Sentiment Analysis of Amazon Reviews

Sai Likhitha (Roll No:S20210020289)

April 21, 2024

Abstract

E-commerce has become a powerful force in the quickly changing digital world by providing consumers with accessibility and convenience for shopping. The number of people who purchase online has increased, greatly increasing the significance of product reviews. Thousands of reviews need to be efficiently analysed and categorised, and this model will help customers choose products. In this study, we used a large-scale Amazon dataset and a supervised learning approach to classify reviews based on sentiment. Our model's accuracy which was in between 73-79% is satisfactory, proving its usefulness in speeding up the review analysis procedure and supporting clients in making defensible decisions.

1 Introduction

With just a few clicks, e-commerce provides consumers with unmatched convenience and accessibility to a wide range of goods and services in the rapidly changing digital era. It has emerged as the cornerstone of consumer transactions. Nevertheless, consumers frequently confront the difficult task of sorting through an abundance of reviews in order to locate the ideal product amid this profusion of options.

The exponential rise in popularity of online shopping has highlighted how important product reviews are in shaping consumer behaviour. Customer reviews are becoming more and more important, and it's important for consumers to be able to analyse and comprehend them because they help them determine the suitability and quality of products.

This project aims to leverage machine learning to streamline the sentiment analysis of Amazon product reviews, addressing the growing need for expedited review processing. By employing supervised learning algorithms and advanced natural language processing techniques, the project seeks to develop a model capable of automatically categorizing reviews into positive, negative, and neutral sentiments. This model not only aids customers in making informed decisions but also provides businesses with valuable insights to enhance their offerings and customer satisfaction.

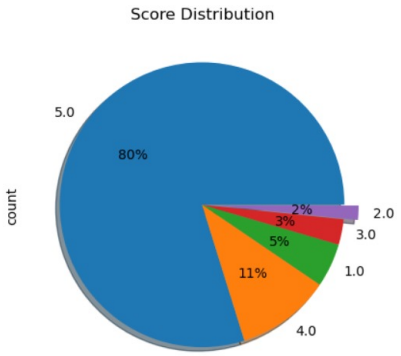

Figure 1: Distribution of Classes in Training and Testing sets

This project demonstrates how machine learning can be used to effectively and practically extract meaningful insights from large volumes of textual data. The project intends to help businesses improve their goods and services while providing consumers with useful information by automating the sentiment analysis process.

2 Explanatory Data Analysis(EDA):

The pie chart[1] shows the distribution of customer ratings on Amazon reviews. The majority of reviews (80%) are 5 star ratings, which corresponds to positive feedback. There are also customers who gave 4 star ratings (11%), 3 star ratings (3%), 2 star ratings (2%), and 1 star ratings (5%).

The two charts in the image [2] show the frequency of reviews by date and by month. The first chart, titled "Frequency of Reviews by Date", shows the number of reviews on the y-axis and the date on the x-axis. The dates are labeled according to the data. There are spikes in the number of reviews on days 1, 7, and 12.

The second chart, titled "Frequency of Re-

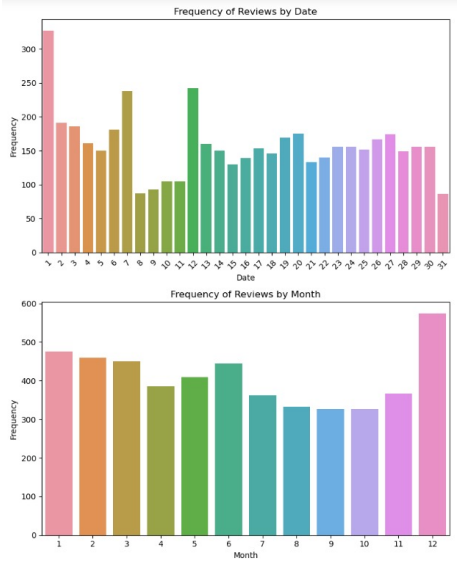

Figure 2: Reviews by Month and Date

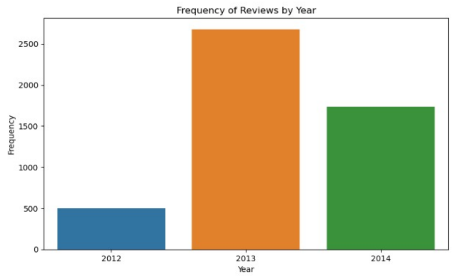

Figure 3: Reviews in a Year

views by Month", shows the number of reviews on the y-axis and the month on the x-axis. The months are labeled from 1 to 12. The chart shows that there were more reviews in December than in others.

The picture[3] "Frequency of Reviews by Year". The x-axis shows the year, ranging from 2012 to 2014. The y-axis shows the frequency of reviews. There are vertical bars for each year. The height of the bar indicates the number of reviews in that year. In 2013, most number of reviews have recorded.

The image [4] illustrates the distribution of review counts over time, depicting the relationship

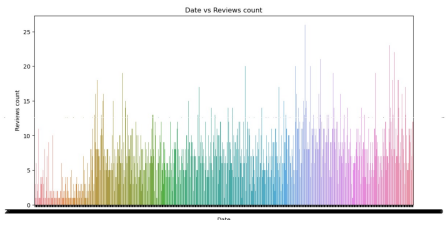

Figure 4: Date vs Review counts Distribution

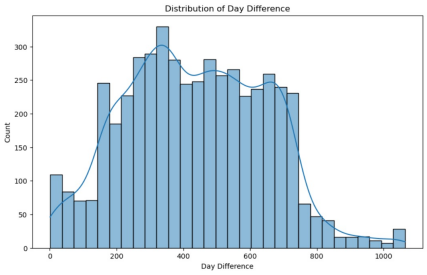

Figure 5: Distribution of Day Difference

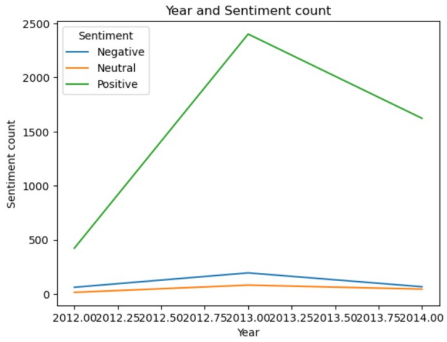

Figure 6: Yearly Sentiment Counts

between the date of reviews and the corresponding number of reviews received. Peaks in the distribution indicate times when a significant number of reviews were posted, while troughs represent periods of relatively fewer reviews. Analyzing this distribution can reveal trends in consumer engagement with products over time, such as spikes in reviews following product launches, promotional events, or seasonal trends

The image [5] illustrates the distribution of day differences between review submissions, providing insights into the frequency of reviews posted relative to each other. It depicts the time intervals between consecutive reviews, indicating patterns in the timing of feedback from consumers. Peaks in the distribution represent common day differences between reviews, suggesting typical intervals at which consumers provide feedback

The image [6] depicts the distribution of sentiment counts (positive, negative, and neutral) over different years, offering insights into the sentiment trends associated with product reviews over time. By examining the distribution, one can discern how the sentiment composition of reviews evolves across various years, identifying shifts in consumer opinions and perceptions of products. Peaks and troughs in sentiment counts for each category (positive, negative, and neutral) indicate periods of heightened or diminished sentiment expression within the dataset.

The image [7] displays the distribution of top

```
Top Reviewers:
reviewerID
A18K1ODH1I2MVB    1
A2RP2S43BN0ZB     1
A3BE8EQ3HG71MK    1
A21MMTDAFAUPQT    1
AM1T7QCP4B8EW     1
Name: count, dtype: int64

Top Reviewers Analysis:
                overall  helpful_yes  total_vote
reviewerID
A18K1ODH1I2MVB      5.0            0           0
A21MMTDAFAUPQT      5.0            0           0
A2RP2S43BN0ZB       5.0            0           0
A3BE8EQ3HG71MK      5.0            0           0
AM1T7QCP4B8EW       5.0            0           0
```

Figure 7: Top Reviewers across years

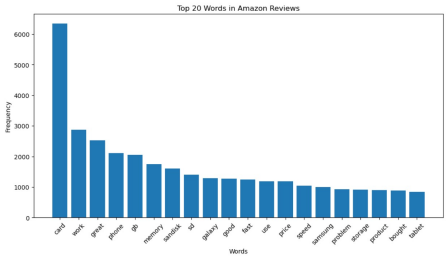

Figure 8: Top 20 most used words across all reviews

reviewers across different years, illustrating the consistency and prominence of certain reviewers over time. By examining this distribution, one can identify reviewers who consistently contribute a significant number of reviews across multiple years, indicating their sustained engagement and influence within the review community

The image [8] visualizes the frequency distribution of the top 20 most used words across all reviews, offering insights into the common themes and topics discussed by reviewers. By examining this distribution, one can identify the key words that appear most frequently in the review dataset, providing valuable information about the prevailing sentiments, opinions, and features highlighted by reviewers.

3 Methodology

One of the biggest online retailers is Amazon, where a vast number of reviews are available for viewing. I made use of dataset from Kaggle, which was called as reviews dataset in the project. The dataset has the following columns:

- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714.
- reviewerName - name of the reviewer

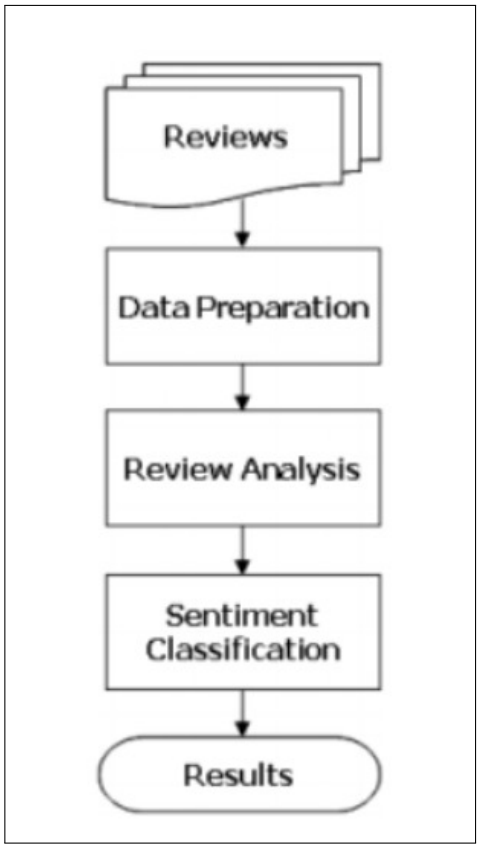

Figure 9: A typical Sentimental analysis model

- helpful - helpfulness rating of the review, e.g. 2/3
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)

3.1 Data collection:

Product reviews data was collected from Kaggle. The dataset encompasses various attributes including reviewer ID, ASIN (Amazon Standard Identification Number) for product identification, reviewer name, a tuple indicating the helpfulness of the review, review text, overall rating on a 5-point scale, a summary of the review, Unix timestamp of the review posting, and the review posting time in a human-readable format. Ratings in the dataset range from 1 to 5, representing satisfaction levels. The dataset was obtained by accessing and downloading it from Kaggle's platform using relevant search queries and filters. Ethical considerations were adhered

to during data collection, ensuring compliance with privacy policies and terms of service. Reviewer identities were anonymized for privacy protection. Prior to analysis, the dataset will undergo preprocessing steps including handling missing values and text normalization. This publicly available dataset serves as a valuable resource for exploring consumer sentiments and preferences.

```
Sentiment
Positive    4447
Negative     324
Neutral      142
Name: count, dtype: int64
```

Figure 10: Imbalance in Dataset

3.2 Data Preprocessing:

3.2.1 Tokenization:

Separate the text data into individual tokens, such as words or phrases, to prepare it for further analysis. Discard unnecessary characters, such as punctuation marks, to clean the text and improve tokenization accuracy

3.2.2 Text Cleaning:

Special characters, URLs, and non-alphabetic characters are eliminated from the text to improve readability and analysis accuracy and lemmatization is applied to normalize words to their base or root form, reducing redundancy in the text data.

3.2.3 Stopword Removal:

Common stopwords are identified by using libraries containing predefined stopword lists for different languages to enhance the accuracy of the analysis, and these stopwords are removed from the text data to focus on meaningful content.

3.3 Feature Engineering:

3.3.1 Bag of Words:

Used in natural language processing, bag of words is a technique for extracting features from simplified text or data. A text or document is modelled in this model as a bag (multiple set) of its words. In sentiment analysis, a "bag of words" is essentially a list of helpful words. The bag of words technique has been utilised by us to extract feature sets. After that, most frequently used terms are obtained from the review dataset.

3.3.2 TF-IDF:

TF-IDF vectorization technique is used to convert raw text data into numerical features. Each review is transformed into a vector where each component represents the importance of a word in that document. TF measures the frequency of a term (word) in a document. It is calculated as the ratio of the number of times a term appears in a document to the total number of terms in the document. IDF measures the importance of a term in the corpus. It is calculated as the logarithm of the ratio of the total number of documents to the number of documents containing the term. The TF-IDF score of a term in a document is the product of its TF and IDF scores. This score reflects how relevant a term is to a specific document in the corpus. The TF-IDF vectors serve as features for our machine learning model. Each review is represented by a TF-IDF vector, capturing the significance of each word in the review. We train the supervised learning models, such as Naive Bayes and Logistic Regression, using the TF-IDF features to predict the sentiment of reviews. The model learns the patterns in the TF-IDF vectors associated with positive, negative, and neutral sentiments.

3.3.3 Data Balancing:

There is a class imbalance in the dataset, so we utilized the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a popular method for generating synthetic samples for the minority class by interpolating new instances between existing ones. This technique helps in improving the performance of our classifier, particularly in scenarios where one class is significantly underrepresented compared to others. But it didn't work as much as expected so 4000 negative and neutral reviews are collected from other dataset and concatenated to the present data to achieve a balanced dataset.

3.3.4 Data Splitting:

Training-Validation-Test Split: The dataset is split into training, validation, and test sets to train, validate, and evaluate the model's performance, respectively.

3.3.5 Label Encoding:

Categorical sentimental variables are encoded into numerical format, using label encoding to

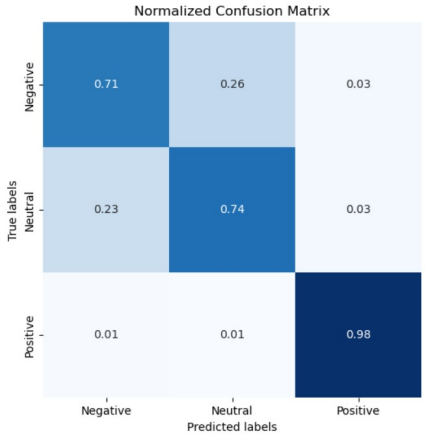

Figure 11: Confusion Matrix Obtained by Naive Baye's Model

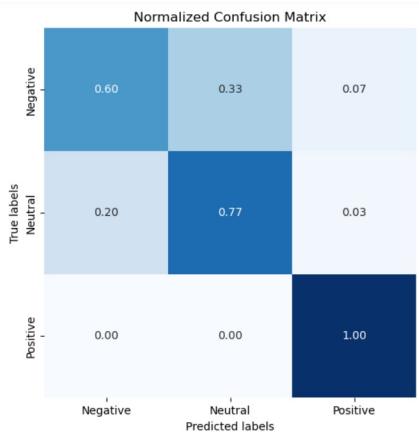

Figure 12: Confusion Matrix Obtained by Logistic Regression Model

make them compatible with Naive Bayes and logistic regression algorithms.

3.3.6 Hyperparameter Tuning and Model Selection:

Hyperparameter tuning using GridSearchCV is performed to find the optimal value of the alpha parameter for the Multinomial Naive Bayes (NB) model, C value for the Logistic regression model. The GridSearchCV function performed a 5-fold cross-validation to evaluate the model's performance for each alpha value and for C values in both the models respectively. The best performing model was selected based on the highest accuracy score achieved during cross-validation.

Finally, the best model is used to make predictions on the test data and evaluate its performance using various metrics, including accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify sentiment in unseen reviews.

4 Results

4.0.1 Naive Bayes model Performance:

The Naive Bayes model [11] trained on the TF-IDF transformed data achieved an accuracy of 78.37% on the test set. The confusion matrix revealed that the model performed well in classifying negative and positive sentiments.

Evaluation Metrics

- Accuracy: 78.37%
- f-score: 0.79

4.0.2 Logistic Regression Performance:

The Logistic Regression model [12] trained on the TF-IDF transformed data achieved an accuracy of 79.45The confusion matrix revealed that the model performed well in classifying negative and positive sentiments.

Evaluation Metrics

- Accuracy: 79.45%
- f-score: 0.73

5 Conclusion

In this project, we investigated the sentiment analysis of Amazon product reviews using machine learning techniques. A dataset of reviews with labels designating positive, neutral, and negative sentiments was utilised. Our objective was to develop and evaluate two sentiment classification models for reviews: Naive Bayes and Logistic Regression.

We prepared the data for modelling by performing data preprocessing, which included text cleaning and feature extraction using TF-IDF. The preprocessed data is then used to train both models, and accuracy metrics are used to assess each model's performance.

The results we obtained demonstrated that in terms of overall accuracy. An invaluable understanding of the most important words for every sentiment category was given by the top words analysis.

In conclusion, this project shows how machine learning can be used to analyse sentiment in Amazon product reviews. This work adds to the expanding body of knowledge in sentiment analysis and has the potential to be expanded upon in the future to include additional datasets and applications.