Enhancing Trust with AI: Product review analysis and segregation system

Dr K Siva Krishna
Associate Professor
MLR Institution of Technology
Hyderabad, India
kshivakrishna@mlrit.ac.in

Dr Sai Prasad Kashi
HOD CSE-AIML
MLR Institution of Technology
Hyderabad, India
saiprasad.kashi@mlrinstitutions.ac.in

Yadagiri Rishyendra Kumar
MLR Institute of Technology
Hyderabad, India
rishyendra09@gmail.com

Guniguntla Rohith
MLR Institute of Technology
Hyderabad, India
rohithgr30@gmail.com

Tammysetty Tarakesh
MLR Institute of Technology
Hyderabad, India
tarakeshtammysetty1@gmail.com

Anurag Thalari
MLR Institute of Technology
Hyderabad, India
anuragreigns366@gmail.com

ABSTRACT

The establishment of trust in the purchasing domain is largely dependent on product reviews. However it can be difficult to sort through the huge number of reviews to discover the ideal product. Our study Enhancing Trust with AI: Product Review Analysis and Segregation System, aimed to address this problem by providing consumers with the information they needed to make correct decisions. We investigated the complex ways that online reviews affect consumers' decisions to buy. Our objective was to simplify the review process for users by classifying reviews as neutral negative or favorable. Our goal is to simplify the huge amount of information available so that consumers can more easily locate relevant insights. Furthermore in order to preserve the accuracy of the data provided we used innovative technology to recognize and eliminate fraudulent reviews. The aim of doing that was to guarantee that customers could rely on the legitimacy of the reviews they read.

Keywords

Sentiment analysis, Fake reviews, NLP Techniques, Visualisation Techniques

CCS CONCEPTS

Online product reviews. Review detection using Machine Learning and AI. Sentiment analysis.

1 INTRODUCTION

The internet's explosive growth has changed many aspects of daily living including work settings information intake and time management. Customers may easily rely on product reviews while making decisions[3][13] in the world of online retail. It is evident that customers in the online retail space depend heavily on product reviews to help them make decisions. While businesses can benefit financially greatly from positive evaluations[12] the increasing number of fake reviews reduces customer confidence. Differentiating between genuine and fraudulent reviews[11] is extremely difficult particularly when trying to separate information that is generated by bots from that that is created by humans. A range of measures including language subtleties and review structure are important indicators for identification. This effort finds powerful support in machine learning which encompasses supervised semi-supervised[8] and unsupervised approaches. Support vector machines and Naive Bayes classifiers[13] are examples of supervised learning algorithms can provide strong frameworks for identifying trends in review data. These complex mathematically supported algorithms enable computers to make well informed decisions which increases confidence and trust in ecommerce platforms. Machine learning is which encompasses supervised semi-supervised and unsupervised[6] techniques proves to be an effective partner in this undertaking. Support vector machines Naive Bayes classifiers[4] are two examples of supervised learning algorithms that provide strong frameworks for identifying trends in review data. These al-gorithms is which are supported by mathematical structures enable computers to make defensible decisions will increase confidence inecommerce platforms.

- Making Informed Decisions: By getting well-organized and analyzed online product reviews consumers can gain valuable insights and sentiments from other users allowing them to make an informed purchasing decision.
- Time Efficiency: Efficiently checking through online reviews saves customers time by processing the decision-making process. They can quickly get a product's pros and cons without being troubled by excessive information.
- Trust and Loyalty: Improving the accessibility and reliability of online reviews provides trust between consumers and brands. This trust gives loyalty and positive relationships. Businesses that design transparency and offer helpful review analysis demonstrate their dedication to customer satisfaction.

In response to these challenges, this paper presents a novel AI-driven framework that harnesses the power of sentiment analysis[5] and natural language processing (NLP)[9]. By categorizing reviews and identifying fraudulent content, this system seeks to fortify consumer trust and confidence in online transactions. Through a comprehensive analysis of review data, leveraging advanced data mining approaches and visualization techniques[17], our framework aims to enhance user convenience and comprehension, thereby fostering a more transparent and trustworthy online retail ecosystem.

1.1 Overview Sentiment Analysis

Imagine being able to quickly and easily understand the core beliefs and feelings that are communicated in the written word. That's exactly what sentiment analysis[18] provides potent instrument that explores the domain of natural language processing to identify the tone or sentiment ingrained in textual information. Sentiment analysis can reveal the attitudes, opinions, and feelings that people are expressed in a variety contexts that include product reviews, social media posts and customer feedback [1].

Sentiment analysis is the process of analyzing text and classifying it as positive, negative, or neutral according to the underlying sentiment using computational methods and algorithms [7]. Sentiment analysis has completely changed how organizations and people perceive and react to textual data by combining artificial intelligence and machine learning[9].

Figure 1: Labeling reviews according to their sentiments

1.2 Support Vector Machine

Support vector machine(SVM) is the most important tool in this process Specifically used for sentiment analysis[14] where the model is trained to classify reviews. The process of converting text data intonumerical vectors using TF-IDF is implemented with the help of feature extraction performed by SVM. Features such as review textand reviewer history are used to distinguish between fake and original reviews. SVM helped to effectively classify reviews[19] extract relevant relevant features and provide users with trustworthy and informative insights for making purchasing decisions. SVM offers techniques like class weights or using different evaluation metrics to handle imbalanced data.

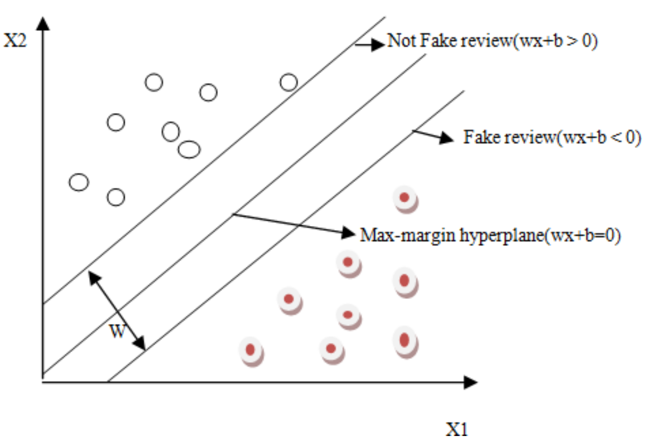

Figure 2: Support vector Machine hyperplane

2 EXISITING PROBLEM

Reviews show a significant impact on people's decisions about product purchases. The primary challenge is in manually identifying fraudulent reviews[16] through reviewers' ability to craft a fake review that matches the real one. Among the millions of evaluations, it is challenging to identify authentic reviews. They discovered several duplicates and almost duplicate reviews in the case of 5.8 million reviews and 2.14 million reviewers on Amazon.com[10] since there is a lack of trustworthy data that has been classified as false or genuine for training. Analyzing posted reviews, each user hasunique strategies and patterns, and it is very difficult to categorize whether posted reviews are bot-generated or human-generated. Most of the fake review detections[15] are manually detected, which consumes and requires additional human power to analyze, so it must switch to dynamic detection of fake reviews[16]. Most of the e-commerce websites hide purchase histories; they disablegetting user information and IP addresses for the majority of thereviews[20].

A single internet review has a big impact. Customers in the digital business research products online before picking which ones to buy by reading a few reviews. Data scarcity is the primary issue, and both language and behavioral aspects are needed. A behavioral technique is suggested to identify review spammers[10] that attempt to influence ratings on certain target products, and the problem is simple but difficult to answer: to find a fake or ambiguous category, one has to go through each review and mark it as such. This procedure must be handled carefully. To fix this,an ML model handling the review section can be trained to flag a particular review. A signal that counts submitted reviews could be used to identify singleton spammers[10] without revealing their presence.

3 IMPLEMENTATION

The solution to the current problem in the E-commerce trust can be replicated with the 4step proposed architecture that provides easy access to the genius reviews as well as the bot generated reviews

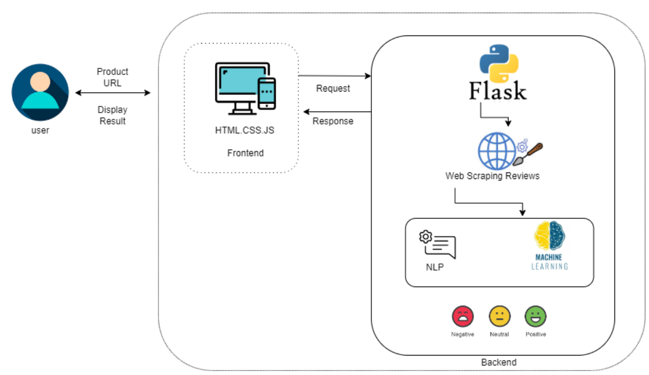

Figure 3: Proposed Architecture

- **User:** The user initiates the web scraping process by providing a URL(Product link) specifying the target source for data extraction.

- **Frontend:** Users interact with this component a graphical interface to input the URL to trigger the web scraping operation and to see the final results.
- **Requests:** The user's input which includes the URL is transmitted to the web scraping server prompting the data extraction process.
- **Backend:** This server-side component houses the logic responsible for executing web scraping tasks including sending requests, fetching web pages, parsing HTML content with BeautifulSoup, and extracting relevant data.
- **Response**Upon successful data extraction using BeautifulSoup a web scraping server delivers the extracted data back to the user. the extracted data undergoes further processing through machine learning (SVM) and NLP (NLTK VADER) algorithms to analyze sentiment and detect computer-generated reviews.

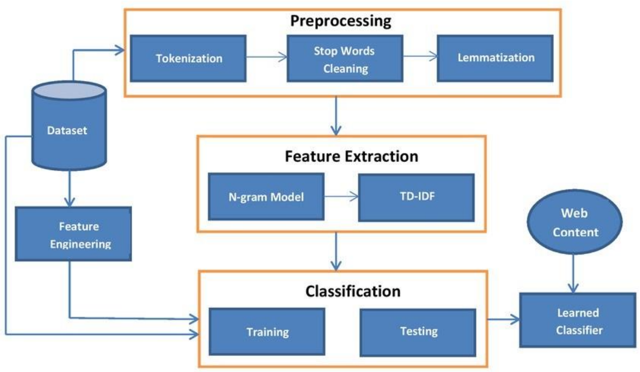

Figure 4: Flowchart outlining the proposed solution overview

3.1 Machine Learning and NLP Integration

- **Sentiment Analysis (NLTK VADER):** NLTK VADER a sentiment analysis tool evaluates the sentiment expressed in the reviews categorizing them as positive, negative, or neutral.
- **Computer Generated Detection (SVM):** SVM, a machine learning algorithm is utilized to classify reviews as either human-written or computer-generated enhancing authenticity assessment.

4 SOLUTION TO THE PROBLEM

4.1 Implementation of Web Scraping

```
# Import packages
Import requests
Import pandas as pd
Import BeautifulSoup from bs4


# Function check_csv_exists
Function check_csv_exists(file_name):
    Return True if file_name exists,
    else False
# Function amazon
Function amazon(url):
    Define headers
    Modify url
    Extract product title
    Construct csv_file_name

    If CSV file exists:
        Read CSV file and return
        Data Frame with Title
    Else:
        Remove unnecessary parts from url
        Construct reviews_url
        Set len_page to 1

        # Function reviewsHtml
        Function reviewsHtml(url, len_page):
            Initialize empty list soups
            Loop from 1 to len_page:
                Send GET request and
         parse response using BeautifulSoup
                Append soup to soups list
            Return soups

        # Function getReviews
        Function getReviews(html_data):
            Initialize empty list data_dicts
            Loop through each review box:
                Extract stars, title,
                and description
                Create dictionary with
                extracted data
            Append dictionary to data_dicts
            Return data_dicts

        # Scraping reviews
Call reviewsHtml and getReviews
to gather review data
    Create DataFrame df_reviews from
    review data

    Return df_reviews and Title
```

The algorithm for extracting product reviews from Amazon is described in pseudocode. Importing the required libraries and creating functions is where everything starts. The programs main function is the amazon function however the check csv exists function checks to see if the CSV file exists. It accepts a URL as input gets it ready for scraping, and determines whether the previous reviews CSV file is there. If not it uses requests to scrape review pages, from which BeautifulSoup pulls review details and builds a DataFrame. Helper methods like reviewsHtml and getReviews are used in this procedure to handle data extraction and HTML parsing. Ultimately the DataFrame with reviews and the product

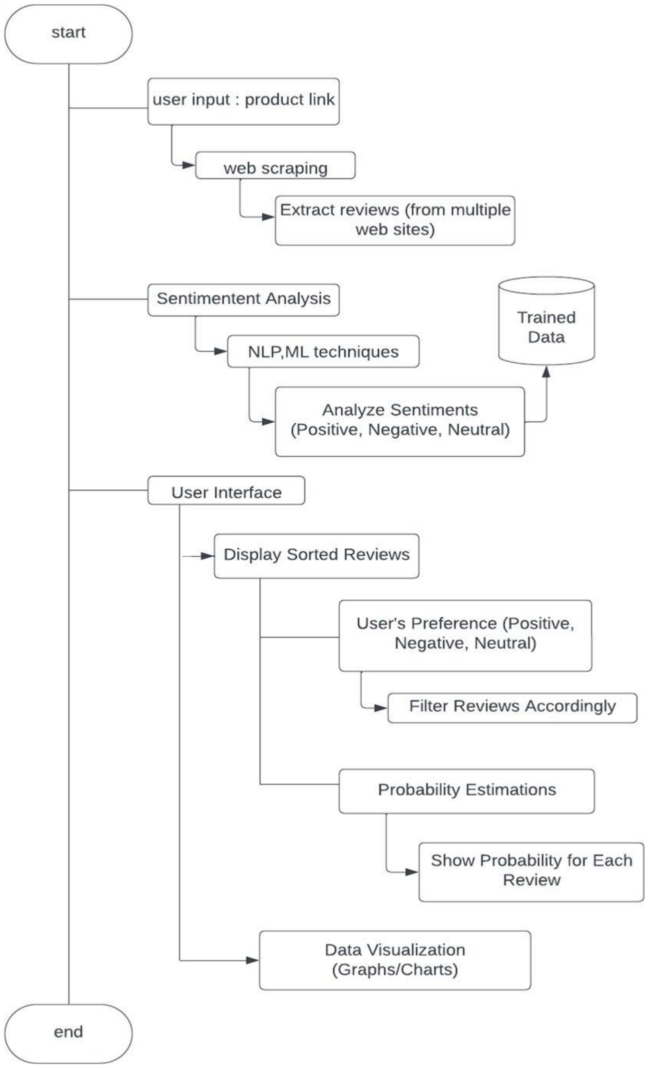

Figure 5: Detailed flowchart of proposed solution

title is returned. This pseudocode offers an organised method for putting the review scraping feature into practice.

4.2 Implementation of Sentiment Analysis

```
# Import SentimentIntensityAnalyzer from
NLTK lib
From nltk . sentiment
import SentimentIntensityAnalyzer
# Function to analyze the sentiments of text
Function analyze_sentiments ( text ):
    Initialize SentimentIntensityAnalyzer
    Get polarity scores for the text
    Extract compound , negative , positive ,
    and neutral scores from the polarity
    scores
```

```
    # Determine sentiment based on
    compound score
    If compound score is less than 0:
        Return 'Negative'
    Else:
        If positive score is greater than or
        equal to neutral score:
            Return 'Positive'
        Else:
            Return 'Neutral'
```

The algorithm that was created to use the SentimentIntensityAnalyzer from the NLTK library to determine the sentiment of the provided text is described in the pseudocode. The SentimentIntensityAnalyzer module from NLTK is imported and the function analyse sentiment(text) which accepts text as input, is defined. The function extracts the compound negative positive and neutral scores as well as initialises the SentimentIntensityAnalyzer and returns the text polarity scores. The function returns Negative if the compound score is negative. It yields Positive otherwise if the positive score is more than or equal to the neutral value.

4.3 Clustering by using SVM Algorithm

```
# Import necessary libraries
Import pandas as pd
From sklearn . pipeline import Pipeline
import CountVectorizer , TfidfTransformer
From sklearn . svm import SVC
Import string
From sklearn . model_selection
import train_test_split
From sklearn . feature_extraction . text
From sklearn . metrics import accuracy_score
Import nltk
From joblib import dump , load
From nltk . corpus import stopwords
# Function to preprocess text
Function convertmyTxt ( rv ):
    Remove punctuation from the text
    Tokenize the text
    Remove stopwords from the tokenized text
    Return the preprocessed text

# Function to train SVM model
Function train_svm_model ( ):
    Read dataset from CSV file into a
    DataFrame
    Drop rows with missing values
    Add a column for text length
    Spliting data into training and testing
    sets# Define pipeline
    Define pipeline with CountVectorizer ,
    TfidfTransformer , and SVC
    Download NLTK stopwords
    Fit pipeline on training data
```

```
    Predict labels for test data
    Calculate accuracy score
    Print accuracy
    Save trained pipeline to a file

    Return trained pipeline

# Function to load SVM model
Function load_svm_model():
    Load trained pipeline from file
    Return loaded pipeline


# Main function
If _name_ == "_main_":
    Call train_svm_model function and
    store the trained pipeline
```

The code above aims to train and apply sentiment analysis to a Support Vector Machine (SVM) model. Importing the required libraries for training and evaluating data processing models is the first step. Stopwords and punctuation are eliminated from text data prior to preprocessing using the convertmyTxt function. After reading a dataset and dividing it into train and testing sets, the trained svm model function builds a pipeline that includes text vectorization TF-IDF transformation and SVM classification. The model is trained its accuracy is checked, and it is then saved to a file. A pre-trained model is loaded from the file using the load svm model function. Lastly, the code trains the SVM model in the main block and saves the trained pipeline for later use.

4.4 Implementation of Flask

This pseudocode outlines the structure of Flask application with three routes: '/' for the home page, '/new page' for processing user input and displaying results, and '/topics detail' for another page. The main function of the application is to render HTML templates based on user requests and display relevant data.

```
import Flask, redirect,
render_template, request, url_for
from main Import main fun as mf
import pandas as pd
# Initializing Flask application
app = Flask(__name_)
#Define routes
@app.route('/')
def index():
    return render_template('index.html')

@app.route('/_new_page',

method=['POST', 'GET'])
def new_page():
    input_text=request.form['keyword']
    csv_path, Title = mf(input_text)
    df = pd.read_csv(csv_path)
    Return render_template
    ('new_page.html',
    submitted_text=input_text,
    csv_data=df.to_html(), Name=Title)

@app.route('/topics_detail')
def topics_detail():
    Return render_template
    ('topics_detail.html')

# Run Flask application now
if _name_ == '_main_':
    app.run(debug=True)
```

Finding real reviews among the millions of evaluations is difficult. Due to a dearth of reliable data that has been identified as authentic or fraudulent for training purposes, a significant number of duplicate and almost duplicate reviews were found in the instance of 4.8 million reviews and 1.14 million reviewers on Amazon.com. Because each user has unique strategies and habits it might be difficult to discern between bot-generated and human-generated content when examining posted evaluations. Most false review detections are done manually. Themajority of false review detections are made by hand.

5 RESULT ANALYSIS

5.1 Analysis Metrics for Sentiment Analysis

5.1.1 Precision and Recall Curve. A basic measure called the ratio of Precision and Recall curve illustrates how varying classification thresholds impact the trade-off between recall and accuracy. Recall is the proportion of accurately predicted positive or negative reviews among all actual positive or negative reviews, while precision is the proportion of correctly predicted positive or negative reviews among all predicted positive or negative reviews. Plotting distinct Precision and Recall curves for neutral, negative, and positive reviews allows us to see how well the model can identify each sentiment category.

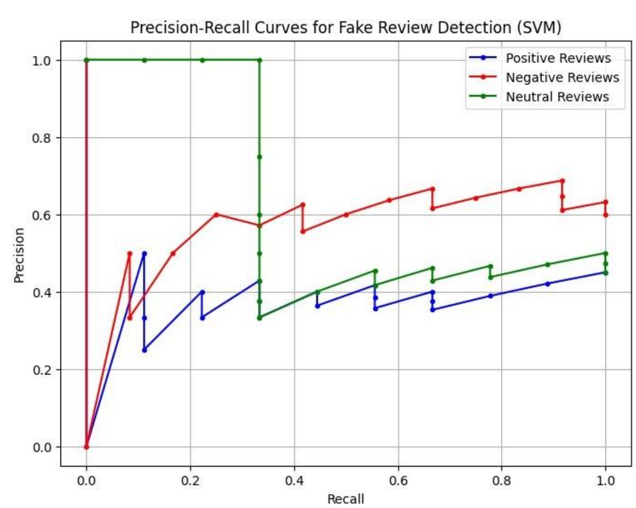

Figure 6: Image of Precision and Recall Curve

5.1.2 Confusion Matrix Confusion Matrix Another important analysis metric is the confusion matrix. By showing the counts of true positive false positive,true negative, and false negative predictions for each sentiment category (positive, negative, and neutral), Confusion Matrix offers a thorough summary of the model performance.. We can determine the particular advantages and dis- advantages of the sentiment analysis model in accurately classifying each sentiment category by looking at the Confusion Matrix for positive, negative, and neutral reviews separately. .

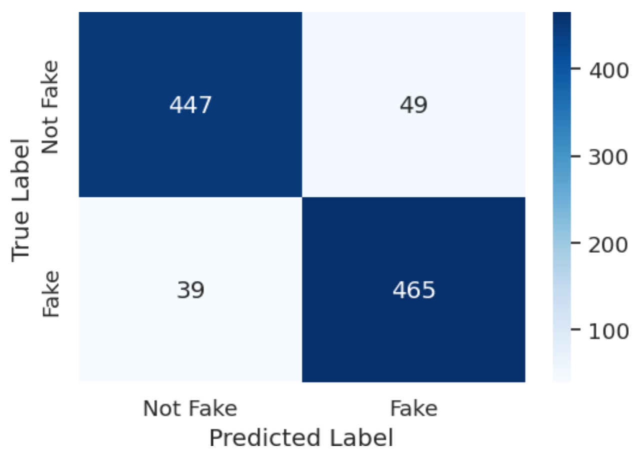

Figure 7: Confusion matrix for Sentiment analysis

5.2 Analysis Metrics for Identifying Fake reviews using SVM

The confusion matrix for fake review detection using Support Vector Machine (SVM) was constructed based on a dataset of 100 reviews in total. Among these 90 reviews were human-generated representing genuine reviews while the remaining 10 reviews were computer-generated representing fake reviews. The SVM models performance was evaluated using this dataset, resulting in a confusion matrix that illustrates its classification outcomes.

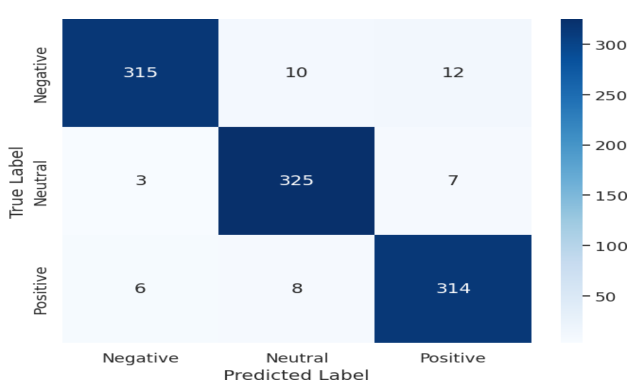

Figure 8: Confusion matrix for SVM

Aspect	Before Implementation	After Implementation
Time spent on Reviews	Varied, with consumers spending an average of 20-30 minutes	Streamlined process, with average time spent reduced by 50 percentnow averaging 10-15 minutes per consumer
Sentiment segregation	Manual Review analysis without segregation by sentiment	Segregation of reviews basedon sentiment, enabling efficient focus on relevant content, with a 75percent reduction in time spent searching for sentiment cues.
CG Reviews Detection	Limited ability to discern between genuine and computer-generated reviews	Enhanced capability to identify and filter out fake reviews, resulting in 90per decrease in exposure to potentially misleading content.
Overall Efficiency	Varied efficiency in a review analysis, with potential for bais and time wastage.	Streamlined and efficient review analysis process, resulting in an overall 60per increase in efficiency and time savings for consumers

Table 1: Comparison table to evaluate aspects before and after implementation

5.3 Labelling and Visualizing the Resulted Data

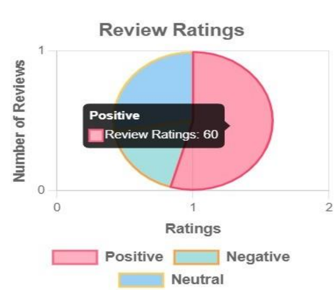

Figure 9: Visualizing data patterns through pie charts

We employed Support Vector Machines (SVM) to detect fake reviews in a dataset. Reviews were analyzed and labeled as either CG (Computer Generated) or OR (Original Reviews). Visualizing the resultant data using a pie chart enables to grasp the summary of the sentiment distribution efficiently. The pie chart provided a clear representation of the proportion of positive, negative, and neutral reviews

Figure 10: Output

6 CONCLUSION AND FUTURE SCOPE

Our research has demonstrated how technology can revolutionize customer decision-making in the e-commerce space. Our model accelerated review aggregation, sentiment analysis, and authenticity assessment by combining web crawling, machine learning, and natural language processing. This approach provides visitors the ability to confidently and quickly browse through a variety of reviews. The advantages are obvious: better understanding, more trust, and well-informed decisions. Future developments in machine learning for more complex sentiment analysis and UI improvements should improve user experiences. This study highlights how technology is changing the way that e-commerce reviews are analyzed, providing insight into a future in which customers have even more power and confidence when making purchases online.

REFERENCES

[1] Aljoharah Almjawel, Sahar Bayoumi, Dalal Alshehri, Soroor Alzahrani, and Munirah Alotaibi. 2019. Sentiment analysis and visualization of amazon books' reviews. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, 1–6.

[2] RUTUJA B ARDAK and GIRISH S THAKARE. 2019. A REVIEW ON FAKE PRODUCT REVIEW DETECTION AND REMOVAL TECHNIQUES. *IJCSE International Journal of Computer Sciences and Engineering* (2019).

[3] Tri Astuti and Irnawati Pratika. 2019. Product Review Sentiment Analysis by Artificial Neural Network Algorithm. *International Journal of Informatics and Information Systems* 2, 2 (2019), 61–66.

[4] Daria Beresneva. 2016. Computer-generated text detection using machine learning: A systematic review. In *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*. Springer, 421–426.

[5] Chhaya Chauhan and Smriti Sehgal. 2017. Sentiment analysis on product reviews. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 26–31.

[6] N Deshai and B Bhaskara Rao. 2023. Deep learning hybrid approaches to detect fake reviews and ratings. *Journal of Scientific and Industrial Research* (2023), 120–127.

[7] Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah. 2018. Senti- ment analysis on large scale Amazon product reviews. In *2018 IEEE international conference on innovative research and development (ICIRD)*. IEEE, 1–6.

[8] Rakibul Hassan and Md Rabiul Islam. 2020. A supervised machine learning approach to detect fake online reviews. In *2020 23rd international conference on computer and information technology (ICCIT)*. IEEE, 1–6.

[9] T Karthikeyan, Karthik Sekaran, D Ranjith, JM Balajee, et al. 2019. Personalized content extraction and text classification using effective web scraping techniques. *International Journal of Web Portals (IJWP)* 11, 2 (2019), 41–52.

[10] S Uma Maheswari and SS Dhenakaran. 2019. Detection of Fake and Genuine Reviews with Hybridization of Fuzzy and Neural Networks Techniques. *IJCSE International Journal of Computer Sciences and Engineering* (2019).

[11] Mr. Navjyotsinh, Chirag visani 2017. *A study on different machine learning techniques for spam reviews detection*.

[12] Rohit Yadav Datta Supekar Prof. P.S. Gaikwad, Kaushal Parmar. 2021. Implementation of web scraping for e-commerce website. *International Journal of Computer Sciences and Engineering* (2021).

[13] Koli. S. 2021. Sentiment Analysis with Machine Learning Techniques and Improved J48 Decision Tree Technique. *International Journal of Computer Sciences and Engineering* (2021).

[14] Raheesa Safrin, KR Sharmila, TS Shri Subangi, and EA Vimal. 2017. Sentiment analysis on online product review. *International Research Journal of Engineering and Technology (IRJET)* 4, 04 (2017).

[15] Yin Shuqin and Feng Jing. 2019. Fake reviews detection based on text feature and behavior feature. In *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2007–2012.

[16] Anusha Sinha, Nishant Arora, Shipra Singh, Mohita Cheema, and Akthar Nazir. 2018. Fake product review monitoring using opinion mining. *International Journal of Pure and Applied Mathematics* 119, 12 (2018), 13203–13209.

[17] Chengai Sun, Qiaolin Du, Gang Tian, et al. 2016. Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering* 2016 (2016).

[18] G Vinodhini and RM Chandrasekaran. 2012. Sentiment analysis and opinion mining: a survey. *International Journal* 2, 6 (2012), 282–292.

[19] Wenyuan Zhao. 2020. Classification of Customer Reviews on E-commerce Platforms Based on Naive Bayesian Algorithm and Support Vector Machine. In *Journal of Physics: Conference Series*, Vol. 1678. IOP Publishing, 012081.

[20] Bichen Zheng, Keith Thompson, Sarah S Lam, Sang Won Yoon, and Nathan Gnanasambandam. 2013. Customers' behavior prediction using artificial neural network. In *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), 700.