INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Text Data : Sentiment Analysis

Contents:

- ❖ Problem Statement
- ❖ Introduction
- ❖ About The Data
- ❖ Basic EDA
- ❖ Data Preprocessing
- ❖ Model Building
- ❖ Evaluation
- ❖ Conclusion

Problem Statement

- The goal of this project is to leverage a large dataset of Amazon fine food reviews to gain insights into consumer preferences and predict the ratings (scores) of products based on textual reviews.

INTRODUCTION

- Sentiment analysis, also known as opinion mining, is a field within Natural Language Processing (NLP) that builds systems to identify, extract, quantify, and study affective states and subjective information.
- It's designed to understand the sentiment behind a series of words, to gain an understanding of the attitudes, emotions, and opinions expressed within an online mention.

Applications of Sentiment Analysis

1. **Customer Feedback Analysis**: Businesses use sentiment analysis to understand customer feedback on products and services, allowing them to improve customer experience and address concerns proactively.
2. **Social Media Monitoring**: Monitoring social media platforms to gauge public opinion about brands, products, or specific topics, helping in reputation management and marketing strategies.
3. **Market Research and Analysis**: Analyzing sentiments in news articles, blogs, and forums to understand market trends and consumer preferences.

About the Dataset

Source of the Dataset:

- The dataset originates from Amazon, one of the largest online retail platforms globally. Specifically, it contains reviews from the Fine Food category, encompassing a wide variety of food items.
- The reviews were collected up to October 2012, providing a temporal snapshot that extends over more than 10 years, starting from October 1999.

Datafields:

Id: A unique Row Number.

Product ID: A unique identifier for the product being reviewed.

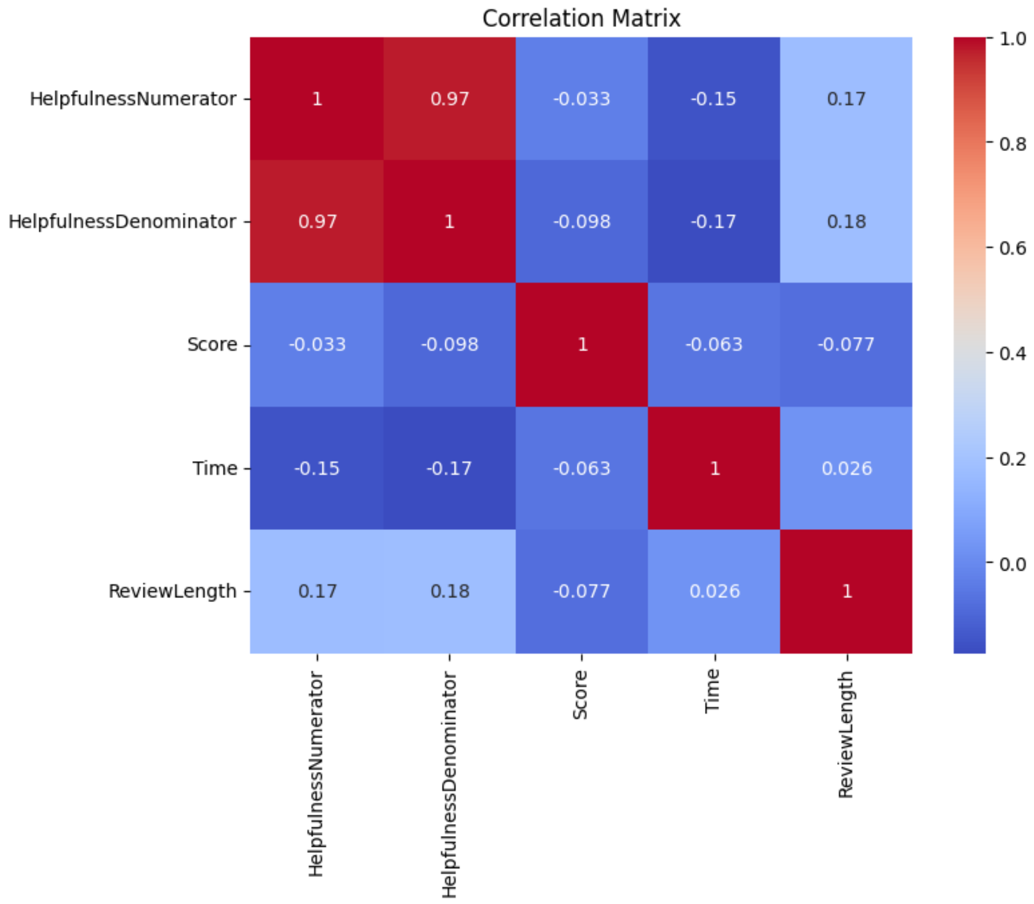
User ID: A unique identifier for the user who wrote the review.

Profile Name: The profile name of the user.
Helpfulness Numerator: The number of users who found the review helpful.
Helpfulness Denominator: The number of users who indicated whether they found the review helpful or not.
Score: The rating between 1 and 5 given by the reviewer.
Time: The Time stamp for the review.
Review Summary: A brief summary of the review.
Review Text: The full text of the review.

BASIC EDA

The plot generates a heatmap visualizing the correlation matrix of a Data Frame, which helps in understanding the relationships between various numerical features within the dataset.

There are Strong Relationships between itself features which some dark red(1.0) and light red(0.8)
There are weak relationships around 0.2.
There is no correaltion around 0.0 and below 0.



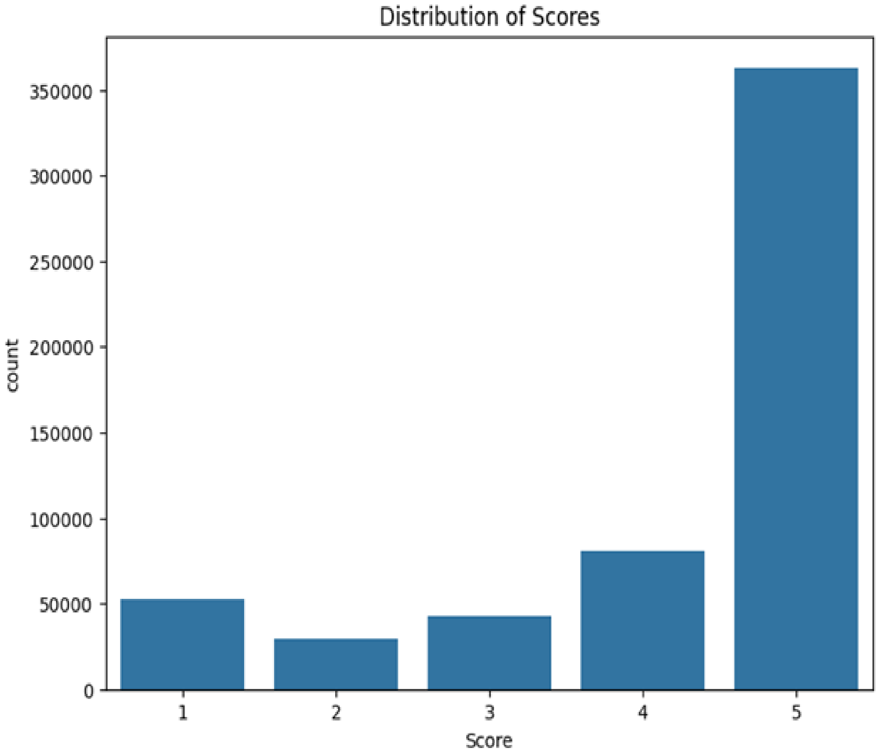
Correlation Matrix

	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	ReviewLength
HelpfulnessNumerator	1	0.97	-0.033	-0.15	0.17
HelpfulnessDenominator	0.97	1	-0.098	-0.17	0.18
Score	-0.033	-0.098	1	-0.063	-0.077
Time	-0.15	-0.17	-0.063	1	0.026
ReviewLength	0.17	0.18	-0.077	0.026	1

It tells about count plot of the distribution of "Score" values within a dataset , visualizing how frequently each score occurs.

The Highest Score is 5 distributing 3,50,000.

The lowest Score is 2 distributing around 25000-30000.



Data Preprocessing

Text Cleaning

Removing Special Characters and Lowercasing

Text data often contains various characters that are not useful for analysis, such as punctuation, special symbols, and numbers. Removing these characters helps in standardizing the text data. Lowercasing all letters ensures that the same words in different cases are treated as identical.

Tokenization

Breaking Text into Tokens (Words or Phrases)

Tokenization is the process of splitting text into individual elements, called tokens. This is a fundamental step for understanding the context or frequency of words within the text.

Stop Words Removal

Eliminating Common Words that Add Little Value

Stop words are common words like "is", "and", "the", etc., that appear frequently in a language but do not contribute much to the meaning of a sentence, especially for the purpose of analysis.

Lemmatization

Converting Words to Their Base Form

Lemmatization is the process of reducing words to their base or root form. Unlike stemming, lemmatization considers the context and converts the word to its meaningful base form, which is a valid word in the language.

Turning Text into Numerical Vectors for Machine Learning

TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF is a numerical statistic that reflects how important a word is to a document in a collection or corpus. It increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

Word Embeddings

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation.

Useful for capturing the context of words within documents, semantic similarity among words, and improving the performance of machine learning models on tasks such as sentiment analysis.

Model Building Algorithms:

1.Logistic Regression

Logistic Regression is a powerful statistical method used for binary classification tasks, capable of predicting the probability of an outcome that can be categorized into one of two groups.

2. Decision Trees

Tree-based model for classification and regression. Splits the data into subsets using feature values, making the decision based on the feature.

3.Random Forest

Ensemble of Decision Trees. It is used for both classification and regression tasks. Builds multiple decision trees and merges them together to get a more accurate and stable prediction.

EVALUATION

1. *Accuracy*

The ratio of correctly predicted observations to the total observations. It's the most intuitive performance measure.

$$Accuracy=TP+TN / TP+TN+\mathrm{FP}+\mathrm{FN}$$

2. Precision (Positive Predictive Value)

The ratio of correctly predicted positive observations to the total predicted positives. It measures the quality of the positive class predictions.

$$Precision=TP / TP+\mathrm{FP}$$

3. Recall (Sensitivity, True Positive Rate)

The ratio of correctly predicted positive observations to all observations in the actual class. It measures the model's ability to capture the positive class.

$$Recall=TP/ TP+\mathrm{FN}$$

4. F1 Score

The weighted average of Precision and Recall. It takes both false positives and false negatives into account.

$$F\mathrm{1}=2\times Precision\times Recall / \mathrm{Precison}+ \mathrm{Recall}$$

LOGISTIC REGRESSION

```
0.729107844948149

             Precision   recall   f1-score    support

1               0.65       0.65     0.65       10597
2               0.41       0.19     0.26        5816
3               0.44       0.26     0.33        8419
4               0.50       0.22     0.30        16122
5               0.78       0.95     0.86        72737


accuracy                             0.73       113691
macro avg       0.56        0.45     0.48        113691
weighted avg 0.69           0.73     0.69        113691
```

DECISION TREE

```
 0.7500857587671848

              precision   recall  f1-score   support

    1            0.63       0.62     0.63      10597
    2            0.51       0.45     0.48       5816
    3            0.53       0.48     0.50       8419
    4            0.55       0.52     0.54      16122
    5            0.84       0.88     0.86      72737


    accuracy 0.75 113691
     macro avg 0.61 0.59 0.60 113691
    weighted avg 0.74 0.75 0.75 113691
```

Random Forest:

```
0.8064842423762655
                 precision   recall   f1-score    support
1                  0.86       0.63     0.73        10597
2                  0.98       0.39     0.56        5816
3                  0.93       0.41     0.57        8419
4                  0.90       0.43     0.59        16122
5                  0.78       0.99     0.88        72737
accuracy                               0.81        113691
macro avg      0.89            0.57     0.66        113691
weighted avg 0.83             0.81     0.78        113691
```

CONCLUSION

- *The Random Forest demonstrates the highest overall accuracy, indicating it is the most effective at correctly classifying instances across all classes.*
- *The decision tree model shows moderate accuracy, and the logistic regression model has the lowest accuracy of the three.*

Precision, Recall, and F1-Score

- *Logistic Regression shows moderate precision and recall across the board , with a notable imbalance in performance across classes, particularly struggling with classes 2, 3, and 4.*
- *Decision Tree shows improved performance , particularly in handling minority classes, suggesting better handling of class imbalance than logistic regression.*
- *The Random Forest Model demonstrates very high precision for classes 2, 3, and 4 but at the cost of recall, indicating a high number of false negatives.*

  However, it excels in identifying the majority class (class 5) with high recall.

Thank you