INFORMATION RETRIEVAL
Assignment -3 Report

Hrishikesh Khambete (MT23038)

Libraries Used:

Pandas: a powerful Python library for data manipulation and analysis, facilitating tasks such as data cleaning, exploration, and transformation through its intuitive DataFrame structure.

Gzip: a compression utility in Python that enables efficient compression and decompression of files using the gzip format, commonly used for reducing file sizes during data transmission or storage.

Re: RegEx can be used to check if a string contains the specified search pattern. RegEx Module. Python has a built-in package called re.

Pickle:A Python module used for serializing and deserializing Python objects, facilitating data storage and retrieval in a compact binary format, commonly employed for object persistence and inter-process communication.

NLTK (Natural Language Toolkit): A Python library for natural language processing (NLP) tasks, offering tools for tokenization, stemming, tagging, parsing, and more, aiding in the analysis and understanding of human language.

Scikit-learn: A Python library for machine learning, providing a wide range of supervised and unsupervised learning algorithms, along with tools for model selection, evaluation, and data preprocessing, enabling developers to build predictive models and perform data analysis tasks efficiently.

Matplotlib: A Python library for creating static, interactive, and animated visualizations, offering a wide range of plotting functions and customization options, empowering users to generate publication-quality graphs and charts for data analysis and presentation purposes.

1. Data Acquisition and Preparation:

i) Obtained the Amazon Reviews dataset, focusing on Electronics, specifically utilizing the 5-core dataset for a smaller subset to facilitate experimentation.
ii) Employed Python to load the dataset into a pandas DataFrame, while keeping product metadata separate for potential use in subsequent analyses.
iii) Carried out preprocessing tasks, including handling missing data, removing duplicates, and other data cleaning procedures to ensure the reliability and integrity of the data.

2. Product Selection:

For the purpose of analysis, the product 'Headphones' was selected from Amazon Electronics dataset.

3. Analysis of 'Headphones' Product:

Filtered the dataset to isolate rows pertaining exclusively to the 'Headphones' product, facilitating targeted analysis.
Determined the total count of rows associated with the 'Headphones' product to gauge dataset representation.

4. Descriptive Statistics of the “Headphone’s Product” :

a. Total Number of Reviews: 411201
b. Average Rating Score: 4.112156828412382
c. Number of Unique Headphones: 8064
d. Number of Good Ratings: 353401
e. Number of Bad Ratings: 57800

Text Preprocessing for Reviews:

a. Eliminating HTML Tags:
● Utilizing appropriate parsing techniques or libraries like BeautifulSoup, any HTML tags within the review text are eliminated**.**

b. Removing Diacritics (accents):
● Diacritics are substituted with their non-diacritic equivalents to standardize text and ensure uniformity.

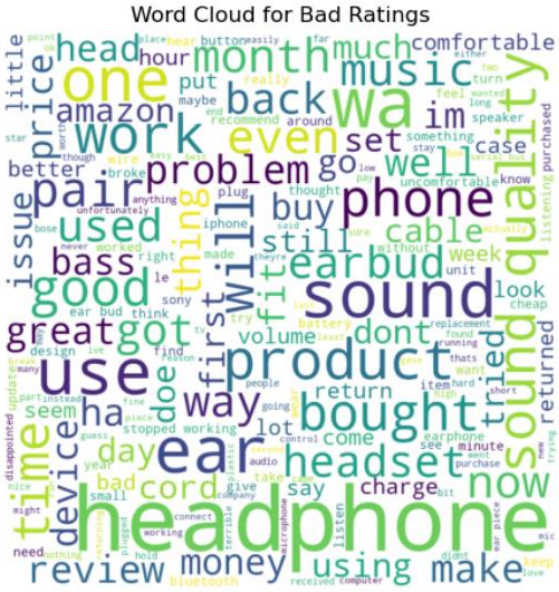
c. Expanding Abbreviations:
● Abbreviations and acronyms commonly found in reviews are expanded to their full forms to enhance comprehension and readability.

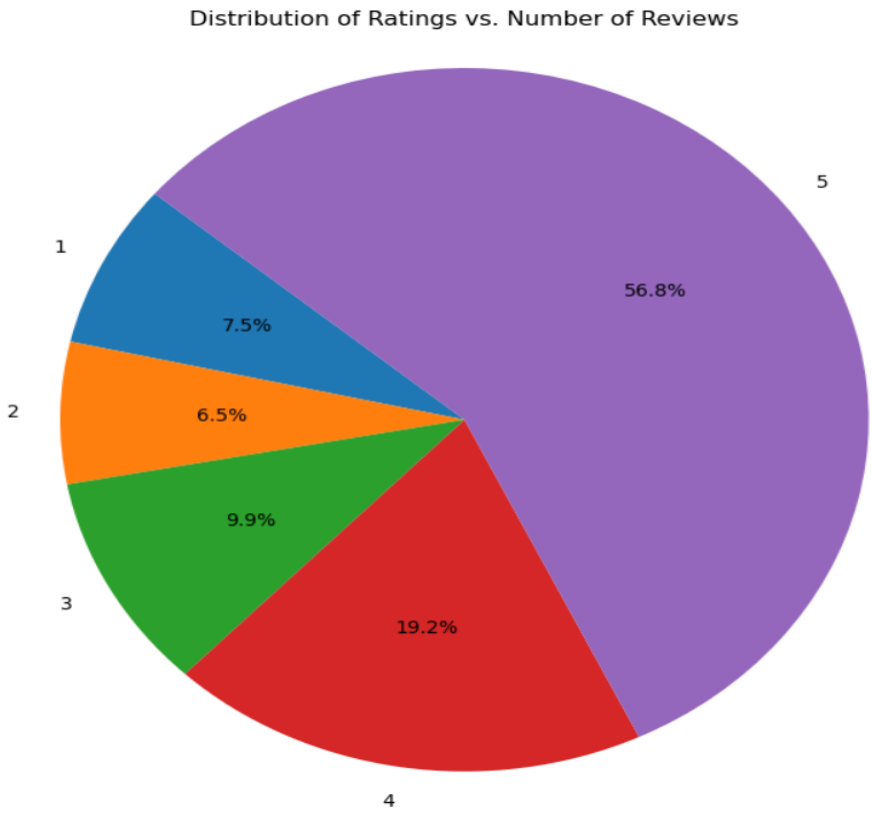
d. Eliminating Special Characters:
● Special characters, symbols, and punctuation marks are eradicated from the review text to emphasize the core content.

e. Lemmatization:
● Employing lemmatization, words within the review text are transformed into their base or dictionary forms, simplifying the text for analysis and interpretation.

f. Text Normalization:
Text normalization methods like converting to lowercase, eliminating stopwords, and tokenization can be employed to further refine and standardize the review text, preparing it for analysis and modeling purposes.

Word Cloud:





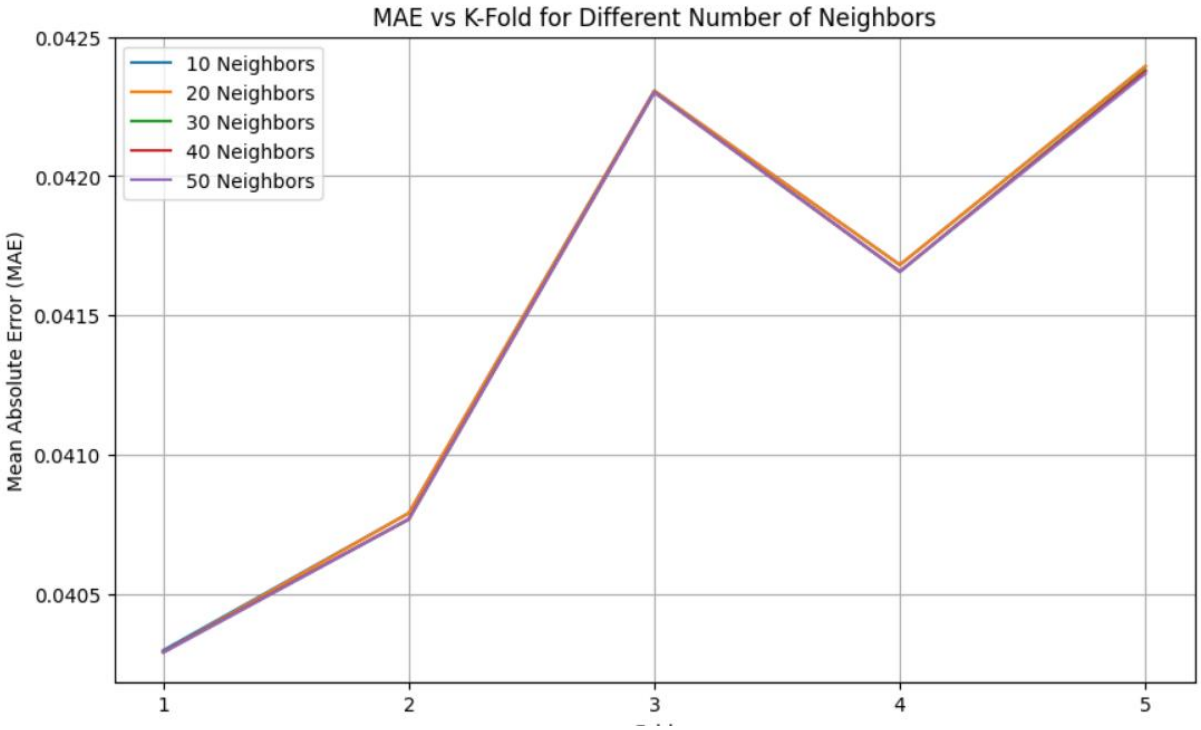
Feature Engineering using Hashed Vector :

	0	1	2	3	4	5	6	7	8	9	...	990	991	992	993	994	995	996	997	998	999
0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.149071	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.301511	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
411196	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
411197	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
411198	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
411199	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
411200	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0

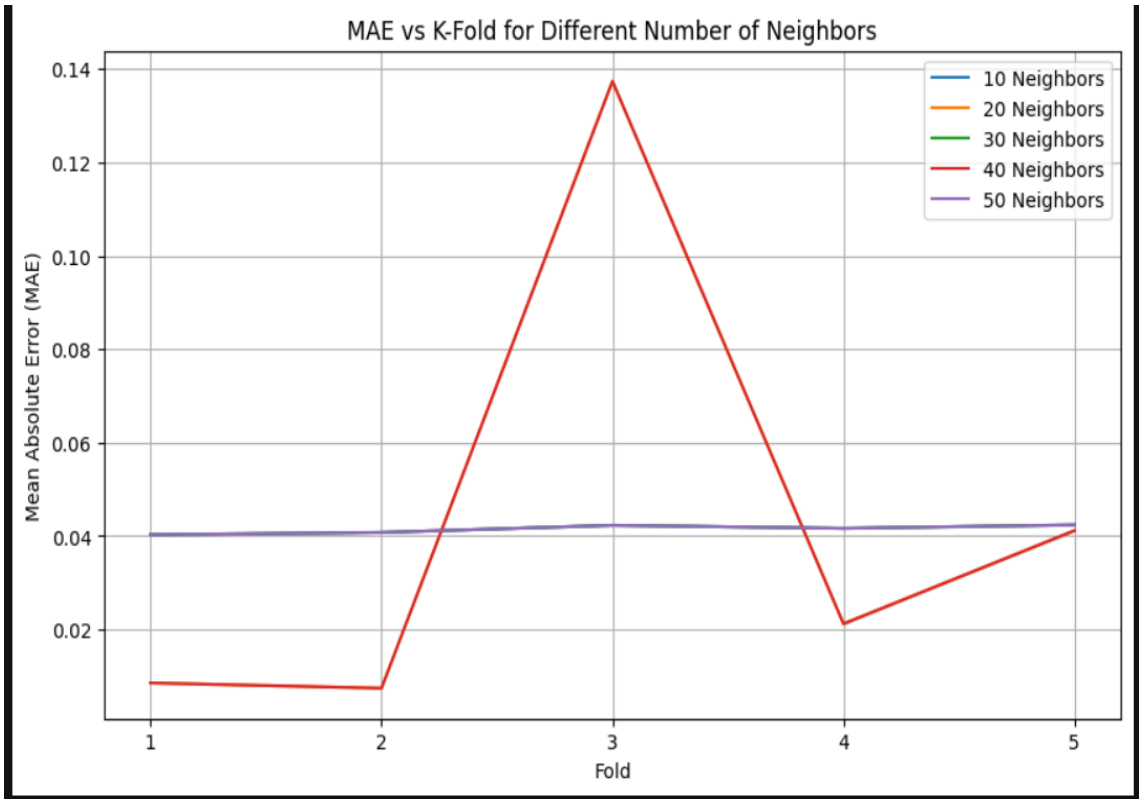
Performance of 5 Machine learning Model

	Classifier	Class	Precision	Recall	F1 Score	Support
0	Random Forest	Good	0.852878	0.251256	0.388161	1592
1	Random Forest	Average	0.896907	0.066565	0.123932	1307
2	Random Forest	Bad	0.801994	0.996875	0.888879	9601
3	Gradient Boosting	Good	0.786996	0.220477	0.344455	1592
4	Gradient Boosting	Average	0.539216	0.042081	0.078070	1307
5	Gradient Boosting	Bad	0.797356	0.992605	0.884332	9601
6	Naive Bayes	Good	0.293434	0.538945	0.379982	1592
7	Naive Bayes	Average	0.151463	0.475134	0.229702	1307
8	Naive Bayes	Bad	0.905223	0.516300	0.657558	9601
9	Logistic Regression	Good	0.658192	0.439070	0.526752	1592
10	Logistic Regression	Average	0.432314	0.075746	0.128906	1307
11	Logistic Regression	Bad	0.831742	0.971045	0.896012	9601
12	Decision Tree	Good	0.383085	0.386935	0.385000	1592
13	Decision Tree	Average	0.248942	0.224943	0.236334	1307
14	Decision Tree	Bad	0.837092	0.846683	0.841860	9601

Collaborative Filtering:



User-item matrix



Item-item matrix