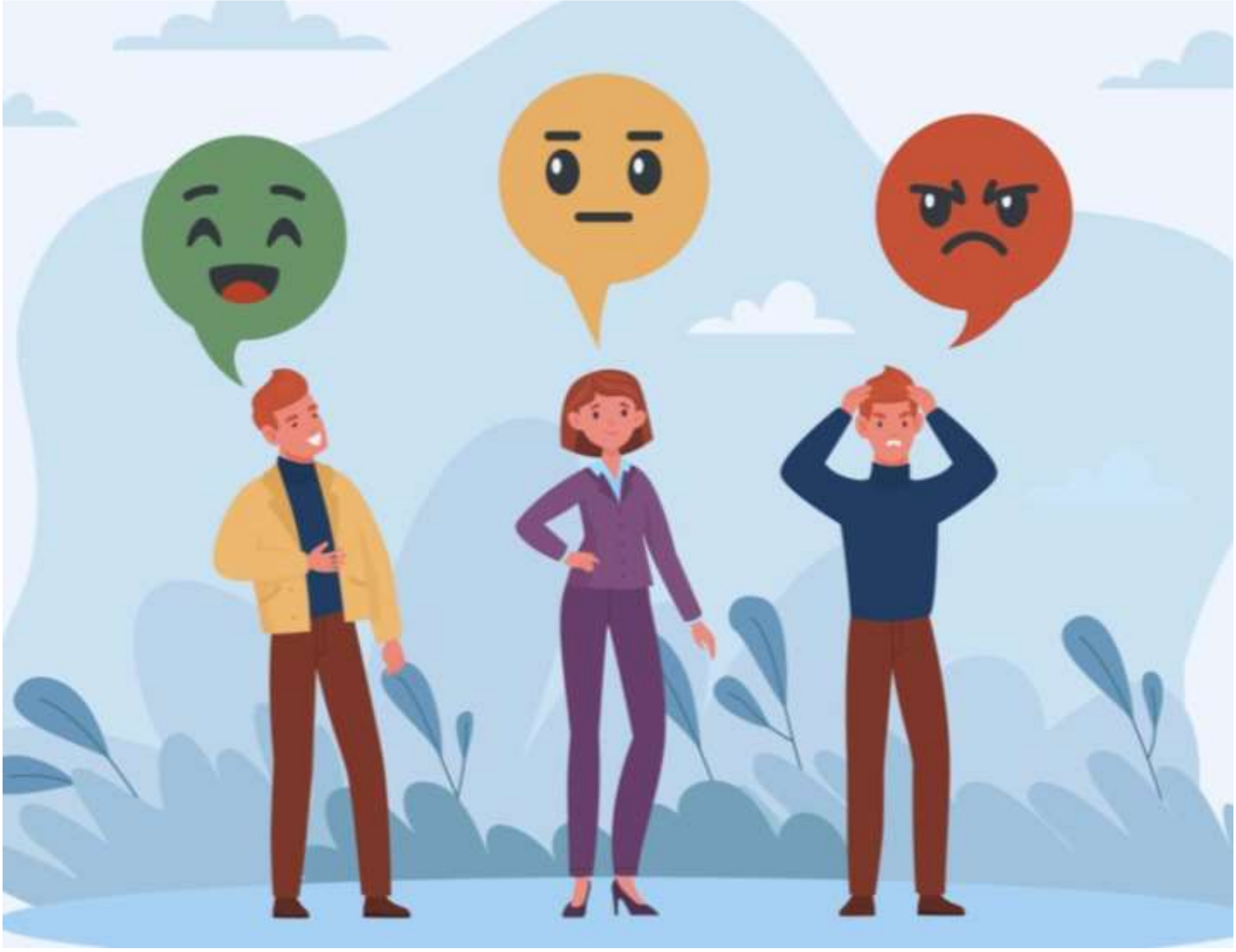


Sentiment Analysis Project

Capstone Project 3

Gokul venugopal 9/24/23 DS AI INTELLIPAAT

Table of Contents

Problem Statement

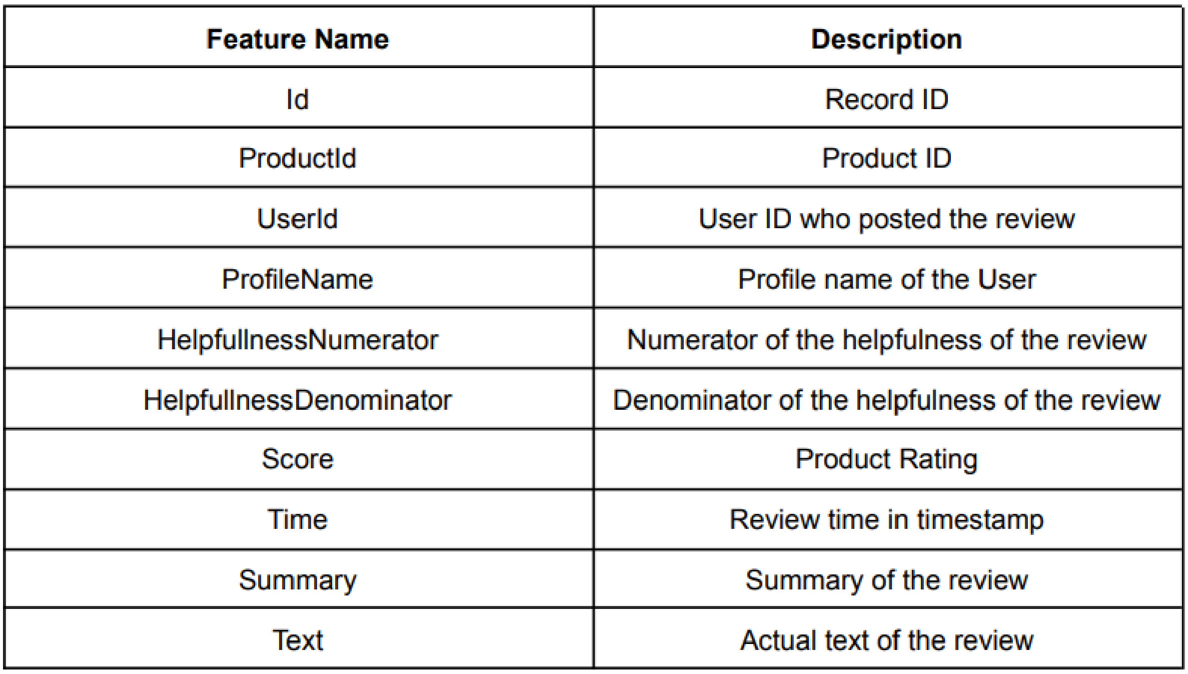
You are working in an e-commerce company, and your company has put forward a task to analyze the customer reviews for various products. You are supposed to create a report that classifies the products based on the customer reviews.

Project Objective

An E-commerce company has put forward a task to analyze the customer reviews for various products. The task is to come up with useful insights for the company and also to classify the customer reviews based on sentiments.

Data Description

The Reviews.csv contains 5,68,454 rows and 10 columns.



Feature Name	Description
Id	Record ID
ProductId	Product ID
UserId	User ID who posted the review
ProfileName	Profile name of the User
HelpfullnessNumerator	Numerator of the helpfulness of the review
HelpfullnessDenominator	Denominator of the helpfulness of the review
Score	Product Rating
Time	Review time in timestamp
Summary	Summary of the review
Text	Actual text of the review

From the given data set of the company, it is observed that the data consist of 5 Lakh sixty eight thousand four hundred and fifty four (5,68,454) records with 10 features as follows.

There are 74,258 unique Products entries with below information.

- ID (Unique Number to identify a single Product review)
- ProductId (Unique Number to identify a product)
- UserId (Unique Number to identify a User)
- ProfileName (Name of the User)
- HelpfullnessNumerator (Number of ratings which was found helpful)
- HelpfullnessDenominator (Number of ratings given by users for the product)
- Score (Rating score given by a customer)
- Time (Time at which User gave the review)
- Summary (Summary of User’s review)
- Text (Entire Text of User’s Review)

Data Preprocessing Steps and Inspiration

The Preprocessing steps included the following steps:

Step 1: Load Data from Kaggle

Step 2: Perform Exploratory Data Analysis and text cleaning

a. Check number of records and its distribution
b. Check Data types
c. Check for missing data, invalid entries and null value entries
d. Identify and remove duplicate reviews by the same user on the same product, keeping only one review per product per user.
e. Convert Ratings to 2 Categories Positive and Negative:
f. Remove HTML Tags
g. Tokenization
h. Eliminate punctuation marks from the text.
i. Filter out common stopwords (e.g., "the," "and," "is").

Step 3: Using different text vectorization methods like Bag of words, TF-IDF, Word2vec.

Step 4: Model Building, three approaches for classifying user review into 2 classes using supervised machine learning model.

Choosing the Algorithm for the Project

Model Selection

Examination of the entries shows supervised data and need of text preprocessing along with vectorization for creating better model. Models like (Logistic Regression, Naïve Bayes, LSTM RNN Model) will be employed for the Sentiment analysis.

Logistic Regression

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.

Assumptions of Logistic Regression

- Independent observations: Each observation is independent of the other. meaning there is no correlation between any input variables.
- Binary dependent variables: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories softmax functions are used.
- Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.
- No outliers: There should be no outliers in the dataset.
- Large sample size: The sample size is sufficiently large

Naive Bayes

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Assumptions made by Naïve bayes models:

- Each feature makes an independent and equal contribution to the outcome.

LSTM RNN

The Long Short-Term Memory, or LSTM, network is a type of Recurrent Neural Network (RNN) designed for sequence problems. Simple RNN models usually run into two major issues. These issues are related to gradient, which is the slope of the loss function along with the error function. Vanishing Gradient problem and Exploding Gradient problem are the issues faced by simple RNN models. LSTM has repeating modules, but the structure is different. Instead of having a single layer of tanh, LSTM has four interacting layers that communicate with each other. This four-layered structure helps LSTM retain long-term memory and can be used in several sequential problems including machine translation, speech synthesis, speech recognition, and handwriting recognition.

Assumptions:

- LSTM assumes that the state at current time step depends on previous time step.

Model Technique and Evaluation

Model Design and Analysis

In Machine Learning, vectorization is a step in feature extraction. The idea is to get some distinct features out of the text for the model to train on, by converting text to numerical vectors. Below are some technique which will be implemented and compared against each other.

- Bag-of-Words(BoW): This vectorization technique converts the text content to numerical feature vectors. Bag of Words takes a document from a corpus and converts it into a numeric vector by mapping each document word to a feature vector for the machine learning model.

- TF-IDF Vectorization: Term frequency-inverse document frequency (TF-IDF) gives a measure that takes the importance of a word into consideration depending on how frequently it occurs in a document and a corpus. TF-IDF gives more weight to less frequently occurring events and less weight to expected events. So, it penalizes frequently occurring words that appear frequently in a document such as "the", "is" but assigns greater weight to less frequent or rare words

- Word2vec : Word2vec is a group of related models that are used to produce word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space.

Model Approach

1. **Data Collection:**
    - Obtain the dataset of customer reviews for Amazon products. Ensure it contains the customer ratings (on a scale of 1 to 5) and the text of the reviews.

2. **Data Preprocessing:**
    - Remove Duplicate Reviews:
        - Identify and remove duplicate reviews by the same user on the same product, keeping only one review per product per user.
    - Convert Ratings to 2 Categories:
        - Ratings 4 and 5 can be labelled as "Positive."
        - Ratings 1, 2, and 3 can be labelled as "Negative."
    - Remove HTML Tags: Use libraries like BeautifulSoup to strip HTML tags from the text if present.
    - Tokenization: Split text into words or tokens.
    - Remove Punctuation: Eliminate punctuation marks from the text.
    - Remove Stopwords: Filter out common stopwords (e.g., "the," "and," "is").

3. **Data Splitting:**
    - Divide your dataset into training and testing sets for model evaluation.

4. **Text Vectorization:**

    Use three different techniques for text vectorization:

    - Bag of Words (BoW):
        - Convert text into a matrix of word frequency counts.
    - TF-IDF (Term Frequency-Inverse Document Frequency):
        - Transform text into numerical vectors that consider word importance.
    - Word2Vec:
        - Represent words in vector form based on their semantic meaning using Word2Vec embeddings.

5. **Model Selection:**

    Choose three different algorithms for sentiment analysis:

    - Logistic Regression:
        - A simple linear model for binary classification.
    - Naive Bayes:
        - Especially suited for text classification tasks.
    - LSTM RNN (Long Short-Term Memory Recurrent Neural Network):
        - A deep learning model that can capture sequential dependencies in text data.

6. **Model Training:**
    - Train each of the selected models using the preprocessed and vectorized data.

7. **Model Evaluation:**

    Evaluate each model using the following metrics:

    - Accuracy: The percentage of correctly classified instances.
    - Precision: The ratio of true positives to the total predicted positives.
    - Recall: The ratio of true positives to the total actual positives.

8. **Hyperparameter Tuning:**

- Fine-tune hyperparameters for each model to optimize their performance.

9. **Comparison and Selection:**

- Compare the performance of the three models based on the evaluation metrics.

10. **Final Model:**

- Select the best-performing model and fine-tuned hyperparameters as your final sentiment analysis model.

Inferences from the Project

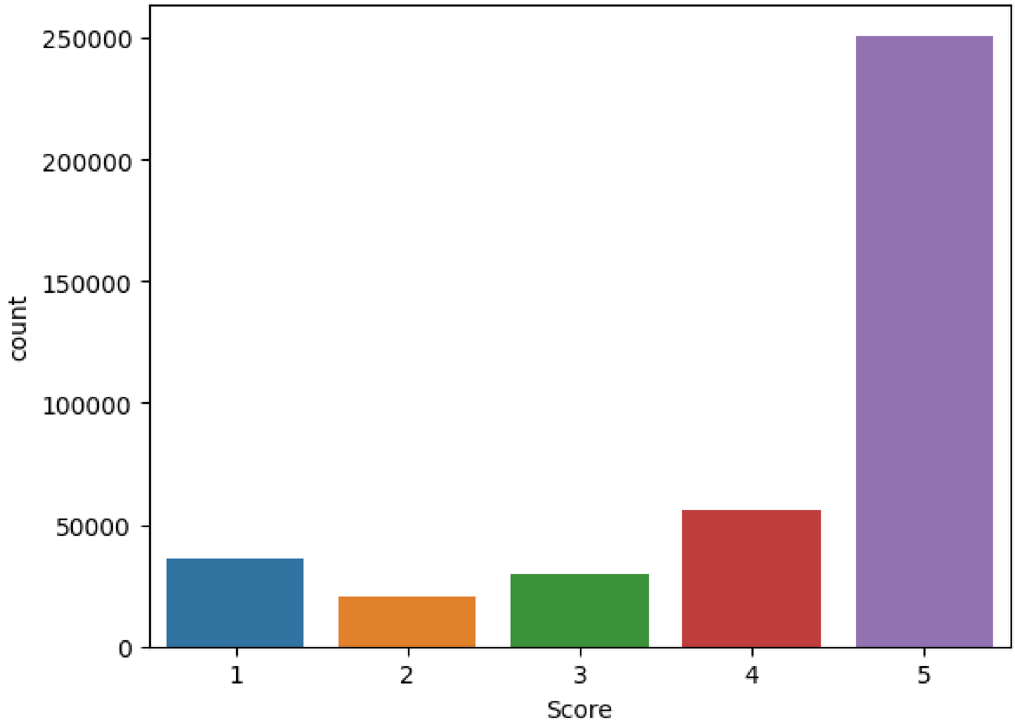
1. Findings from the Dataset.

Invalid data entries and Duplicate product reviews by same user

Observation:

- **Found duplicates of products with repeating ProductId. Taking only one review per user each product**
- **Found 2 invalid entries where numerator values exceeded denominator value and they are removed from dataset.**

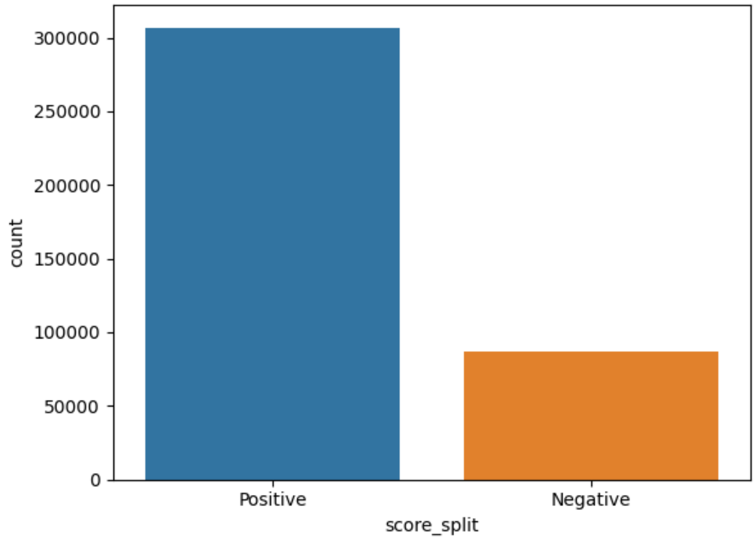
Distribution of 5-Rating score in the dataset



Observations:

- **Rating 5 is the highest given score in the dataset.**

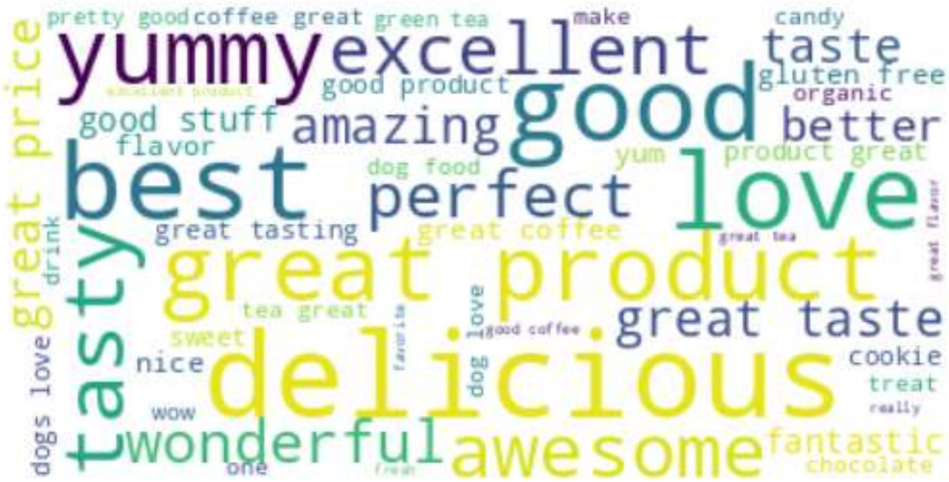
Distribution of 2 category score



Observations:

- **Positive reviews are the most common review given by customers**

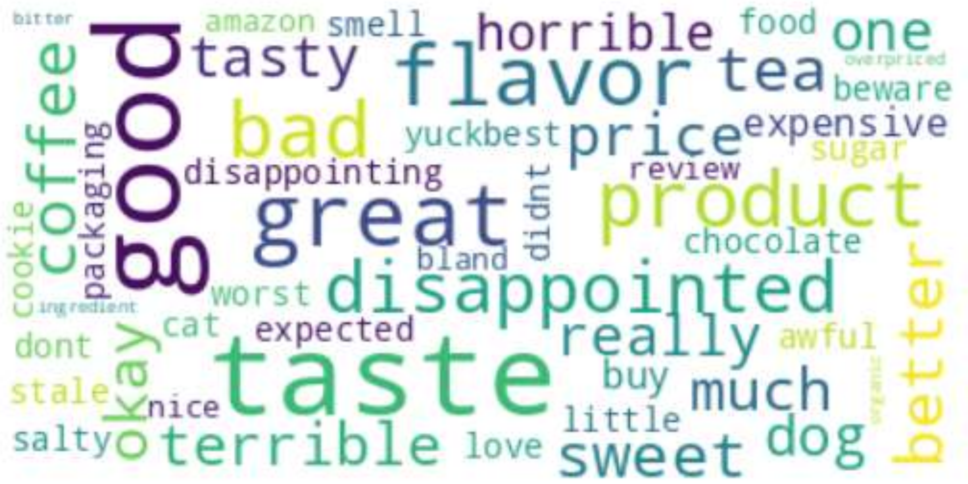
Top words seen in Positive reviews



Observation

- **Some Positive words are great. Good, best, delicious, excellent, tasty, perfect, etc.**

Top words seen in Negative reviews



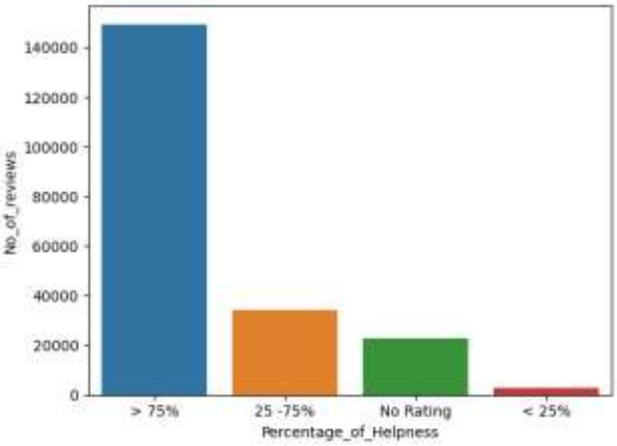
Observation:

- **Some negative words are don't, not good, bad, not great, disappointing, terrible, okay**

Number of Reviews rated by users

Observations: 209307 out of 393917 rated reviews - 53.13 %

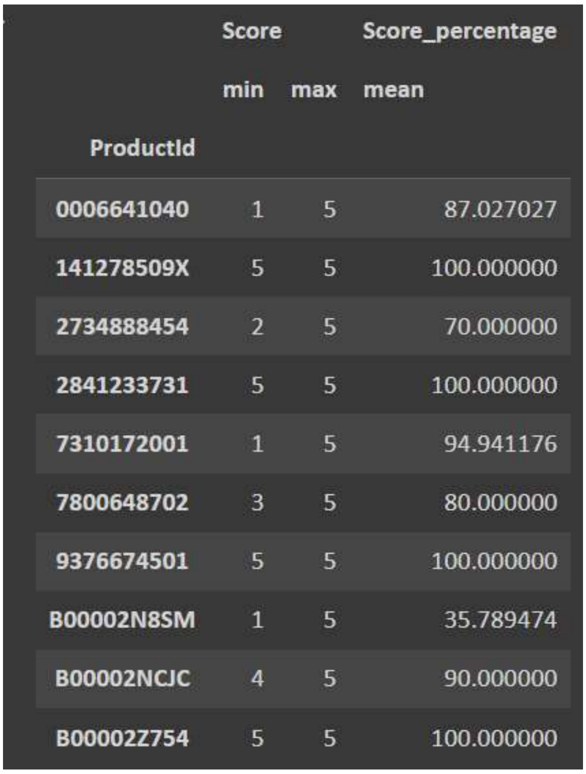
Distribution of helpfulness rating to the users



Observations:

- **More than 75% percent of the customers found the rating helpful.**

Highest and lowest rating for the products. Percentage wise product ratings for the entire data.



	Score		Score_percentage
	min	max	mean
ProductId			
0006641040	1	5	87.027027
141278509X	5	5	100.000000
2734888454	2	5	70.000000
2841233731	5	5	100.000000
7310172001	1	5	94.941176
7800648702	3	5	80.000000
9376674501	5	5	100.000000
B00002N8SM	1	5	35.789474
B00002NCJC	4	5	90.000000
B00002Z754	5	5	100.000000

Total number of reviews by unique profiles

Observations:

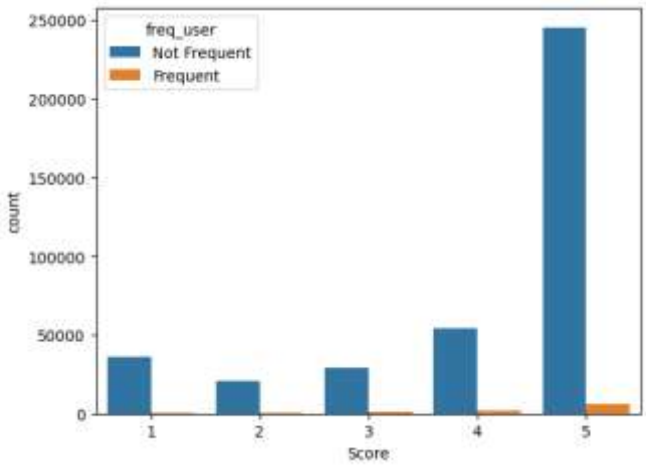
- **Total number of reviews by unique profiles: 175383 / 393917**

Total number of customers or profiles who have given reviews for more than one product

Observations

- **Total number of customers or profiles who have given reviews for more than one product: 218534 / 393917**

<u>Graph of frequent reviewers</u>



Observations:

- **Frequent customers do not rate much as compared to infrequent customers.**
- **Infrequent often provide ratings to the product which they have bought.**

2. Model Building and Comparison:

Supervised classification was performed on the dataset using different text vectorization techniques and these vectors were then used to build models using 3 different algorithms.

Observations:

- Bag of words (BoW) approach works well when used with logistic regression for the current dataset. Good Precision and Recall are observed along with good accuracy score.
- TF-IDF approach works well when used with logistic regression for the current dataset. Good Precision and Recall are observed along with good accuracy score.
- Word2vec did not perform with the current dataset. Observed very low recall scores for negative review category
- LSTM model worked well with this dataset. Highest Precision and Highest Recall was observed using this RNN model. 89% test accuracy was observed.

Best Model: LSTM model (More finetuning can also be done)

The Evaluation metrics are summarized in the Table below.

Bag of Words					
Classification Model	Precision		Recall		Accuracy
	0	1	0	1	
Logistic Regression	0.79	0.90	0.64	0.95	0.882
Naïve Bayes	0.50	0.90	0.68	0.81	0.779
TF-IDF					
Classification Model	Precision		Recall		Accuracy
	0	1	0	1	
Logistic Regression	0.78	0.92	0.70	0.94	0.891
Naïve Bayes	0.63	0.92	0.74	0.88	0.845
Word2vec					
Classification Model	Precision		Recall		Accuracy
	0	1	0	1	
Logistic Regression	0.32	0.79	0.11	0.93	0.750
Naïve Bayes	0.00	0.78	0.00	1.00	0.780
Keras Tokenizer					
Classification Model	Precision		Recall		Accuracy
	0	1	0	1	
LSTM RNN Model	0.77	0.92	0.73	0.94	0.892

Future Possibilities

Sentiment analysis in online ecommerce markets is on the brim of exciting developments. One significant trend is the move toward more personalized shopping experiences. By harnessing sentiment analysis, online retailers can gain insights into individual customer preferences, allowing them to recommend products and services tailored to each shopper. This not only enhances customer satisfaction but also boosts conversion rates and revenue.

Real-time customer feedback is another promising avenue. With the growing shift to online shopping, there is an increasing need for instant feedback and support. Sentiment analysis can play a vital role in monitoring customer sentiment in real-time through chatbots and automated customer service agents. This enables retailers to address issues promptly, prevent negative reviews, and cultivate a positive brand image.

Additionally, sentiment analysis is becoming a key tool for market research and product development. Ecommerce businesses can analyze reviews, comments, and social media discussions to gather valuable market intelligence. This data informs product design and marketing strategies, helping companies stay competitive and aligned with customer demands. As ecommerce continues to evolve, these developments in sentiment analysis will shape the way businesses engage with customers and adapt to shifting market dynamics.

Conclusion

The project undertook a study of online ecommerce company with 3,93,917 unique reviews and 74,258 different unique products. Some important findings form the report include the following.

1. Reviews were classified into 2 categories (Positive and Negative) and LSTM model gave the best accuracy score with good precision and recall.
2. More hyperparameter finetuning can give even better model.
3. To improve the sales, the following steps are recommended:
    a. Determining the key factors influencing sentiment by analyzing the text data to identify recurring themes or topics in positive and negative reviews. These insights can guide product improvements or marketing strategies.
    b. Addressing issues highlighted in negative reviews to enhance product features, quality, or customer support. Customers are more likely to make purchases if they see that their concerns are being addressed.
    c. Highlighting positive sentiment and customer testimonials in marketing materials and product listings. Positive reviews can serve as powerful social proof to influence potential buyers.
    d. Using sentiment analysis to provide personalized product recommendations to customers. Recommend products that align with their preferences based on their past reviews and sentiments.
4. Collect feedback from customers is key step to identify pain points and areas for improvement.

References

1. https://www.kaggle.com/code/akshaysharma001/naive-bayes-with-hyperpameter-tuning

2. https://github.com/MilaNLProc/contextualized-topic-models/issues/114

3. https://keras.io/api/layers/core_layers/embedding/

4. https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer

5. https://www.analyticsvidhya.com/blog/2022/01/sentiment-analysis-with-lstm/

6. https://www.datacamp.com/tutorial/naive-bayes-scikit-learn

7. https://www.kaggle.com/code/akshaysharma001/naive-bayes-with-hyperpameter-tuning

8. https://www.geeksforgeeks.org/long-short-term-memory-lstm-rnn-in-tensorflow/

9. https://www.analyticsvidhya.com/blog/2021/06/lstm-for-text-classification/

10. https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

11. https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/

12. https://www.youtube.com/watch?v=Nk8nM2anJTw