Sentiment Analysis of Product Based Review Using Machine Learning

Minor Project-I Report

Submitted for the partial fulfillment of the degree of

Bachelor of Technology

In

Mathematics and Computing

Submitted By

Pranshoo Patel

0901MC221054

UNDER THE SUPERVISION AND GUIDANCE OF

Dr. Vijay Shankar Sharma

Assistant professor

Department of Engineering Mathematics & Computing

माधव प्रौद्योगिकी एवं विज्ञान संस्थान, ग्वालियर (म.प्र.), भारत

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR (M.P.), INDIA

Deemed University

(Declared under Distinct Category by Ministry of Education, Government of India)

NAAC ACCREDITED WITH A++ GRADE

July-December 2024

Declaration by the Candidate

I hereby declare that the work entitled **"Sentiment Analysis of Product Based Review Using Machine Learning"** is my work, conducted under the supervision of **Dr. Vijay Shankar Sharma, Assistant Professor,** during the session July-Dec 2024. The report submitted by me is a record of bonafide work carried out by me.

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Pranshoo Patel
0901MC221054

Date: 02/12/24
Place: Gwalior

This is to certify that the above statement made by the candidates is correct to the best of my knowledge and belief.

Guided By: Sharma 2/12/24

Dr. Vijay S. Sharma
Assistant Professor
Engineering Mathematics & Computing
MITS-DU, Gwalior

Departmental Project Coordinators

Atul 02/12/24
Dr. Atul Kumar Ray
Assistant Professor
Engineering Mathematics & Computing,MITS-DU, Gwalior

for Atul 02/12/24
Dr. Minakshi Dahiya
Assistant Professor
Engineering Mathematics & Computing, MITS-DU, Gwalior

Approved by HoD

02/12/24
Dr. D. K. Jain
Professor and Head
Engineering Mathematics & Computing, MITS-DU, Gwalior

PLAGIARISM CHECK CERTIFICATE

This is to certify that I/we, a student of B.Tech. in **Engineering Mathematics And Computing** have checked my complete report entitled **"Sentiment Analysis of Product Based Review UsingMachine Learning"** for similarity/plagiarism using the "Turnitin" software available in the institute.

This is to certify that the similarity in my report is found to be …… which is within the specified limit (30%).

The full plagiarism report along with the summary is enclosed.

Pranshoo Patel

0901MC221054

Checked & Approved By:

Dr. Barkha Tiwari
Assistant Professor
Engineering Mathematics & Computing
MITS-DU, Gwalior

ABSTRACT

Sentiment analysis, also known as opinion mining, is an important area of research in text mining and computational linguistics. It involves extracting and analyzing opinions or sentiments from textual data. This document discusses the methodologies and approaches employed in a project called sentiment classification of Amazon product reviews. It classifies reviews into either positive or negative sentiments by the sentiments portrayed in the text.

The entire data preparation process starting from data scraping and preprocessing along with the feature extraction part is followed by applying machine learning models such as Naïve Bayes, Logistic Regression, and Support Vector Machine for sentiment classification. Techniques used to improve the accuracy of classification include part-of-speech tagging, negation phrase identification, and bi-gram analysis. This implies that performance evaluation metrics are F1 Score, Accuracy, Precision, Recall, and ROC AUC. The results indicate that the trained models have the ability to predict correct sentiment polarity correctly and indicate the preprocessing and feature selection steps as vital towards obtaining improvements in classifying by performance-related aspects.

The potential for future progress in sentiment analysis is highlighted at the end of the study, pointing toward its application in varied fields-related customer response analysis, social media monitoring, and other market trends. Such contributions further the ever-expanding domain of sentiment analysis, giving light to potential further development concerning the performance of models and lines of potential exploration in the future.

ACKNOWLEDGEMENT

I extend my sincere gratitude to the esteemed Vice Chancellor of Madhav Institute of Science and Technology, Deemed University, **Dr. R. K. Pandit**, and the respected Dean Faculty of Engineering & Technology, **Dr. Manjaree Pandit**, for their invaluable support and guidance throughout this project.

Further, I thank my project Supervisor **Dr. Vijay Shankar Sharma**, Assistant Professor, from the Department of Engineering Mathematics & Computing. Their expert guidance, continuous support, and encouragement in my journey, help me to overcome challenges and achieve my target.

I am also deeply grateful to the esteemed faculties of the Department of Engineering Mathematics & Computing for their invaluable feedback, insightful suggestions, and continuous encouragement, which have played a pivotal role in shaping my project.

To my friends, I extend my deepest thanks for your constant support, encouragement, to-the-point comments on the draft, subject matter knowledge and maturity in helping me realise my blind spots remained a source of motivation to stay focused and at it.

Pranshoo Patel

0901MC221054

CONTENT

Table of Contents

ACRONYMS

SA- Sentiment Analysis

SO- Sentimental Orientation

NB - Naïve Bayes

SVM - Support Vector Machine

ROC AUC - Receiver Operating Characteristic Area Under Curve

TSI- Total Sentiment Index

ML -Machine Learning

NLTK - Natural Language Toolkit

JSON - JavaScript Object Notation

F1- F1 Score (Harmonic Mean of Precision and Recall)

LIST OF FIGURES

Page No.

CHAPTER 1: INTRODUCTION

1.1 Overview of Sentiment Analysis

Sentiment analysis, otherwise known as opinion mining, is a critical field in the area of natural language processing (NLP)[9] and text mining[1]. It is the identification, extraction, and analysis of opinions, sentiments, and emotional tones within text. The analysis provides information on whether the sentiment in the text is negative, positive, or neutral, hence making it a very important tool for determining consumer opinion, tracking public sentiment, and hence informing business strategies.

The growth in online platforms such as social media, e-commerce websites, forums, and blogs of unstructured text data has been so immense that one can only imagine such growth. Sites such as Amazon have millions of product reviews, where the customers give their free expression of thought. With such enormous data, it is impossible to analyze manually. This is exactly where sentiment analysis comes in, which would use algorithms and models to extract and classify sentiments automatically.

1.2 Value of Sentiment Analysis for E-commerce

Sentiment analysis has become a foundational element for e-commerce ventures:

- **Customer Feedback Analysis** : Understand the customer experience via product reviews
- **Improvement of Services** : Identifying the shortcomings in products or services
- **Reputation Management :** Public perception of brands
- **Market Trends Analysis** : Understanding emerging trends based on customer sentiment.

For instance, the perception of a review such as, "The phone's battery life is excellent, but the camera is mediocre," calls for both positive and negative sentiments to prevail in a single review. These subtle findings help businesses identify exactly which product attributes need improvement.

1.3 Problems in Sentiment Analysis

- **Ambiguity and Subjectivity:** Human language is inherently ambiguous. Sentiments can be variable based upon context, tone, and individual perception.

  Example: The sentence "The product is too good to be true" seems positive but has an undertone of skepticism.

- **Negation and Double Negatives:** The phrase "not bad" or "not terrible" is semantically positive but would require very sophisticated models to interpret appropriately.
- **Context Sensitivity:** Any given word may change meaning based on the context, like "hot" meaning fabulous for fashion, but hot weather isn't fabulous.
- **Unstructured Data:** Most online reviews are unstructured, full of grammatical mistakes, emojis, abbreviations, and slang, making them complex to analyze.
- **Scalability:** Analysis involving millions of Amazon reviews demands computational power.

CHAPTER 2: LITERATURE SURVEY

2.1) Evolution of Sentiment Analysis

The Sentiment Analysis process has dramatically changed from lexicon-based techniques to more advanced machine learning and deep learning techniques. In the early stages of research, dictionaries or rule-based systems were primarily used for classification purposes, but with increased computational power and availability of data, more dynamic techniques emerged.

2.2) Important Contributions:

2.1.1) Feature Extraction Techniques[6,2]:

- Pang and Lee (2004): Presented methods for filtering subjective sentences using the minimum-cut approach, greatly improving the quality of sentiment classification.
- Gann et al. (2013): Introduced the Token Sentiment Index (TSI) that measures word's score based on frequency of occurrences in positive and negative reviews, improving token-level sentiment detection.

2.2.2) Classification Models:

- **Naïve Bayes (NB)[**6,4]: It is a probabilistic model that has been used for text classification purposes because it is simple in nature and efficient.
- **Logistic Regression (LR**)[9,7]: It produced highly accurate results in binary sentiment classification.
- **Support Vector Machines (SVM)**[10,9]: Used for linear and nonlinear data classification, offering high accuracy in sentiment analysis.

2.3.3) Negation Handling[9]:

- Advanced algorithms identify negation phrases, such as "not good" or "never exciting,"to ensure accurate polarity detection.

2.4.4) Applications:

- Dave et al. (2003)[6]: Demonstrated sentiment analysis's potential in product reviews.

- Zhu et al. (2011)[7]: Highlighted aspect-based opinion mining for customer feedback, emphasizing the importance of identifying sentiments associated with specific product attributes.

2.3) Challenges Addressed in Literature

Managing Ambiguity Researchers developed techniques such as part-of-speech (POS) tagging, which filter out unnecessary words in the phrases leaving the sentiment-bearing phrases. Context Sensitivity Bi-gram and n-gram models helped capture context since they analyze sequences of words rather than individual words themselves.

Scalability: Techniques for big data sentiment analysis, such as parallel processing and distributed systems, were developed to manage huge Amazon review kind of datasets.

CHAPTER 3: OBJECTIVE

3.1) Primary Objective

The primary purpose of this project is the development of an automated sentiment analysis model[3]. This will classify reviews on Amazon products as positive or negative. This will help businesses and researchers acquire insightful knowledge of customer opinions.

3.2) Detailed Objectives

3.2.1) Data Collection and Preparation:

- Gather and sort product review data from Amazon.
- preprocess collected data to remove unimportant or noisy content.

3.2.2) Sentiment analysis:

- Perform document-level sentiment analysis and make an overall judgment of the sentiments for each review.
- Apply richer text preprocessing techniques, including POS tagging[10], bi-gram analysis[9] to account for more subtle sentiments.

3.2.3) Classification of Reviews:

- Design and test various machine learning models: Naïve Bayes, Logistic Regression, Support Vector Machines.
- Tune the hyper-parameters with appropriate metrics like Accuracy, Precision, Recall, F1 Score, ROC AUC for accurate classification.[6,10]

3.2.4) Address Challenges:

- Negation and Vagueness[3] : Deals with negation and ambiguous phrases.
- Ensure scalability to handle large datasets efficiently.

2.5) Insights and Applications:

- Offer actionable insights on product improvement and customer service.

- Indicate how it could be used in the future to analyze market trends, track reputation, and review customer feedback.

3.3) Deliverables

At the end of this project, the following deliverables will have been achieved:

- A fully trained sentiment analysis model.
- Evaluation metrics indicating performance of the model.
- Detailed analysis of results as well as insights generated from the dataset.
- Recommendations for future improvements in sentiment analysis methodologies.

CHAPTER 4: SYSTEM DESIGN

4.1) Software Requirements:

- Programming Language: Python 3.x
- Libraries and Tools: Anaconda, Natural Language Toolkit (NLTK), Scikit-learn
- Data Format: JSON file with review text, ratings, and metadata.

4.2) Hardware Requirements:

- Processor: Core i5/i7
- Memory: Minimum 8 GB RAM
- Storage: At least 50 GB SSD for fast data processing.

4.3) Data Information:

There are approximately 35 million reviews in the dataset for 18 years. Attributes include:

- Reviewer ID
- Product ID
- Review text and rating (1 to 5 stars)

4.4) System Workflow:

4.4.1) Data Preparation

- Read in the JSON data.
- Extract the relevant field like review text and ratings.

4.4.2) Preprocessing

- Remove 3-star ratings (neutral review).
- Tokenization[6] and Remove Stop Words[9]
- Identify phrases carrying sentiment along with negation.

4.4.3) Model Training

- Train the classifiers like Naïve Bayes, Logistic Regression, and SVM.

4.4.4) Evaluation

- Validate the models by Accuracy, Precision, Recall, F1 Score, and ROC AUC.

CHAPTER 5: METHODOLOGY FOR IMPLEMENTATION (FORMULATION/ALGORITHM)

5.1) Dataset Preparation

Labeling Reviews:

- Reviews with ratings of 1–2 stars were labeled as negative.
- Reviews which scored 4–5 stars were classified as positive.
- Neutral reviews (3 stars) was excluded as it was too vague about the sentiment.

Stored as JSON files, hence parsed and manipulated efficiently in Python.

5.2) Preprocessing

- The raw dataset was subjected to several preprocessing steps before applying feature extraction and training of a machine learning model.

5.2.1) Text Cleaning:

- Special characters, HTML tags, emojis, numbers removed
- Text all lowered to uniformity
- Example
  Raw: "INCREDIBLE product!!! Worth EVERY penny.????????????"
  Cleaned: "incredible product worth every penny"

5.2.2) **Tokenization**[6,9]:

- Split the sentences into individual tokens, or words.
- Done using the NLP libraries like spaCy or NLTK.

5.2.3) **Stopword Removal**[8]:

- Stopped common words: "is," "the," "and", etc. Do not help to determine the sentiment.

5.2.4) Part-of-Speech (POS) Tagging:

- Tagged the token to determine its Part of Speech.
- Identified adjectives and adverbs usually characterize sentiment as in "great" or "terrible".

5.2.5) **Handling Negations**[10]:

- Detected the phrases with negations "not bad", "never boring"

- Made sure that the polarity was correctly represented.

5.2.6) **Bigram and Trigram Features**[6]:

- Also, sequences of two or three words that provide contextual meaning were taken into consideration.
- Example:

  Text: "not bad at all"

  Features: {not_bad: 1, bad_at: 1, at_all: 1}

5.2.7) **Balancing Classes**[7]:

- Ensured fair representation of both positive and negative reviews to avoid model bias.

5.3) Feature Extraction

5.3.1) **Bag of Words (BoW)**[8]:

- Represented text as a vector of word frequencies.
- Disadvantage: Does not consider word order or context

5.3.2) **TF-IDF (Term Frequency-Inverse Document Frequency)[**6]**:**

- Weighted term importance on its frequency in the document as against the dataset.
- Reduced the impact of common words like "product" or "review."
- Formula:

$$\text{TF-IDF(w)=TF(w)*log}(\frac{N}{DF(w)})$$

Where:

TF(w):Term frequency of word www.

DF(w): Document frequency of word www.

N: Total number of documents.

5.4) Sentiment Classification Algorithms

5.4.1) **Naïve Bayes Classifier**[6,9]

Naïve Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem, which is frequently used in classification tasks. It particularly well works with text data because of its simplicity and efficiency. Given the "naïvety" assumption it makes about the independence of features, it surprisingly still works well in many applications, such as sentiment analysis.

Bayes' Theorem

Bayes' Theorem introduces a mathematical representation through which the probability of a class C given the feature X

$$\mathrm{P(C|X)} = \frac{P(X|C).P(C)}{P(X)}$$

Where:

$P(C|X)$: Posterior probability of class Cgiven the feature set X.

P(X|C): likelihood of feature X given in class C.

P(C): probability prior of class C

P(X): probability of feature set X under all class.

Training Phase:

- Calculate P(C), the prior probabilities for every class (positive or negative sentiment).
- Calculate P(X|C), the probability of each word given the class. In general, it is done with the help of word frequency from the training data.

Prediction Phase:

- For a given review, for each class, compute the posterior probability P(C|X)
- Assign to the review the class that maximizes the posterior probability.

5.4.2) **Logistic Regression**[9,3]

Logistic Regression is a statistical technique of binary classification that predicts the probability of an event occurring. It is very popular for sentiment analysis because it is simple yet strong.

Sigmoid Function

Logistic regression models use the sigmoid function to transform any real-valued input into a probability between 0 and 1:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Where: z= $w^{T}x + b$: linear combination of input feature x with weight w and bias b.

The sigmoid function bounds the output between 0 and 1, which makes the output interpretable as a probability.

Training Phase:

Fit the model to the training data by learning the best values of the weights w and bias b which minimize the loss function, for instance cross-entropy loss.

Prediction Phase:

Compute the probability P(y|X) that the review is in the positive class.

Classification: assign the review to the positive class if P(y|X)>0.5 otherwise, classify it as a negative class.

Loss Function

The loss function used for logistic regression is the log-loss (or cross-entropy loss), that penalizes incorrect predictions more heavily:

$$\text{loss}= -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(z_i) + (1-y_i)\log(1-z_i)]$$

where: y_i is true loble for ith review and z_i predict probability for i-th reiew

5.4.3) Support Vector Machines (SVM)[10,6]

SVM is a supervised learning algorithm which mainly operates by identifying an optimal hyperplane that separates classes. It is best suitable for high dimensional datasets.

Linear SVM

Given linearly separable data, SVM produces the hyperplane that maximizes the margin for positive and negative classes. The decision boundary is given by:

$$w^t x + b = 0$$

Where, w: Weight vector perpendicular to the hyperplane.

x; Input feature vector. b; Bias term.

Non-Linear SVM

For non-linear data, SVM employs kernel functions to transform data into a high-dimensional space where a linear hyperplane could easily separate the classes.

Radial Basis Function (RBF) Kernel:

$$\mathrm{K}(x_i, x_j) = \exp(-\gamma \left||x_i - x_j|\right|^2$$

Where:

x_i, x_j: input feature vector, γ kernal coefficient

Training Phase:

- In this step, it solves an optimization problem that maximizes the margin and minimizes the classification errors.
- The approach also puts kernel functions when dealing with non-linear data.

Prediction Phase:

- Compute the distance of the input vector from the hyperplane.
- classify this review as positive or negative, depending on which side of the hyperplane it lies in.

CHAPTER 7: RESULT AND SAMPLE OUTPUT

7.1) **Accuracy**[9]

Accuracy is the percentage of correctly classified instances from the entire instances. It measures how correct the model is when it makes predictions.

TP: Those instances where the model correctly predicts the positive class of instances.

True Negatives (TN): Instances where the model correctly predicts the negative class.

False Positives (FP): Number of times when the model got the positive class wrong.

False Negatives (FN): The model misclassifies examples into a negative class.

$$Accuracy = \frac{True\ positive\ (TP) + True\ Negative(TN)}{Total\ instance}$$

7.2) **Precision**[2]

Precision calculates the percentage of correct true positives out of all positives. This represent s the number of the model's true positives out of all of its positives.

$$pricision = \frac{true\ positive}{true\ positive + false\ positive}$$

7.3) **Recall**[1,3],

Recall, also referred as sensitivity or true positive rate, calculates the number of correct positive instances out of actual positive ones.

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

7.4) **F1 score** (also **F-score** or **F-measure**)[1,3]

The F1 score is a harmonic mean between precision and recall; therefore, if both metrics are important, then it provides a balanced measure. It penalizes extreme values in either precision or recall. where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$F_1 = \frac{2}{\frac{1}{precesion} + \frac{1}{recall}}$$

7.5) ROC AUC (Receiver Operating Characteristic - Area Under Curve)[4,8]

In ROC AUC, it evaluates the model's ability to distinguish between the positive and negative classes. It depends upon the overall performance of the model for all classification thresholds.

True positive rate (TPR) = $\frac{TP}{TP+FN}$

False Positive rate (FPR)=$\frac{FP}{FP+TN}$

ROC Curve

ROC curve: plot of TPR vs FPR for multiple classification thresholds.

Threshold = 1: the model classifies all instances as negative. (low TPR, low FPR)

Threshold = 0: the model classifies all instances as positive. (high TPR, high FPR)

AUC (Area Under Curve)

AUC provides an assessment of model performance overall:

AUC = 1.0: perfect classifier.

AUC = 0.5: random guess (no discriminative ability).

Results obtained using Hold-out Strategy(Train-Test split)

Classifier Name	F_1	Accuracy	Precision	Recall	ROC AUC
Multinomial NB	**86.50%**	**87.31%**	**84.76%**	**89.15%**	**86.41%**
Logistic Regression	**87.82%**	**83.07%**	**88.84%**	**87.78%**	**89.55%**
Linear SVC	**89.82%**	**87.71%**	**88.29%**	**87.60%**	**86.81%**

Fig. 7.1: Result of different models

The Matrix of confusion Format is as follows

True Negative (TN)	False Positive (FN)
False Negative (FP)	True Positive(TP)

The Matrix of confusion of Each Classifier are as follows:

68529	11457
12014	67969

Classifier 1: Multinomial NB

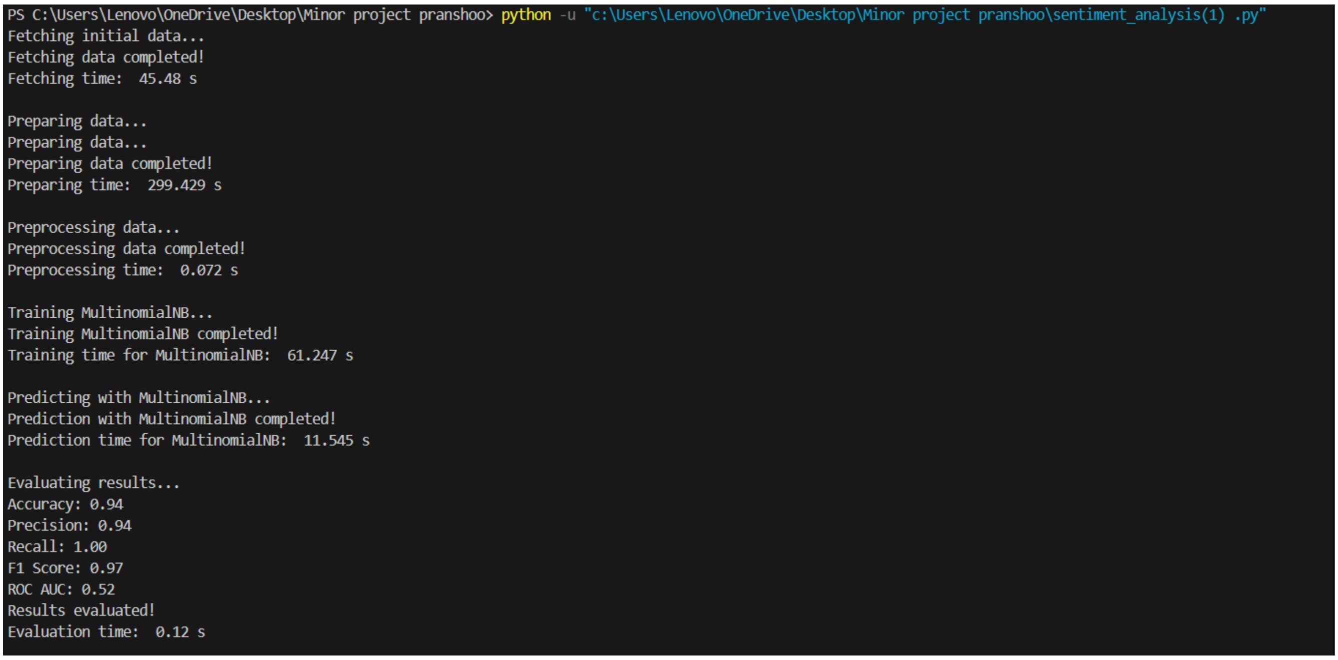
69942	10099
9039	70883

Classifier 2: Logistic Regression

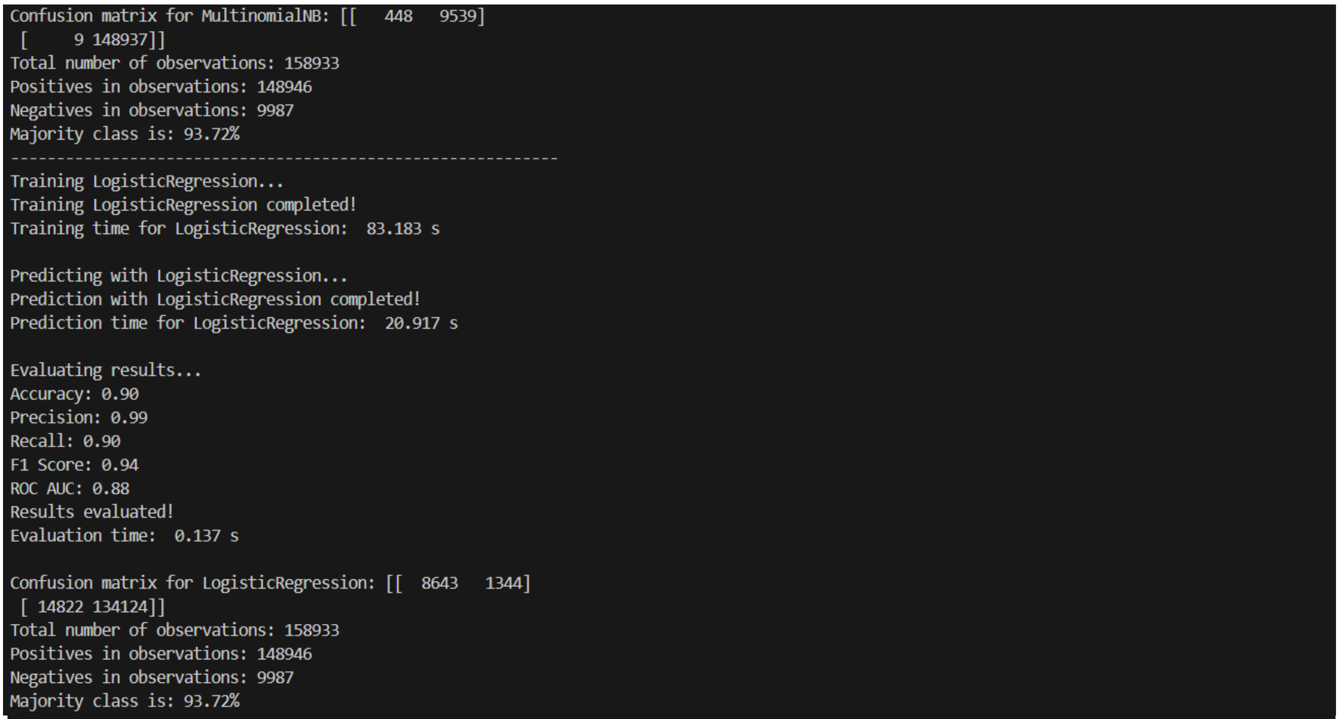
69957	10047
8977	17042

Classifier 3: Linear SVC

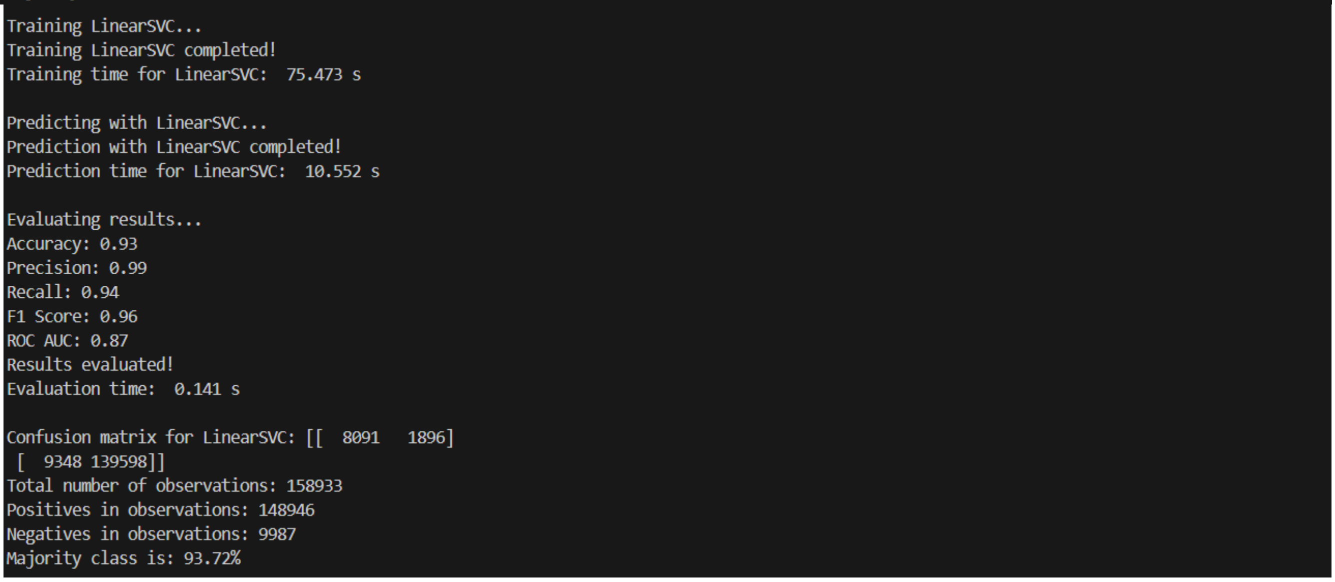
These are the sample image of the produced output:



```
PS C:\Users\Lenovo\OneDrive\Desktop\Minor project pranshoo> python -u "c:\Users\Lenovo\OneDrive\Desktop\Minor project pranshoo\sentiment_analysis(1) .py"
Fetching initial data...
Fetching data completed!
Fetching time:  45.48 s

Preparing data...
Preparing data...
Preparing data completed!
Preparing time:  299.429 s

Preprocessing data...
Preprocessing data completed!
Preprocessing time:  0.072 s

Training MultinomialNB...
Training MultinomialNB completed!
Training time for MultinomialNB:  61.247 s

Predicting with MultinomialNB...
Prediction with MultinomialNB completed!
Prediction time for MultinomialNB:  11.545 s

Evaluating results...
Accuracy: 0.94
Precision: 0.94
Recall: 1.00
F1 Score: 0.97
ROC AUC: 0.52
Results evaluated!
Evaluation time:  0.12 s
```



```
Confusion matrix for MultinomialNB: [[   448   9539]
 [     9 148937]]
Total number of observations: 158933
Positives in observations: 148946
Negatives in observations: 9987
Majority class is: 93.72%
------------------------------------------------------------
Training LogisticRegression...
Training LogisticRegression completed!
Training time for LogisticRegression:  83.183 s

Predicting with LogisticRegression...
Prediction with LogisticRegression completed!
Prediction time for LogisticRegression:  20.917 s

Evaluating results...
Accuracy: 0.90
Precision: 0.99
Recall: 0.90
F1 Score: 0.94
ROC AUC: 0.88
Results evaluated!
Evaluation time:  0.137 s

Confusion matrix for LogisticRegression: [[  8643   1344]
 [ 14822 134124]]
Total number of observations: 158933
Positives in observations: 148946
Negatives in observations: 9987
Majority class is: 93.72%
```



```
Training LinearSVC...
Training LinearSVC completed!
Training time for LinearSVC:  75.473 s

Predicting with LinearSVC...
Prediction with LinearSVC completed!
Prediction time for LinearSVC:  10.552 s

Evaluating results...
Accuracy: 0.93
Precision: 0.99
Recall: 0.94
F1 Score: 0.96
ROC AUC: 0.87
Results evaluated!
Evaluation time:  0.141 s

Confusion matrix for LinearSVC: [[  8091   1896]
 [  9348 139598]]
Total number of observations: 158933
Positives in observations: 148946
Negatives in observations: 9987
Majority class is: 93.72%
```

The rating of dataset is given in the bar graph:

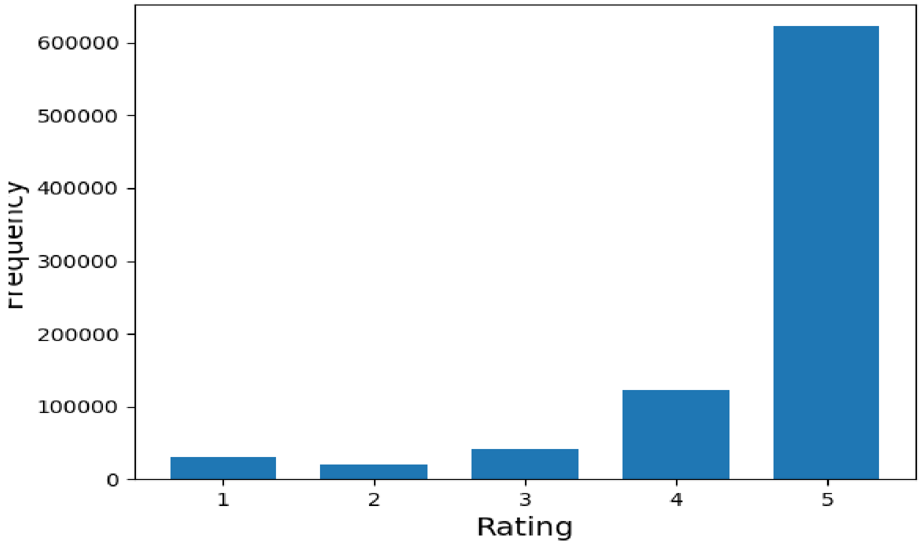


Fig. 7.2: rating of dataset

After successful training the below graph is obtained. The parameters be: F1 Score, Accuracy, Precision, Recall and Roc-Auc.

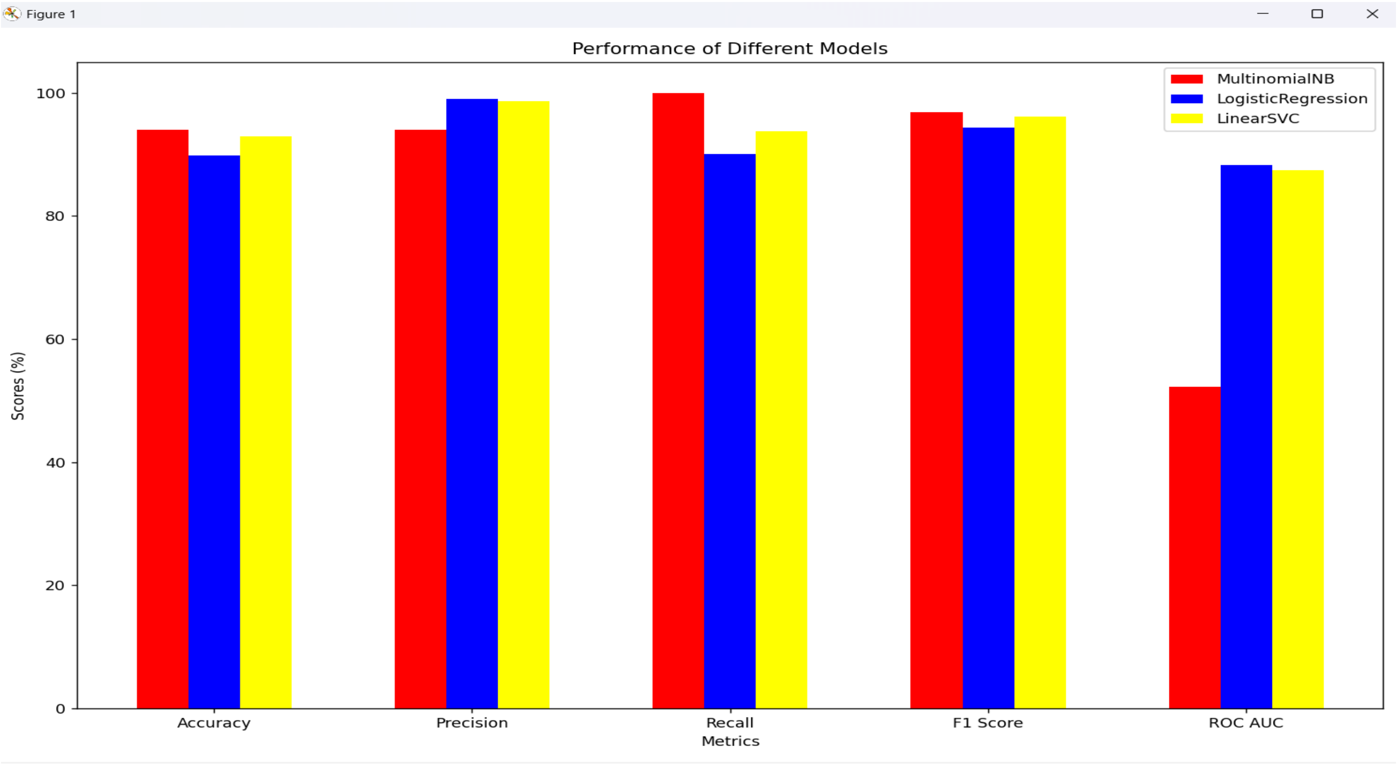


Fig 7.3: bar graph of result

CHAPTER 8: CONCLUSION AND FUTURE SCOPE

Sentiment analysis is a key tool in natural language processing, enabling information extraction from textual data. The project aimed at classification of Amazon product reviews as positive or negative using robust preprocessing techniques, effective feature extraction methods like TF-IDF[2,5], and three machine learning models: Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM). Each model has its specific strengths in a classification task. Naive Bayes excelled in terms of speed and simplicity, logistic regression achieved interpretability and robustness, and finally SVM clearly demonstrated good performance at handling high-dimensional data.

The system implemented data cleaning, tokenization, negation handling, and n-grams to capture the contextual relationship of words. The evaluation of models using Accuracy, Precision, Recall, F1 Score, and ROC AUC was useful as it offered comprehensive insight into their performance, explaining the trade-offs between precision and recall and the significance of choosing the appropriate model for the use case.

This project demonstrates the importance of machine learning in taking enormous masses of unstructured data and turning them into intelligible insights. The outcomes were able to impress the usability of sentiment analysis in applications such as assessment of the customer feedback, product improvement, and market trends. By addressing issues like data imbalance and linguistic nuances, the performance of the models was high and practically viable.

FUTURE SCOPE

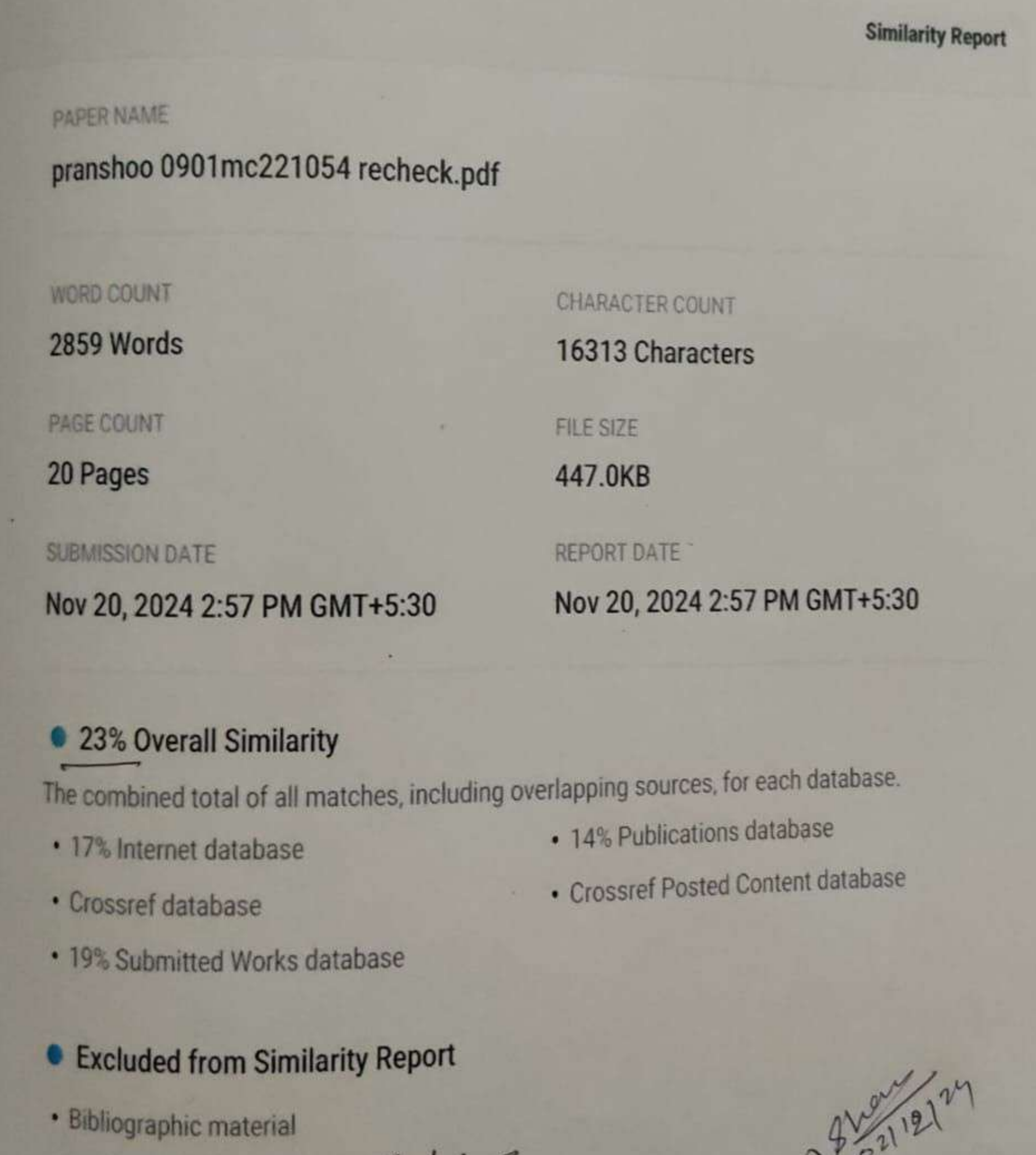
- Enhanced Feature Extraction[2]: Further research could be in the sophisticated techniques with respect to feature or opinion extraction from text, such as product-specific attributes or deeper context understanding. Improved linguistic and semantic analysis would improve the accuracy of classifications.
- Multimodal Data Handling[4,8]: In addition to textual reviews, one can use multimodal data like images, videos, or audio in which this will provide a wider range of sentiments across different platforms.

- Real-Time Sentiment Analysis[1,3,8]: As data starts streaming in real-time, the need for optimization of algorithms towards efficient and accurate real-time sentiment analysis will come to play for applications such as live monitoring of customer feeds.
- Cultural and Language Diversity[3]: Expanding the scope of sentiment analysis to support more languages and dialects, with cultural nuances, makes models applicable to a global audience.
- Sentiment Analysis in Voice-Enabled Systems[1]: Integrating sentiment analysis into voice-enabled assistants can enable them to empathize with the emotions of the users based on tone and language.

REFERENCES

1. S. ChandraKala and C. Sindhu, “Opinion Mining and Sentiment Classification: A Survey,” *International Journal of Computer Applications*, vol. 3, no. 1, pp. 420–427, Oct. 2012.
2. G. Angulakshmi and R. ManickaChezian, “An Analysis on Opinion Mining: Techniques and Tools,” *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, vol. 3, no. 7, 2014.
3. C. Rain, “Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning,” Swarthmore College, Department of Computer Science.
4. P. P. Tribhuvan, S. G. Bhirud, and A. P. Tribhuvan, “A Peer Review of Feature Based Opinion Mining and Summarization,” *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 5, no. 1, pp. 247–250, 2014.
5. G. Carenini, R. Ng, and E. Zwart, “Extracting Knowledge from Evaluative Text,” in *Proc. Third Int. Conf. Knowledge Capture (K-CAP)*, 2005.
6. D. Dave, A. Lawrence, and D. Pennock, “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews,” in *Proc. Int. World Wide Web Conf. (WWW)*, 2003.
7. J. Zhu, et al., “Aspect-Based Opinion Polling from Customer Reviews,” *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 37–49, 2011.
8. J.-C. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou, “Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews,” *Advances in Knowledge Organization*, vol. 9, pp. 49–54, 2004.
9. T. Nasukawa and J. Yi, “Sentiment Analysis: Capturing Favorability Using Natural Language Processing,” in *Proc. 2nd Int. Conf. Knowledge Capture (K-CAP)*, ACM, 2003, pp. 70–77.
10. S. Li, Z. Wang, S. Y. M. Lee, and C.-R. Huang, “Sentiment Classification with Polarity Shifting Detection,” in *Proc. Int. Conf. Asian Language Processing (IALP)*, 2013, pp. 129–132.

Turnitin Plagiarism Report



Similarity Report

PAPER NAME

pranshoo 0901mc221054 recheck.pdf

WORD COUNT 2859 Words	CHARACTER COUNT 16313 Characters
PAGE COUNT 20 Pages	FILE SIZE 447.0KB
SUBMISSION DATE Nov 20, 2024 2:57 PM GMT+5:30	REPORT DATE Nov 20, 2024 2:57 PM GMT+5:30

● 23% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 17% Internet database
- 14% Publications database
- Crossref database
- Crossref Posted Content database
- 19% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material

02/12/24