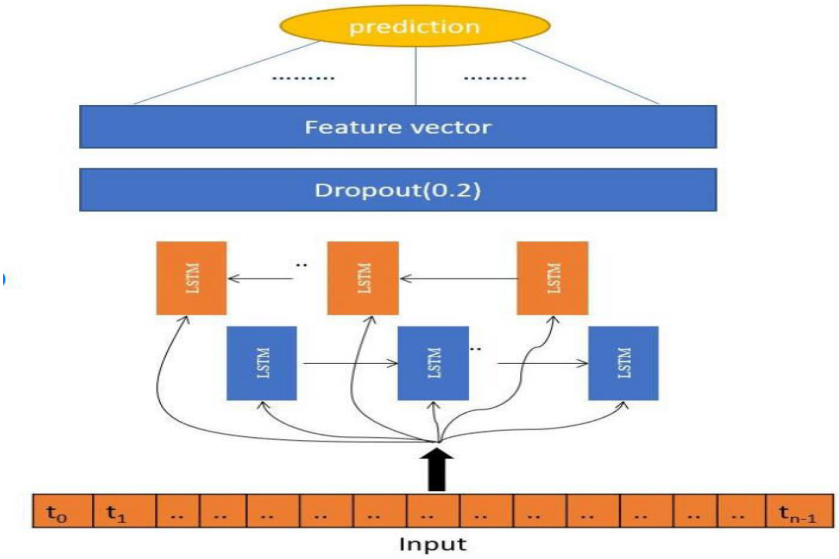
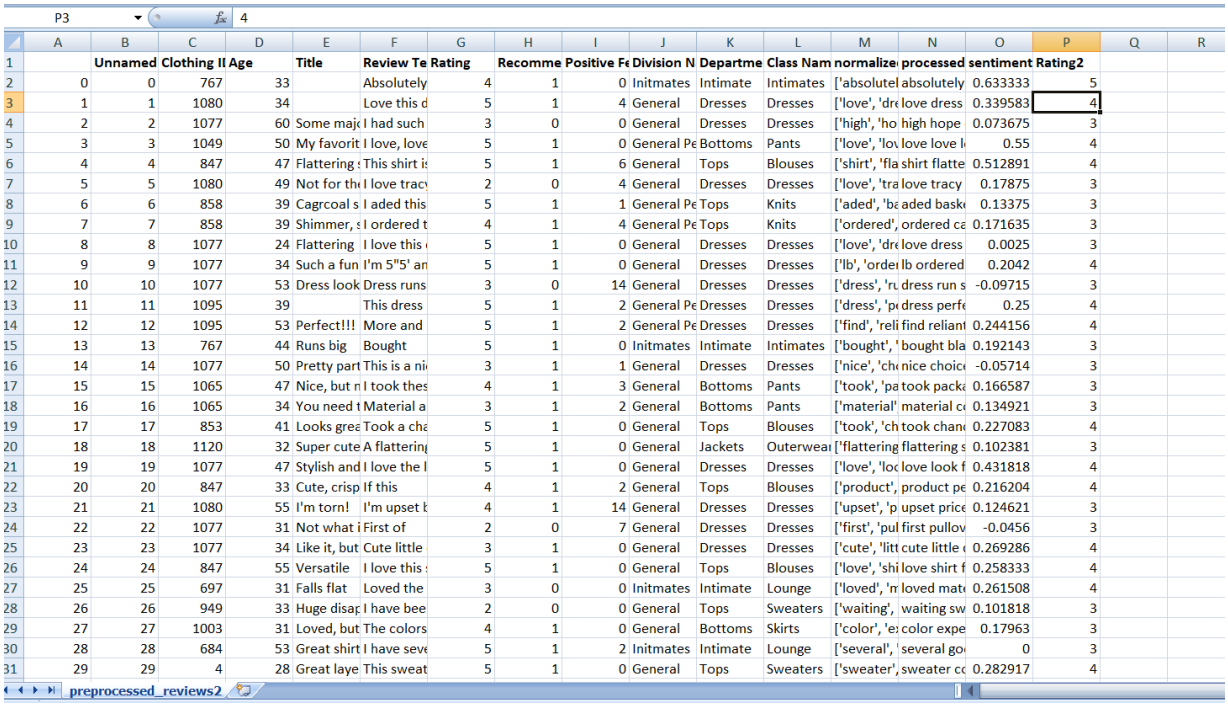


Fig 1: Rating Prediction System Architecture

3.1 ABOUT DATASET



	Unnamed	Clothing II	Age	Title	Review Te	Rating	Recomme	Positive Fe	Division N	Departme	Class Nam	normalize	processed	sentiment	Rating2
0	0	767	33		Absolutely	4	1	0	Initmates	Intimate	Intimates	['absolutel	absolutely	0.633333	5
1	1	1080	34		Love this d	5	1	4	General	Dresses	Dresses	['love', 'dr	love dress	0.339583	4
2	2	1077	60	Some maj	I had such	3	0	0	General	Dresses	Dresses	['high', 'ho	high hope	0.073675	3
3	3	1049	50	My favorit	I love, love	5	1	0	General Pe	Bottoms	Pants	['love', 'lo	love love l	0.55	4
4	4	847	47	Flattering	This shirt is	5	1	6	General	Tops	Blouses	['shirt', 'fla	shirt flatte	0.512891	4
5	5	1080	49	Not for th	I love tracy	2	0	4	General	Dresses	Dresses	['love', 'tra	love tracy	0.17875	3
6	6	858	39	Cagrcoal s	I aded this	5	1	1	General Pe	Tops	Knits	['aded', 'ba	aded bask	0.13375	3
7	7	858	39	Shimmer, s	I ordered t	4	1	4	General Pe	Tops	Knits	['ordered',	ordered ca	0.171635	3
8	8	1077	24	Flattering	I love this	5	1	0	General	Dresses	Dresses	['love', 'dr	love dress	0.0025	3
9	9	1077	34	Such a fun	I'm 5"5' an	5	1	0	General	Dresses	Dresses	['lb', 'orde	lb ordered	0.2042	4
10	10	1077	53	Dress look	Dress runs	3	0	14	General	Dresses	Dresses	['dress', 'ru	dress run s	-0.09715	3
11	11	1095	39		This dress	5	1	2	General Pe	Dresses	Dresses	['dress', 'pe	dress perfe	0.25	4
12	12	1095	53	Perfect!!!	More and	5	1	2	General Pe	Dresses	Dresses	['find', 'reli	find reliant	0.244156	4
13	13	767	44	Runs big	Bought	5	1	0	Initmates	Intimate	Intimates	['bought', '	bought bla	0.192143	3
14	14	1077	50	Pretty part	This is a ni	3	1	1	General	Dresses	Dresses	['nice', 'ch	nice choic	-0.05714	3
15	15	1065	47	Nice, but n	I took thes	4	1	3	General	Bottoms	Pants	['took', 'pa	took packa	0.166587	3
16	16	1065	34	You need t	Material a	3	1	2	General	Bottoms	Pants	['material'	material c	0.134921	3
17	17	853	41	Looks grea	Took a cha	5	1	0	General	Tops	Blouses	['took', 'ch	took chan	0.227083	4
18	18	1120	32	Super cute	A flattering	5	1	0	General	Jackets	Outwear	['flattering	flattering s	0.102381	3
19	19	1077	47	Stylish and	I love the l	5	1	0	General	Dresses	Dresses	['love', 'loo	love look f	0.431818	4
20	20	847	33	Cute, crisp	If this	4	1	2	General	Tops	Blouses	['product',	product pe	0.216204	4
21	21	1080	55	I'm torn!	I'm upset b	4	1	14	General	Dresses	Dresses	['upset', 'p	upset price	0.124621	3
22	22	1077	31	Not what i	First of	2	0	7	General	Dresses	Dresses	['first', 'pul	first pullov	-0.0456	3
23	23	1077	34	Like it, but	Cute little	3	1	0	General	Dresses	Dresses	['cute', 'litt	cute little	0.269286	4
24	24	847	55	Versatile	I love this	5	1	0	General	Tops	Blouses	['love', 'shi	love shirt f	0.258333	4
25	25	697	31	Falls flat	Loved the	3	0	0	Initmates	Intimate	Lounge	['loved', 'm	loved mat	0.261508	4
26	26	949	33	Huge disap	I have bee	2	0	0	General	Tops	Sweaters	['waiting',	waiting sw	0.101818	3
27	27	1003	31	Loved, but	The colors	4	1	0	General	Bottoms	Skirts	['color', 'ex	color expe	0.17963	3
28	28	684	53	Great shirt	I have seve	5	1	2	Initmates	Intimate	Lounge	['several',	several go	0	3
29	29	4	28	Great laye	This sweat	5	1	0	General	Tops	Sweaters	['sweater',	sweater co	0.282917	4

Fig .2: Dataset

4.RESULTS AND DISCUSSION

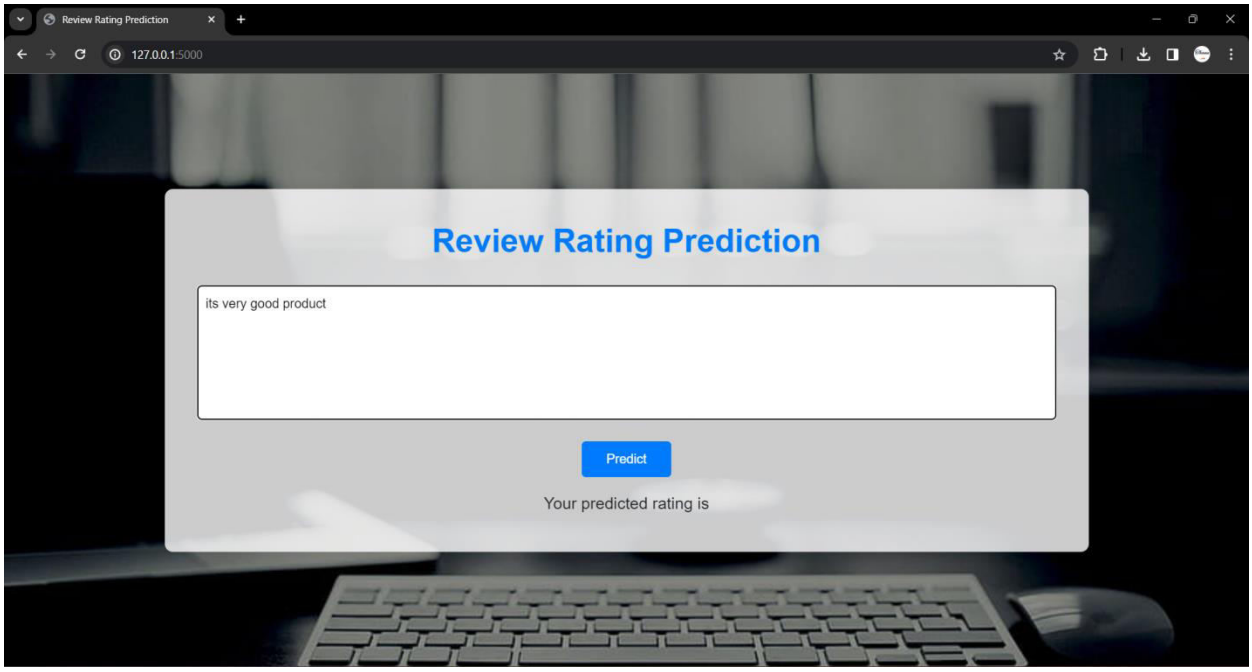


Figure 3:

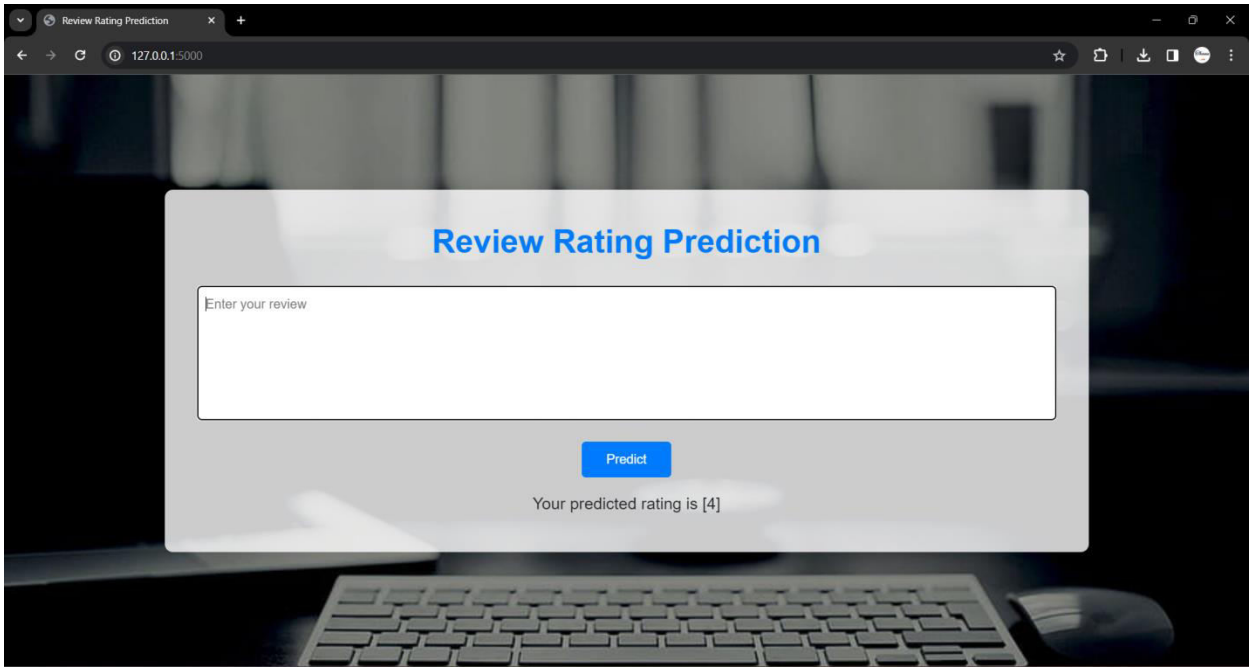


Figure 4:

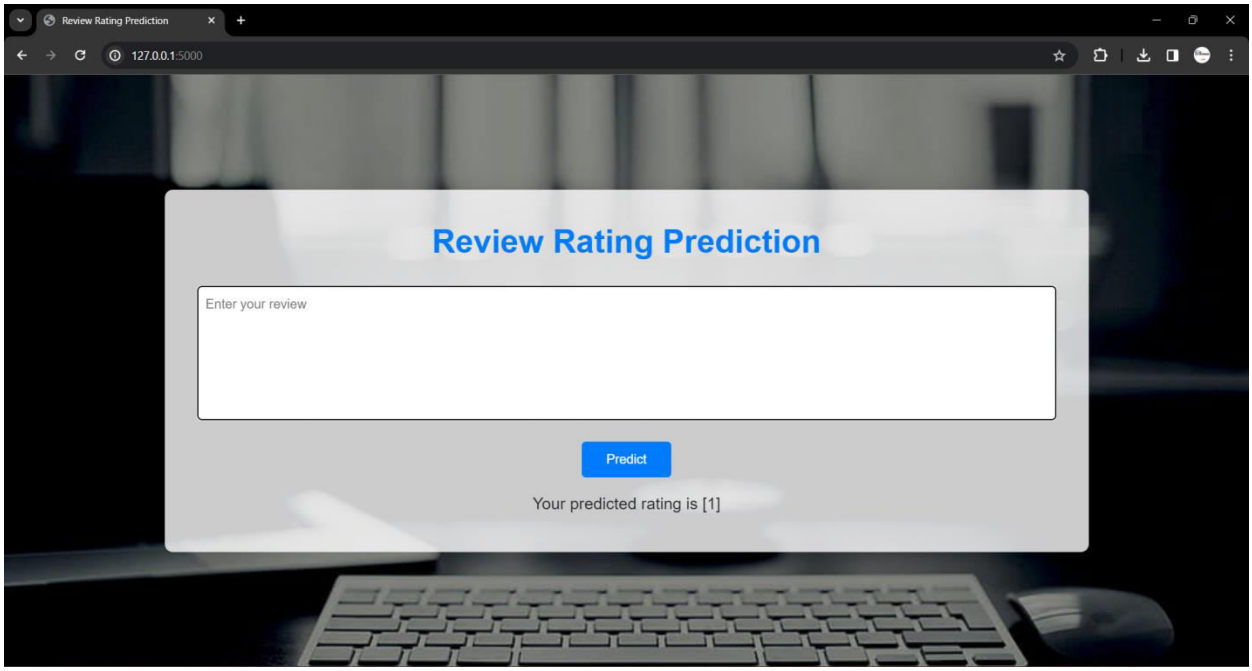


Fig 5:

5.CONCLUSION

- This research used two machine learning algorithms: Random Forest and XGBoost to classify customer review texts.

- We concentrated on online reviews features such as class name, review texts.

Moreover, we compared our results with previous research and our results indicated that XGBoost was the preferred classifier.

- Previous studies such as (Agarap and Grafilon; 2018) used Bidirectional Recurrent Neural Network to do sentiment analysis, our research used two machine learning algorithms and was able to achieve better result, achieving more than 94% accuracy for all the algorithms.

- Agarap and Grafilon; 2018 and research done by Ireland study used the same data set, however, when explored data these used different methods.

- Agarap and Grafilon; 2018 used NLTK to do sentiment analysis. Compared with Agarap and Grafilon; 2018,Ireland researchers used a heat map and data exploration to do feature selection and then built models.

- f it is possible, they should focus on unbalanced and balanced data set to explore the sentiment analysis.

- In our project we have changed the complete dataset. We used a real time amazon
dataset which is available in AWS.

- We completely changed our dataset to avoid unbalanced dataset.

- We have applied only two algorithms on our dataset and achieved accuracy of more than 90%.

- We used Random Forest and achieved 91% and used XGBoost and achieved accuracy of 99%.

- We used lemmitizer, Textblob in NLP to normalize the review text.

- We have used polarity to predict the rating score.

REFERENCES

[1] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro,

Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian

Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, RafalJozefowicz,

Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mane, Rajat Monga, Sherry

Moore, Derek Murray, Chris Olah,

Mike Schuster 2015 https://www.tensorflow.org/

Software available from tensorflow.org.

[2] Nick Brooks. 2018. Guided Numeric and Text Exploration E-Commerce. (2018).

https://www.kaggle.com/nicapotato/ guided-numeric-and-text-exploration-e-commerce.

[3] Nick Brooks. 2018. Women's E-Commerce Clothing Reviews. (2018). https:

//www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

[4] François Chollet et al. 2015. Keras. https://github.com/keras-team/keras. (2015).

[5] Ian Goodfellow, YoshuaBengio, and Aaron Courville. 2016. Deep Learning. MIT Press.

http://www.deeplearningbook.org.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural

computation 9, 8 (1997), 1735–1780.

[7] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. Computing In Science &

Engineering 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[8] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In

Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching

Natural Language Processing and Computational Linguistics. Philadelphia: Association

for Computational Linguistics.