Dissertation Submitted for the partial fulfillment of the **M.Sc. as a part of M.Sc. (Integrated) Five Years Program AIML/Data Science** degree to the Department of AIML & Data Science.

Project Dissertation

Women's E-commerce Clothing Reviews Classification

Submitted to

By

Agarwal Shruti Hemant (DS 01)

Sharma Radhika Shivkumar (DS 13)

Sandhu PrableenKaur CharanjitSingh (DS 17)

Esha Mishra (AIML 18)

Semester-IX

Under the guidance of

Rashmi Pandey[1], Dr. Ravi Gor[2]

M.Sc. (Integrated) Five Years Program AIML/Data Science

Department of AIML & Data Science.

School of Emerging Science and Technology

Gujarat University

October 2023

DECLARATION

This is to certify that the research work reported in this dissertation entitled "**Women's E-commerce Clothing Reviews Classification**" for the partial fulfillment of M.Sc. as a part of M.Sc. (Integrated) Data Science degree is the result of investigation done by myself.

Place: Ahmedabad Agarwal Shruti Hemant

Date: 13th October 2023

DECLARATION

This is to certify that the research work reported in this dissertation entitled "**Women's E-commerce Clothing Reviews Classification**" for the partial fulfillment of M.Sc. as a part of M.Sc. (Integrated) Data Science degree is the result of investigation done by myself.

Place: Ahmedabad Sharma Radhika Shivkumar

Date: 13th October 2023

DECLARATION

This is to certify that the research work reported in this dissertation entitled "**Women's E-commerce Clothing Reviews Classification**" for the partial fulfillment of M.Sc. as a part of M.Sc. (Integrated) Data Science degree is the result of investigation done by myself.

Place: Ahmedabad Sandhu PrableenKaur CharanjitSingh

Date: 13th October 2023

DECLARATION

This is to certify that the research work reported in this dissertation entitled "**Women's E-commerce Clothing Reviews Classification**" for the partial fulfillment of M.Sc. as a part of M.Sc. (Integrated) AIML degree is the result of investigation done by myself.

Place: Ahmedabad Esha Mishra

Date: 13th October 2023

ACKNOWLEDGMENT

We are writing this acknowledgment to express our sincere gratitude to everyone who has supported us throughout our project.

Firstly, we would like to extend our heartfelt thanks to Dr. Ravi Gor, Co-Ordinator of Department of AIML and Data Science, School of Emerging Science and Technology at Gujarat University for providing us with the necessary resources and guidance required for completing this project.

We would also like to express our gratitude to our mentor Mrs. Rashmi Pandey who helped us to navigate through the challenges faced during the project. Your expertise and knowledge in the subject matter were instrumental in shaping our ideas and concepts.

Our sincere thanks also go to the admin staff for their tireless efforts in providing administrative support and facilitating the smooth functioning of the project.

Last but not the least, we would like to thank our friends and family for their unwavering support and encouragement throughout this journey. Your constant motivation and belief in us have been the driving force behind our success. Only because of them, we were able to create this project and make it a good and enjoyable experience.

Thank you everyone for everything.

~ Agarwal Shruti Hemant

~ Sharma Radhika Shivkumar

~ Sandhu PrableenKaur CharanjitSingh

~ Esha Mishra

TABLE OF CONTENT

Chapter 1: Abstract & KeyWords

ABSTRACT

Our project is about understanding the feedback on women's clothes purchased online. We have analyzed and sorted women's e-commerce clothing reviews into positive and not positive sentiments. We're using various machine learning models like Logistic Regression, Recurrent Neural Network (RNN), Decision tree, and Random Forest to find the most accurate sentiment classification model for our women's clothing reviews.

But it's not just about classifying reviews. We've also created a user-friendly dashboard that presents insights and trends from the review data in a simple and clear way. We've designed a Graphical User Interface (GUI) to clearly show the sentiment of each review. This not only ensures accurate sentiment analysis but also makes it easy for both businesses and consumers to understand the results in the fast-paced online clothing world.

By combining advanced machine learning with user-friendly tools for visualization, we're taking a step towards understanding the complexities of women's clothing reviews in the ever-changing e-commerce world. Our project will provide valuable insights and practical information for online retailers and consumers.

KEYWORDS

Sentiment analysis, E-Commerce, Classification, Reviews, Women's clothing, NLP

Chapter 2: Introduction

2.1 Background

The retail landscape has witnessed a profound transformation in recent years, primarily catalyzed by the exponential growth of online shopping. This paradigm shift has not only redefined the way individuals purchase goods but has also significantly impacted the fashion industry, particularly women's clothing purchases. With the rise of digital platforms, consumers now have unparalleled access to an extensive array of products and brands at their fingertips.

2.2 Problem Statement

In today's digital age, consumers increasingly rely on online feedback to inform their purchasing decisions. Therefore, developing an automated system to understand and categorize these sentiments is paramount. The problem revolves around the analysis and classification of customer reviews for women's clothing brands, specifically focusing on distinguishing between positive and negative sentiments expressed in these reviews.

2.3 Objective

This shift towards online shopping and the reliance on product reviews necessitate a nuanced understanding of consumer sentiments. Analyzing these sentiments, differentiating between positive and not positive feedback, is not only a challenge but also an opportunity to enhance the online shopping experience for consumers and empower businesses to meet customer expectations effectively. The ultimate goal is to provide value to both consumers and brand managers. Consumers can use this classification to make informed purchase decisions, while brand managers can use it to identify areas for improvement in their products and services.

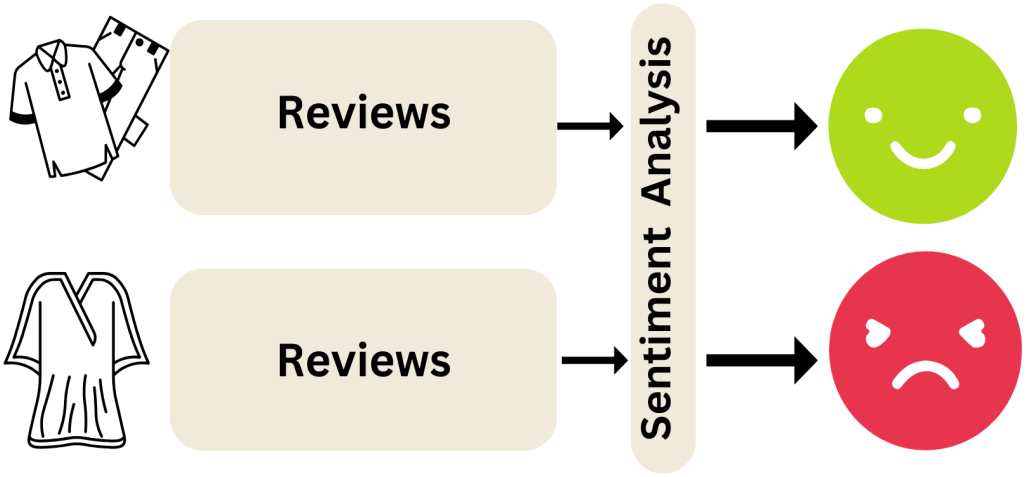
2.4 Introduction

In this digital age, product reviews have emerged as crucial influencers in the decision-making process of online shoppers, serving as beacons of guidance amid the vast virtual marketplace. Particularly in the realm of women's clothing, these reviews act as vital sources of information, offering insights into product quality, sizing, and overall customer satisfaction. As more consumers rely on these reviews to inform their purchasing decisions, the challenge lies in comprehensively understanding the sentiments expressed within these diverse and often nuanced opinions.

Sentiment Analysis

Sentiment Analysis is a use case of Natural Language Processing (NLP) and comes under the category of text classification. To put it simply, Sentiment Analysis involves classifying a text into various sentiments, such as positive or negative,etc. Thus, the ultimate goal of sentiment analysis is to decipher the underlying mood, emotion, or sentiment of a text. This is also known as Opinion Mining.

Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.



Sentiment analysis Approaches:

Rule-Based Approach: It uses a predefined list of positive and negative words to determine sentiment based on word counts.

Machine Learning Approach: This method involves training models on datasets and doing prediction analysis using techniques like Logistic Regression, Support Vector Machines, and more for sentiment classification.

Neural Network Approach: It utilizes artificial neural networks, inspired by the human brain, to classify sentiments and can handle sequential data, eg: RNN

Hybrid Approach: This combines rule-based and machine learning methods to achieve higher accuracy by leveraging the strengths of both.

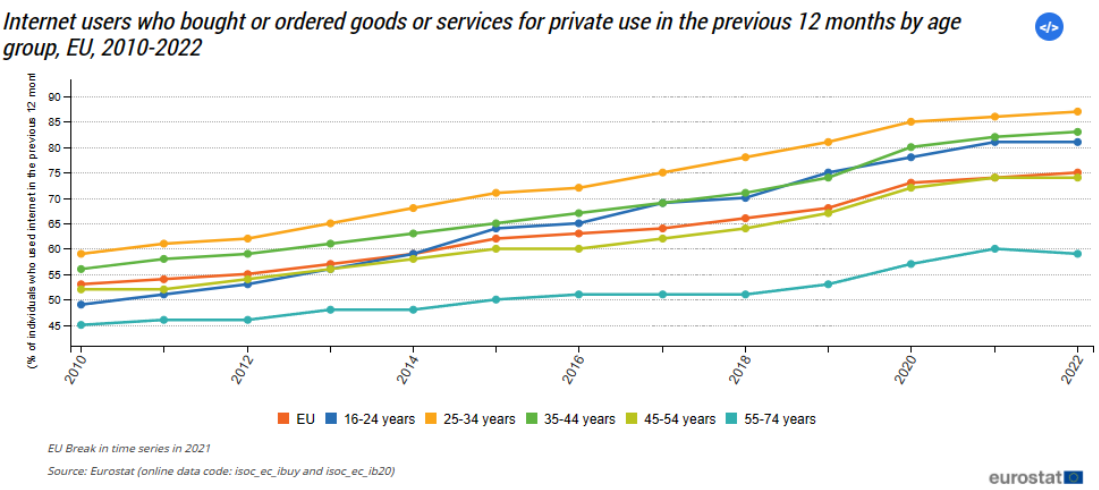
Chapter 3: Literature Review

The proliferation of e-commerce has undeniably transformed the dynamics of the clothing industry. Online retail, particularly in the realm of women's clothing, has seen significant growth, altering the traditional landscape of retail shopping. In this context, customer feedback and product reviews play a pivotal role in shaping the online shopping experience. Understanding the sentiment expressed in these reviews is vital for both businesses and consumers. This literature review explores the background and previous research on sentiment analysis in e-commerce, with a specific focus on women's clothing.

Size and Scale :
According to the 2019 Businesswire report, the global clothing and apparel market reached a value of nearly $758.4 billion in 2018, having grown at a compound annual growth rate (CAGR) of 7.5% since 2014, and is expected to grow at a CAGR of 11.8% to nearly $1,182.9 billion by 2022. There are three major segments of the clothing market: women, men and children. With the women's clothing segment having the biggest market share of 55.7%

Fashion Trends :
Fast Fashion trends fueled by social media and cheap e-commerce clothing have led to weekly trends in the fashion industry. The fast fashion industry is $106 billion Catering to these trends requires fast identification of customer demands. Some rising trends in the clothing industry are: Atheliesurewear, Sustainable and ethical, Fast fashion and Genderless.

Age:
According to an EU e-commerce report on data between 2010-22, The individuals of the age groups 16-24 years and 25-54 years filled in the demand of online clothing with the proportion of 51 % and 49 % respectively of online buyers in 2022. Although it is important to notice the difference in the size of these age groups, the proportion of individuals aged 25-54 years were only topped by the group aged 16-24 years in the purchase of the following items: clothes, deliveries from restaurants, and musical articles (CDs and vinyl, etc.). For all the remaining products, the proportion of individuals aged 25-54 years was the highest, except for the items related to computers, tablets and films or series, where the percentage was the same as for the age group 16-24 years.



Challenges in Sentiment Analysis:
The complexities of understanding consumer sentiment, nuances in language, and varying scales of positivity or negativity make the task inherently intricate. Researchers and businesses need robust methodologies and tools to extract valuable insights from these reviews.

Chapter 4: Basic Terminologies

1. **Sentiment Analysis**: The process of determining the sentiment or emotional tone expressed in a piece of text, often categorized as positive, negative, or neutral.

2. **User Interface (UI):** The graphical interface through which users interact with the system, which may include features like sentiment analysis results display.

3. **Dashboard:** A visual representation of data, often with charts and graphs, to provide an overview of insights from the analysis.

4. **E-commerce:** E-commerce, short for electronic commerce, refers to the buying and selling of goods and services over the internet. In the context of your project, it relates to the online retail environment where consumers purchase clothing products.

5. **Clothing:** Clothing pertains to garments and attire that people wear for various purposes, such as fashion, protection, or modesty. Your project focuses on analyzing reviews related to clothing products in the e-commerce domain.

6. **Analysis:** Analysis involves the examination and interpretation of data to extract meaningful insights or draw conclusions. In our project, sentiment analysis refers to the process of analyzing reviews to determine the sentiment expressed, whether it's positive, negative, or neutral.

7. **Classification:** categorizing reviews into predefined groups, such as "positive" or "negative," based on the sentiment expressed in the text. It's a key task in sentiment analysis.

Chapter 5: Methodology

5.1 Selection of OS

Microsoft Windows was used for this project because it is user friendly & it's robust. The code works on all OS.

5.2 Tool Used

- **Python**

Python is a high-level, general-purpose programming language that is widely used for web development, data analysis, scientific computing, and many other purposes. It is known for its simplicity, readability, and flexibility, as well as its large and active developer community. Some of the key features of Python include A large standard library that supports many common programming tasks, such as connecting to web servers, reading and writing files, and working with data, An interactive interpreter, which allows you to try out code snippets and experiment with the language in an interactive environment, Support for object-oriented, imperative, and functional programming styles, Dynamically-typed, which means that you don't have to specify the data type of a variable when you declare it, Cross-platform compatibility, which means that Python programs can run on multiple operating systems.

- **Tableau**

Tableau is a powerful data visualization and business intelligence software that enables users to transform complex datasets into interactive and easily understandable visualizations. It's widely used for creating informative dashboards and reports, making data-driven decisions, and sharing insights with a non-technical audience. Tableau connects to various data sources, facilitating data analysis and exploration, and is

employed in diverse industries, including finance, healthcare, and marketing, to gain actionable insights from data.

5.3 Libraries Used

- **Numpy:**

  NumPy is a library for the Python programming language that is used for scientific computing and data analysis. It provides functions and utilities for working with large, multi-dimensional arrays and matrices of numerical data, as well as for performing mathematical operations on these data. One of the main advantages of NumPy is that it is very efficient for performing operations on large arrays and matrices, as it is implemented in C and can make use of the performance of compiled code.

- **Pandas:**

  Pandas is a library for the Python programming language that is used for data manipulation and analysis. It provides functions and utilities for working with tabular data, such as data stored in a spreadsheet or a database table. One of the main advantages of Pandas is that it provides a high-level interface for working with data, making it easy to perform tasks such as filtering, aggregating, and transforming data. It also integrates well with other libraries for data analysis, such as NumPy and Matplotlib.

- **Scikit-learn:**

  sklearn is a popular open-source machine learning library for Python, offering a versatile set of tools and algorithms for a wide array of machine learning tasks. It is renowned for its user-friendly interface, making it accessible to both novice and experienced data scientists. Scikit-learn features an extensive collection of machine learning algorithms, tools for data preprocessing, model evaluation, and seamless integration with other Python libraries like NumPy and SciPy. It is actively maintained and has a vibrant community, ensuring it stays up-to-date and well-supported for diverse machine learning and data science projects.

- **Keras:**

  Keras is a high-level, open-source neural networks library written in Python. It serves as an interface for artificial neural network development, allowing users to efficiently build, train, and evaluate deep learning models. Keras is renowned for its simplicity and user-friendliness, making it an excellent choice for both beginners and experienced machine learning practitioners. The library offers an extensive set of pre-processing tools, optimizers, and activation functions, streamlining the process of designing and training neural networks.

- **Pickle:**

  Pickle is a Python module used for serializing and deserializing Python objects. It allows you to convert complex data structures into a binary format for easy storage and transmission. While it's convenient and readily available in Python installations, it's essential to exercise caution when deserializing data, as Pickle can execute arbitrary code, potentially posing security risks. Pickle is commonly used for tasks like saving and loading machine learning models, caching data, and persisting complex data structures.

- **Gradio:**

  Gradio is a Python library that simplifies the creation of web-based interfaces for machine learning models. It offers an easy-to-use API, supports various input and output data types, integrates with popular machine learning libraries, and allows for customization. Gradio enables developers to quickly share their models as interactive web applications, making it a handy tool for showcasing and distributing machine learning projects.

- **TensorFlow:**

  TensorFlow is a popular open-source machine learning framework developed by Google.TensorFlow is used for developing and deploying machine learning models, with a focus on deep learning. It's employed in various applications, including image and speech recognition, natural language processing, recommendation systems, and more. TensorFlow is used in research, industry, and academia to build and train advanced artificial intelligence models for a wide range of tasks.

- **Matplotlib:**

  Matplotlib is a Python library renowned for its ability to generate diverse static, animated, and interactive visualizations, making it an invaluable tool for data scientists, researchers, and analysts. This versatile library enables the creation of charts, plots, graphs, and more, with extensive customization options for controlling visual aesthetics. Matplotlib is used across various domains, including data analysis, scientific research, engineering, finance, academia, and the development of web-based dashboards, helping professionals and educators to visualize and convey data effectively.

- **Seaborn:**

  Seaborn is a Python data visualization library built on top of Matplotlib, designed to make creating informative and attractive statistical graphics more straightforward. Seaborn is known for its ability to generate complex visualizations with concise code, making it a valuable tool for data analysts and researchers.

- **Plotly:**

  Plotly is a dynamic Python library renowned for its ability to create interactive and web-based data visualizations. It excels in generating visually engaging and interactive charts and dashboards that can be explored and manipulated by users. Plotly finds applications in various domains, including business intelligence, data analytics, data journalism, finance, healthcare, and scientific research, where interactive and dynamic visualizations are pivotal for data exploration, analysis, and storytelling.

- **Imblearn:**

  The imbalanced-learn (imblearn) library is a Python package designed to address the challenges posed by imbalanced datasets in machine learning. It provides a collection of techniques for mitigating class imbalance, including oversampling, undersampling, and generating synthetic data. Imblearn is valuable for improving model performance when dealing with datasets where one class significantly outnumbers another, as it helps in achieving more accurate and equitable predictions in imbalanced classification problems.

- **VaderSentiment**

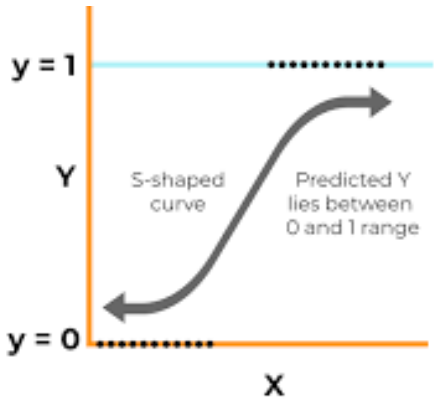
  VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is fully open-sourced and available as a Python library.
  VADER works by assigning a sentiment score to each word in a given text, based on a lexicon of sentiment-related words and a set of rules that account for factors such as capitalization, punctuation, and negation. The sentiment scores for individual words are then aggregated to produce a composite sentiment score for the entire text.
  VADER is particularly well-suited for sentiment analysis of social media text, as it is

able to handle informal language and slang. It is also able to capture sentiment that is expressed in subtle ways, such as through the use of emojis or sarcasm.

5.4 Algorithms Used

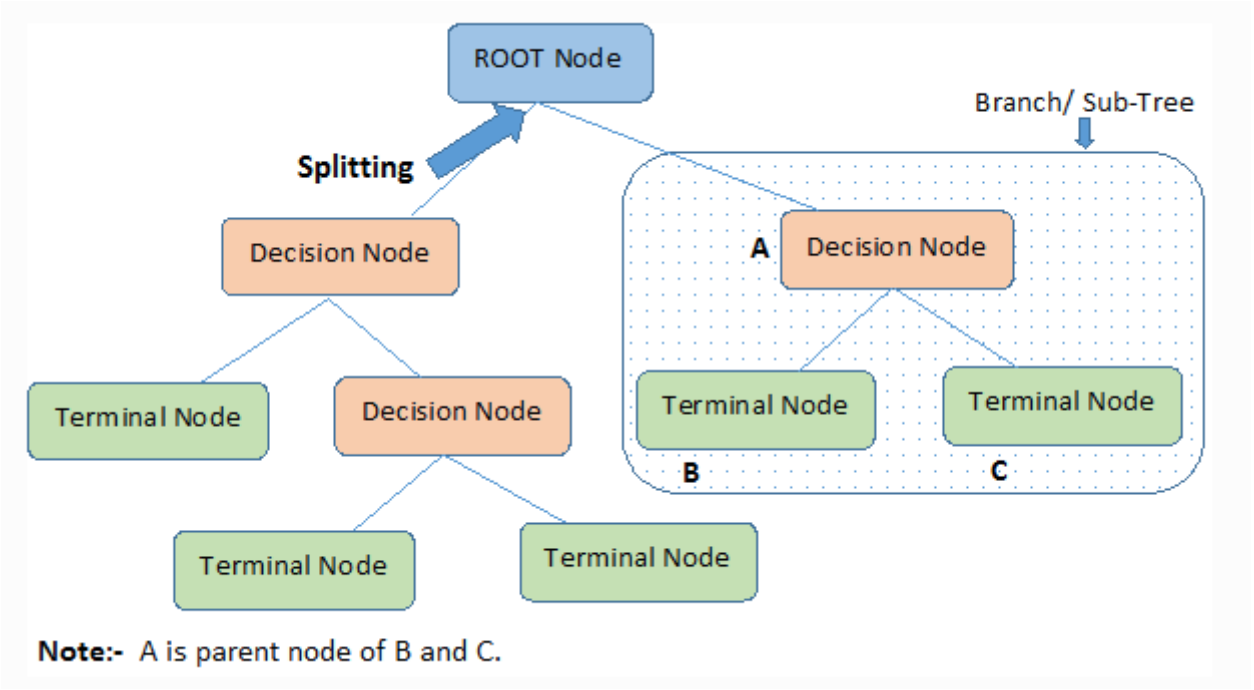
- **Logistic Regression**

  Logistic Regression is a machine learning algorithm used for binary classification tasks. It predicts the probability of an input belonging to one of two classes. It's simple, yet effective, and is widely employed in various fields for tasks like disease prediction and sentiment analysis.



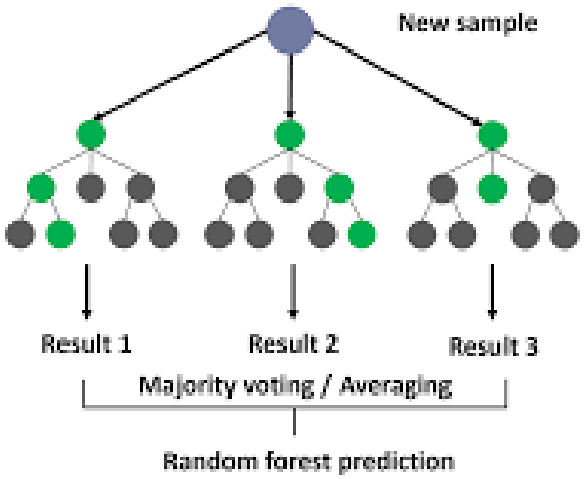
- **Decision Tree**

  Decision Tree is a versatile machine learning algorithm used for both classification and regression tasks. It operates by recursively splitting data into subsets based on the most significant attributes, ultimately forming a tree-like structure of decisions. Each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class label (in classification) or a numerical value (in regression). Decision trees are easy to understand and interpret, making them valuable for gaining insights into the factors influencing predictions. They are employed in various fields, including finance, healthcare, and recommendation systems, and can serve as the foundation for more complex ensemble methods like Random Forest and Gradient Boosting.



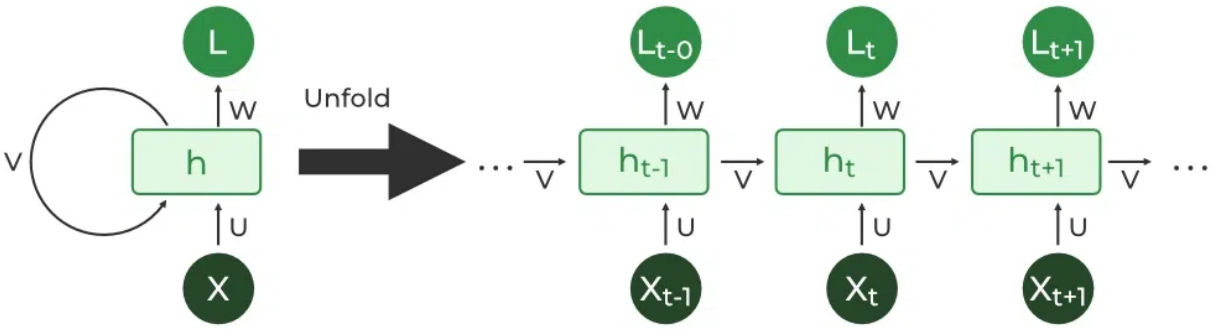
- **Random Forest**

  Random Forest is a powerful ensemble learning algorithm that builds multiple decision trees during training and combines their predictions to improve accuracy and reduce overfitting. It excels in both classification and regression tasks, making it a versatile choice for a wide range of machine learning problems. Random Forest introduces randomness by selecting a random subset of features and data points for each tree, which enhances the diversity of the individual trees in the ensemble. This diversity contributes to the model's robustness and generalization performance. Random Forest is widely used for tasks such as image classification, anomaly detection, and feature selection, thanks to its strong predictive capabilities and ability to handle high-dimensional data.



- **RNN**

  Recurrent Neural Network (RNN) is a type of artificial neural network designed for processing sequential data, making it suitable for tasks where context and order matter. RNNs have an internal memory that allows them to maintain information about previous inputs, enabling them to capture dependencies over time. They are widely used in natural language processing, speech recognition, and time series analysis. However, RNNs can suffer from vanishing gradient problems, limiting their ability to capture long-term dependencies. This limitation has led to the development of more advanced RNN architectures, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), which mitigate the vanishing gradient issue and have become popular choices for sequential data processing.



5.5 Model Training

- **Logistic Regression**

3. Logistic Regression

```
{'newton-cholesky':83, 'lbfgs', 'newton-cg', 'sag':83, 'saga':83, 'liblinear'}
```

Python

```
from sklearn.linear_model import LogisticRegression
logistic_classifier = LogisticRegression(C=0.001, penalty='l1', solver='liblinear')  # You can choose a different solver
logistic_classifier.fit(X_train_scaled, y_train)
```

[93] Python

```
LogisticRegression
LogisticRegression(C=0.001, penalty='l1', solver='liblinear')
```

```
Eval(logistic_classifier, X_test_scaled, y_test)
```

[94] Python

```
0.8703054581872809
```

- **Decision Tree**

4. Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
decision_tree_classifier = DecisionTreeClassifier(max_depth=1)
decision_tree_classifier.fit(X_train, y_train)
```

[53] Python

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=1)
```

```
Eval(decision_tree_classifier, X_test_scaled, y_test)
```

[54] Python

```
c:\Users\eon8w\anaconda3\envs\rl\Lib\site-packages\sklearn\base.py:465: UserWarning: X does not have valid feature names, but DecisionTreeClassifier wa
  warnings.warn(
0.8703054581872809
```

- **Random Forest**

1. Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
model_f = RandomForestClassifier(n_estimators = 40, max_depth=20, min_samples_split=10)
Forest = model_f.fit(X_train_scaled, y_train)
```

[43] Python

```
Eval(Forest, X_test_scaled, y_test)
```

[44] Python

```
0.8708062093139709
```

- **RNN**

```
# Build an RNN model with LSTM
model = Sequential()
model.add(Embedding(input_dim=max_words, output_dim=128, input_length=max_sequence_length))
model.add(LSTM(128))
model.add(Dense(1, activation='sigmoid'))

# Compile the model
model.compile(loss='binary_crossentropy', optimizer=Adam(learning_rate=0.001), metrics=['accuracy'])

# Train the model
model.fit(X_train_padded, y_train, batch_size=64, epochs=5, validation_split=0.2)

# Evaluate the model
y_pred = (model.predict(X_test_padded) > 0.5).astype(int)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# You can also print a classification report for more detailed metrics
print(classification_report(y_test, y_pred))
```

```
Epoch 1/5
91/91 [==============================] - 49s 452ms/step - loss: 0.5919 - accuracy: 0.7174 - val_loss: 0.4339 - val_accuracy: 0.8125
Epoch 2/5
```

5.6 Model Evaluation

Accuracy: Measures the proportion of correct predictions ((True Positives + True Negatives) / Total Predictions).

Precision: Evaluates the accuracy of positive predictions, minimizing false positives (True Positives / (True Positives + False Positives)).

Recall: Assesses the ability to identify all relevant instances, minimizing false negatives (True Positives / (True Positives + False Negatives)).

F1 Score: Strikes a balance between precision and recall, using their harmonic mean (2 * (Precision * Recall) / (Precision + Recall)).

Chapter 6: Data Analysis

6.1 Data Collection

The dataset was collected from Kaggle, a well-known online platform for data science and machine learning resources.

6.2 About Data

	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

- **Details Of Columns In Data:**
  - **Clothing ID:** Integer Categorical variable that refers to the specific piece being reviewed.
  - **Age:** Positive Integer variable of the reviewers age.
  - **Title:** String variable for the title of the review.
  - **Review Text:** String variable for the review body.
  - **Rating:** Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
  - **Recommended IND:** Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
  - **Positive Feedback Count:** Positive Integer documenting the number of other customers who found this review positive.
  - **Division Name:** Categorical name of the product high level division.
  - **Department Name:** Categorical name of the product department name.
  - **Class Name:** Categorical name of the product class name.

6.2 Data Preprocessing

- **Dropping NA values**: Eliminating missing or incomplete data points from the dataset.
- **One hot encoding:** Converting categorical variables into binary (0 or 1) format for machine learning models.
- **Text vectorization:** Converting text data into numerical format for analysis and modeling.
- **Under-sampling:** Balancing imbalanced datasets by reducing the number of samples in the majority class.

- **Standard Scaling:** Transforming numerical features to have a mean of 0 and a standard deviation of 1 for consistent feature scaling.
- **Train-test split:** Dividing the dataset into two parts for model training (train) and evaluation (test) to assess its performance.

```
[13]
no_text = data.drop(columns=['Class Name','Department Name','Division Name','Review Text' ])
```
Python

```
from sklearn.feature_extraction.text import TfidfVectorizer
# Create a TF-IDF vectorizer
tfidf_vectorizer = TfidfVectorizer(max_features=1000)  # Adjust max_features as needed

# Fit and transform the 'review' column
tfidf_vectors_review = tfidf_vectorizer.fit_transform(data['Review Text'])
#tfidf_title = tfidf_vectorizer.fit_transform(data['Title'])

# Create a DataFrame from the TF-IDF vectors
review  = pd.DataFrame(tfidf_vectors_review.toarray(), columns=tfidf_vectorizer.get_feature_names_out())
#title = pd.DataFrame(tfidf_title.toarray(), columns=tfidf_vectorizer.get_feature_names_out())
# Concatenate the TF-IDF DataFrame with the original DataFrame
#all_data = pd.concat([no_text, review, title ], axis=1)
all_data = pd.concat([no_text, review ], axis=1)
```
[14] Python

Undersampling Function

```
from imblearn.under_sampling import RandomUnderSampler

def undersample(df, target  ):

    X = df.drop(target, axis=1)
    y = df[target]

    # Define the undersampler
    undersampler = RandomUnderSampler(sampling_strategy='auto', random_state=42)

    # Apply undersampling to balance the classes
    X_resampled, y_resampled = undersampler.fit_resample(X, y)

    # Create a new DataFrame with the resampled data
    #resampled_df = pd.concat([X_resampled, y_resampled], axis=1)

    return  X_resampled, y_resampled
```
Python

Scaling

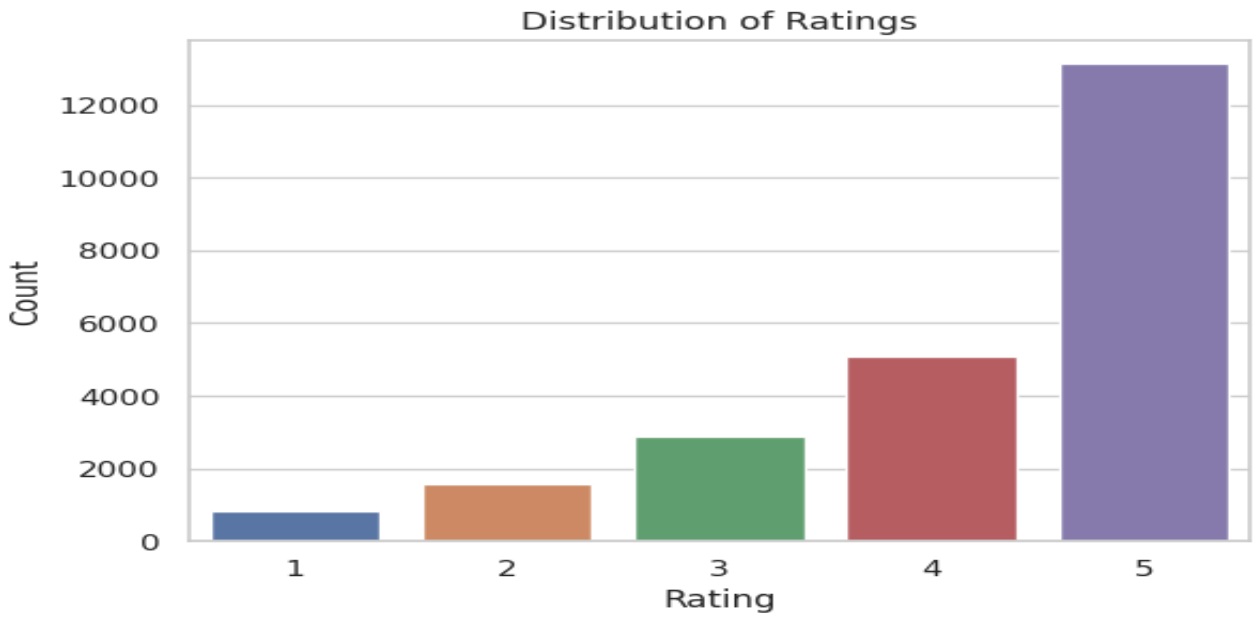
```
#Standard Scaling
from sklearn.preprocessing import StandardScaler

# Create the StandardScaler
scaler = StandardScaler()

# Fit the scaler on the training data and transform it
X_train_scaled = scaler.fit_transform(X_train)

# Use the same scaler to transform the test data
X_test_scaled = scaler.transform(X_test)
```
Python

Evaluation Function

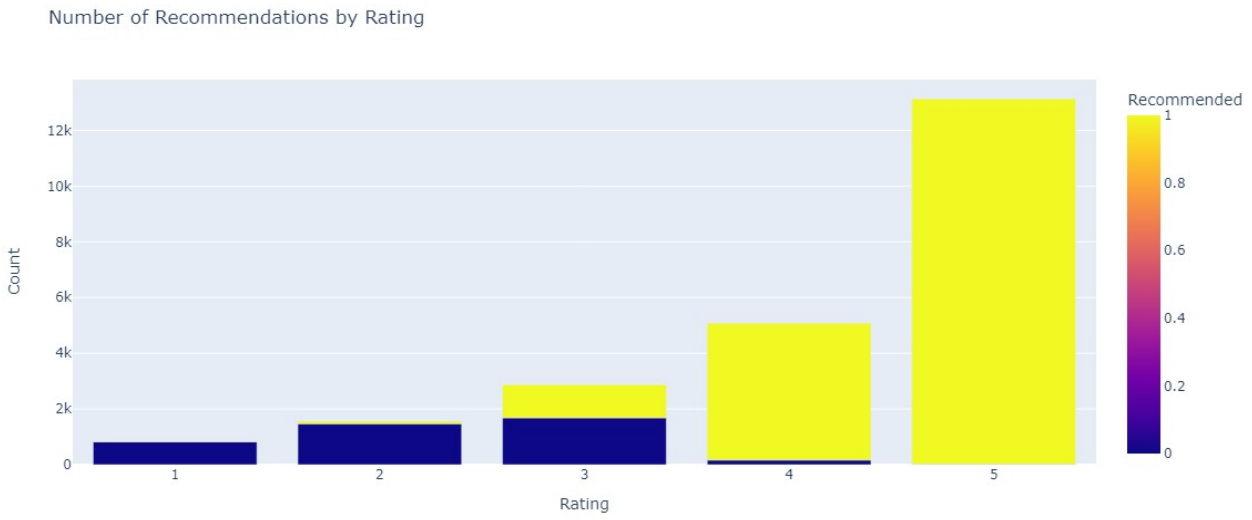
```
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
import seaborn as sns
```

6.3 Data Insights

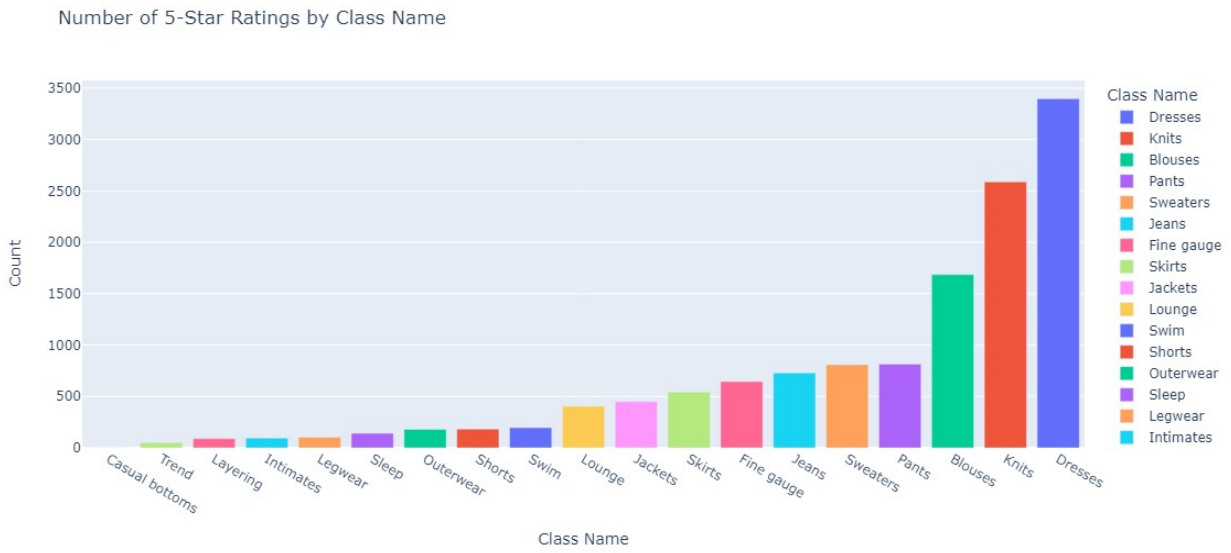
- Looking at the rating distribution chart, it's evident that there are more 5-star ratings than any other rating category in the data.



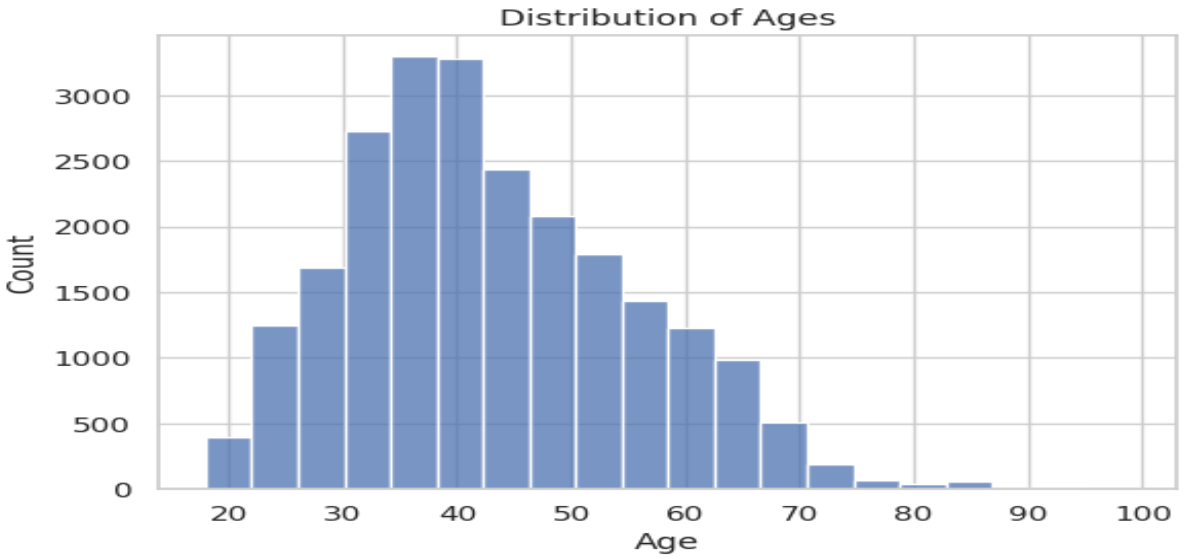
- This chart indicates that customers who give higher ratings are also more likely to recommend the product.



- This chart highlights the top products rated with 5 stars across all classes providing valuable information for making decisions about sales and supplies.



- A significant portion of the customers falls within the age group of 30-40 years old.



Chapter 7: Result & Discussion

7.1 Model Performance

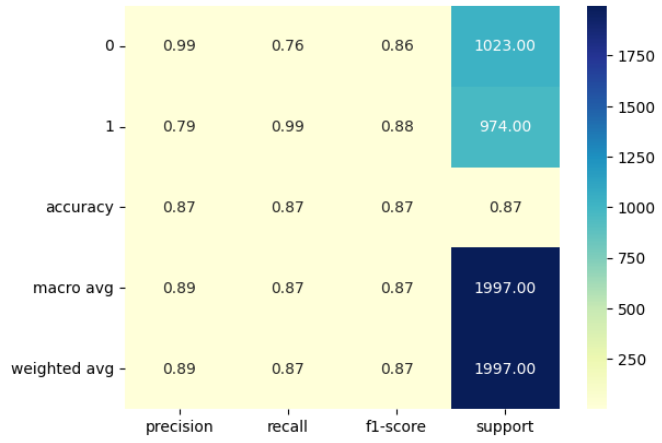
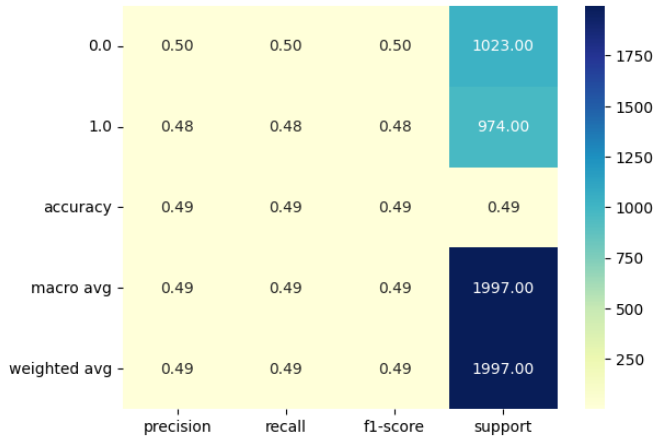
7.1.1 Model Accuracy

Model	Accuracy (all data)	Accuracy (review text)
Random Forest	0.870305	0.491737
Decision Tree	0.870305	0.506259
Logistic Regression	0.870305	0.512268
RNN	0.473684	0.841085

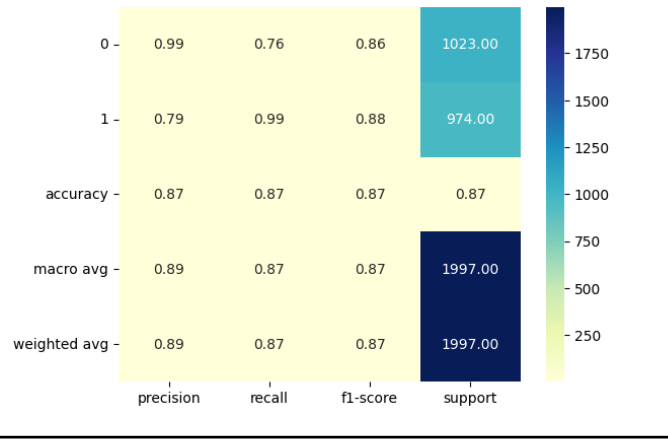
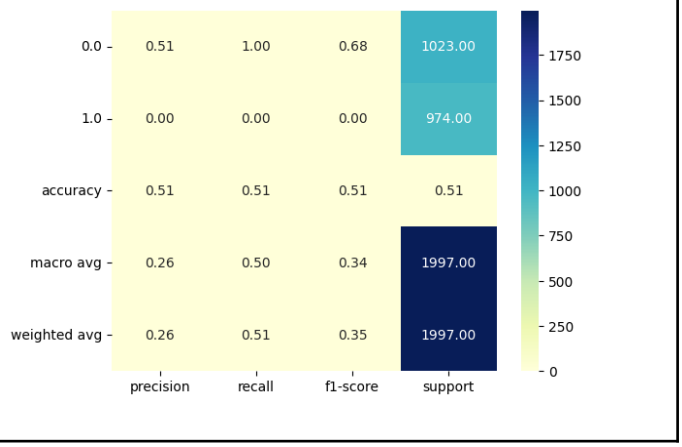
7.1.2 Classification Report

Following classification report is calculated on test data. Model was trained on all data and review text.

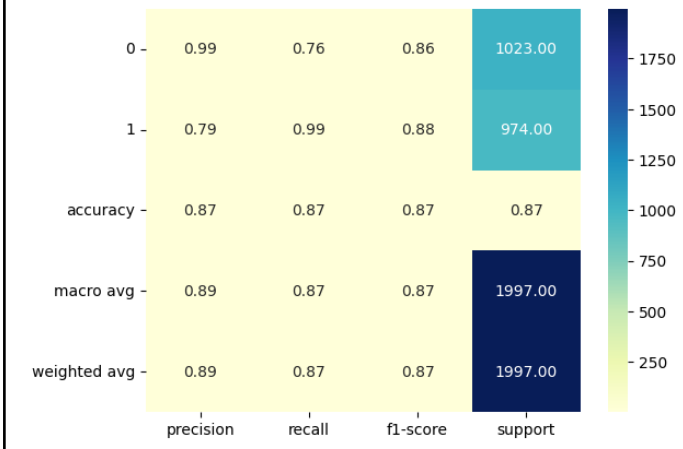
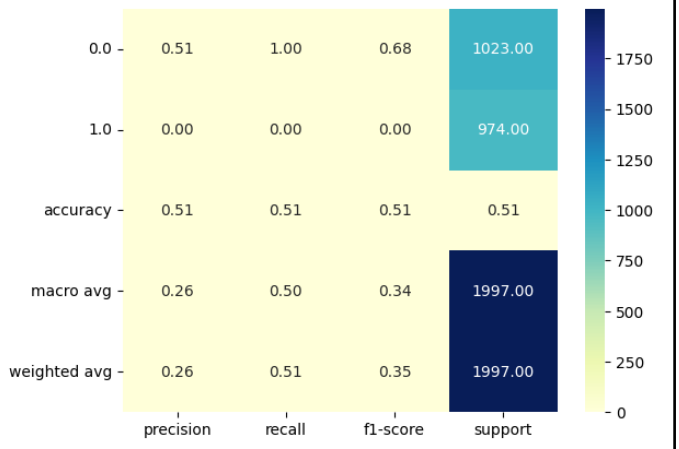
7.1.2.1 Random Forest

All Data	Review Text
	

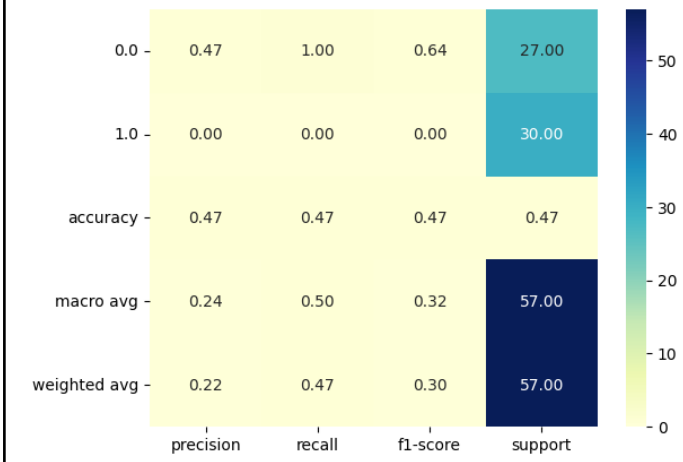
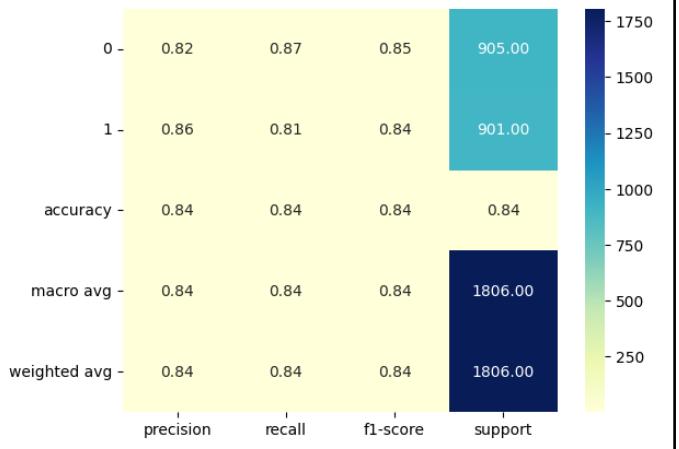
7.1.2.2 Decision Tree

All Data	Review Text
	

7.1.2.3 Logistic Regression

All Data	Review Text
	

7.1.2.4 RNN

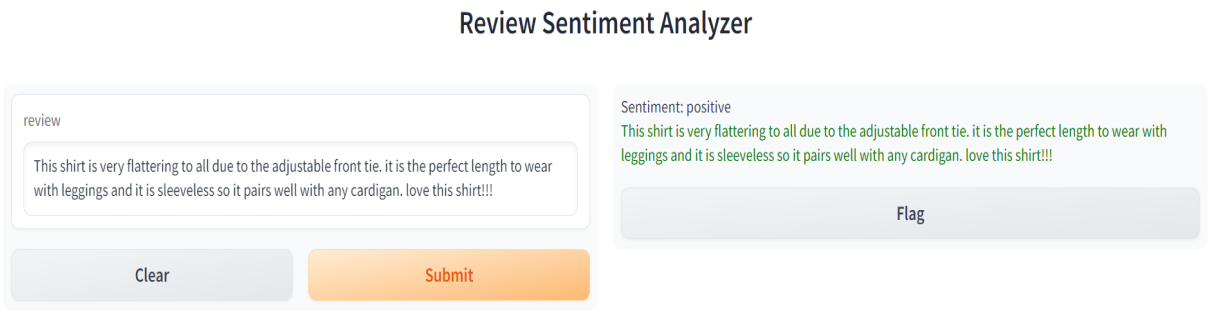
All Data	Review Text
	

7.2 GUI

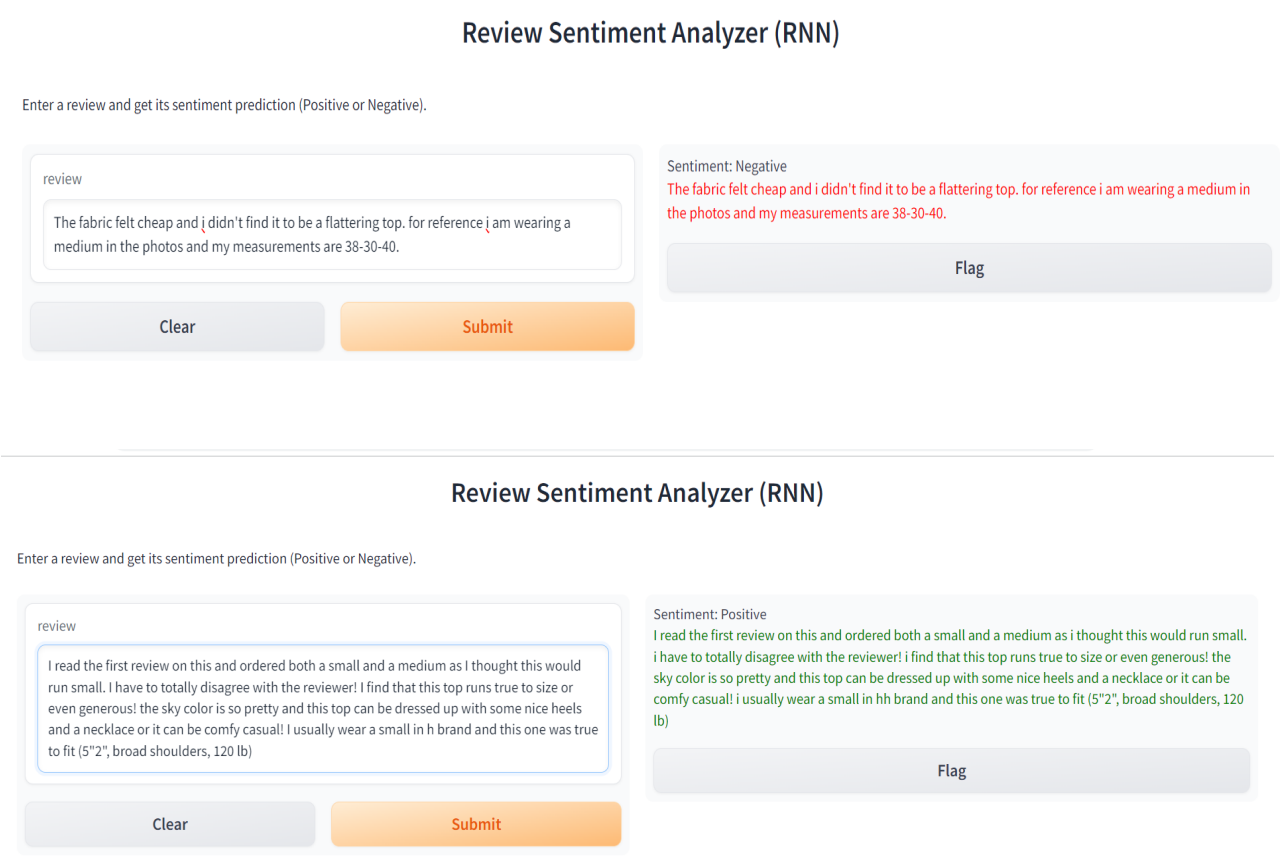
In order to provide an intuitive interface for sentiment analysis, we employed the Gradio library to design and develop a Graphical User Interface (GUI). This GUI allows users to input text-based reviews and promptly receive sentiment analysis results. The system classifies the reviews as either 'positive' or 'negative' and takes it a step further by visually representing the results.

Positive reviews are highlighted in a vivid shade of green, while negative reviews are highlighted in a noticeable shade of red.

This visual representation provides a quick and intuitive way for users to gauge the sentiment of the text they input.



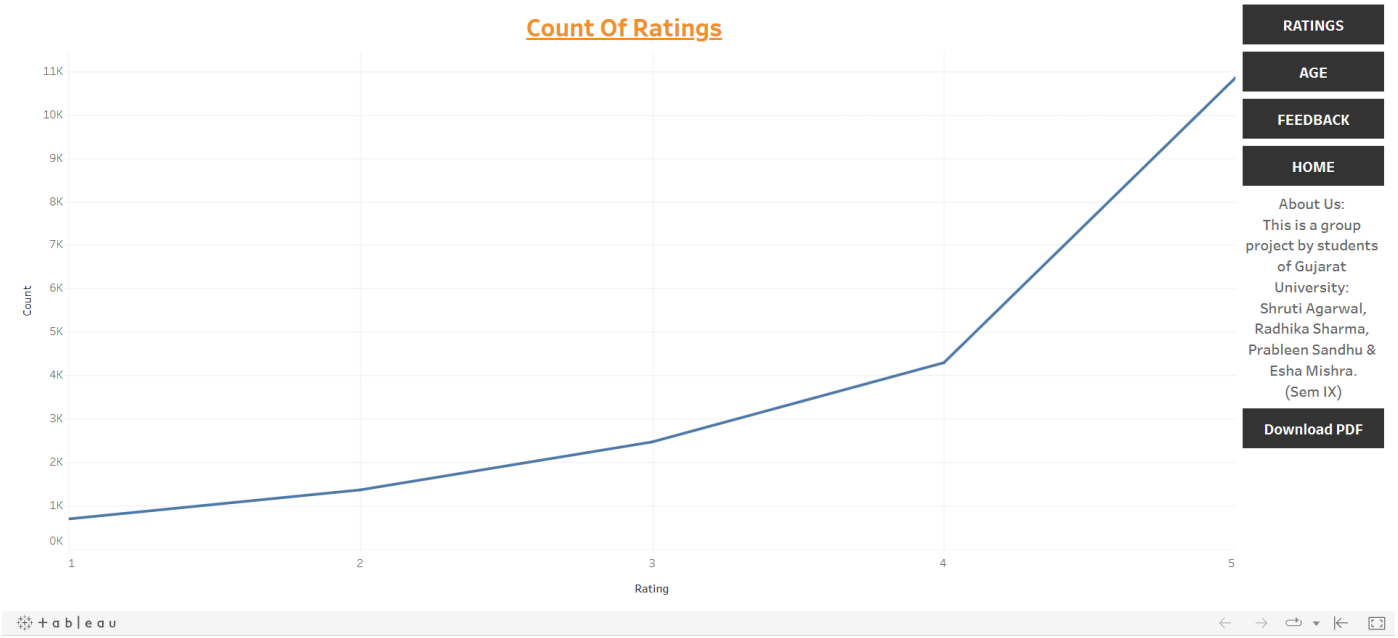
This GUI utilizes the VADER sentiment analyzer, a general-purpose tool, as it was not specifically trained on our dataset.



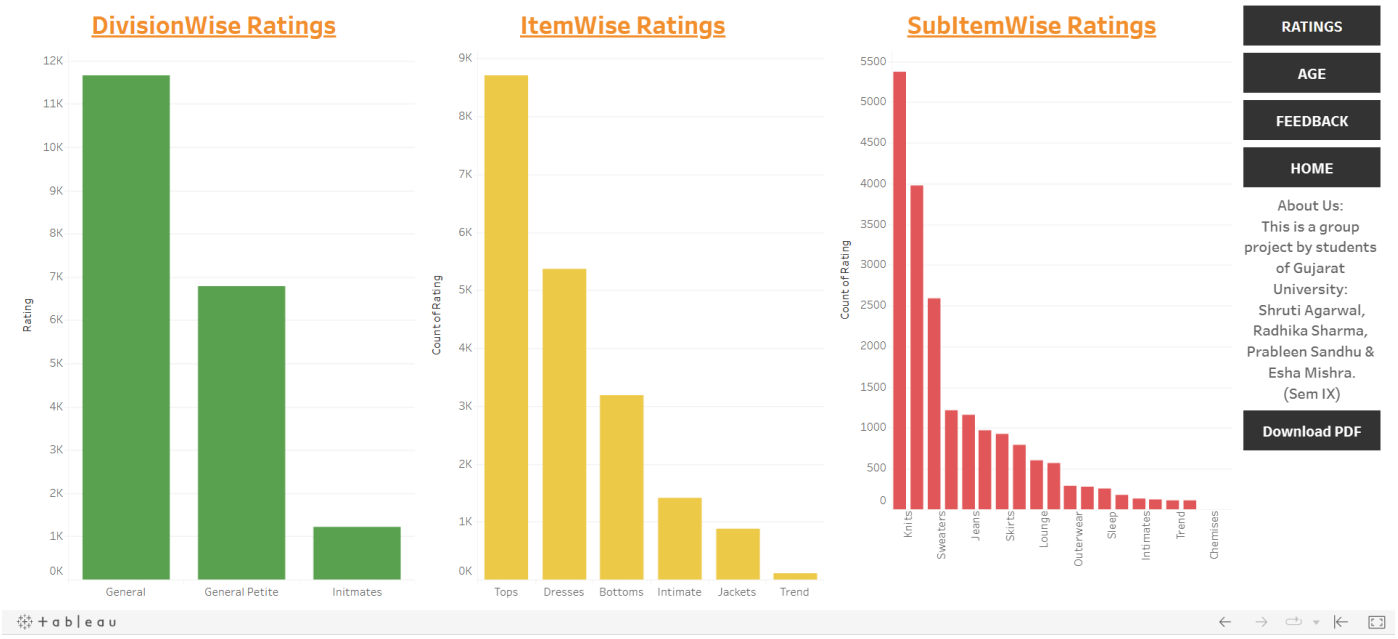
This GUI was developed using an RNN model that was trained on our specific dataset.

7.3 Dashboard

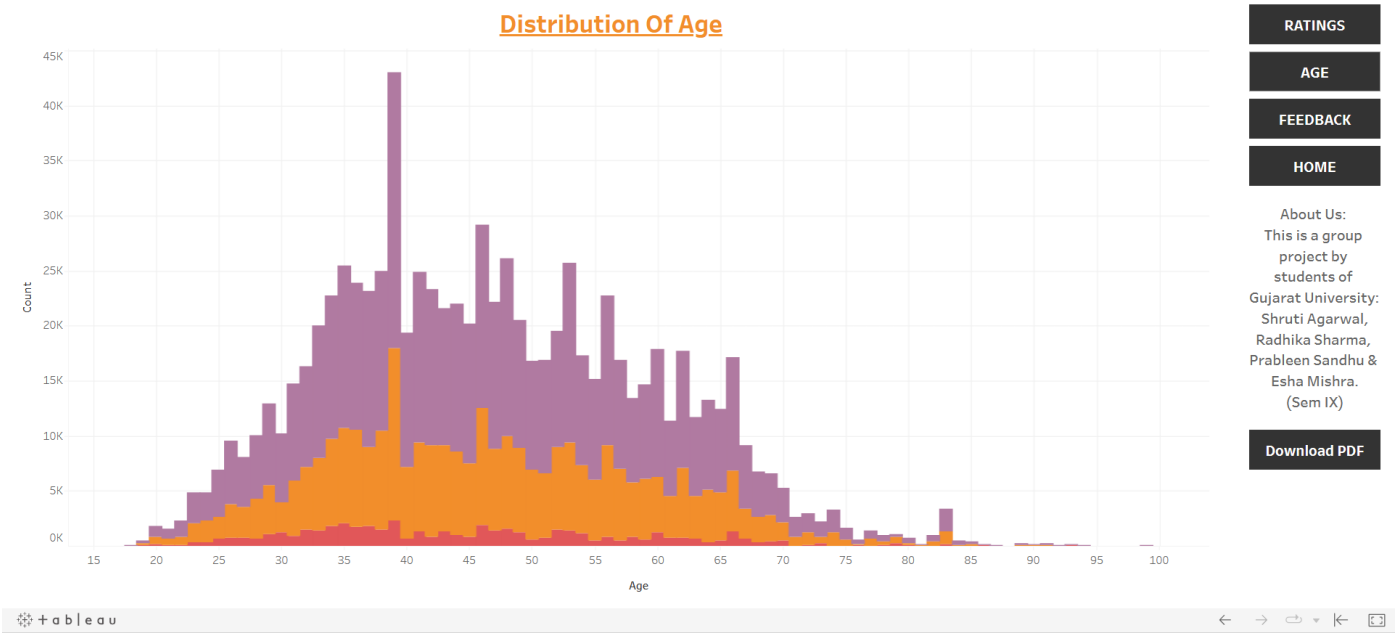
Link to dashboard:
https://public.tableau.com/app/profile/shruti6130/viz/WomensECommerceClothingReviewsAnalysis/HOME?publish=yes



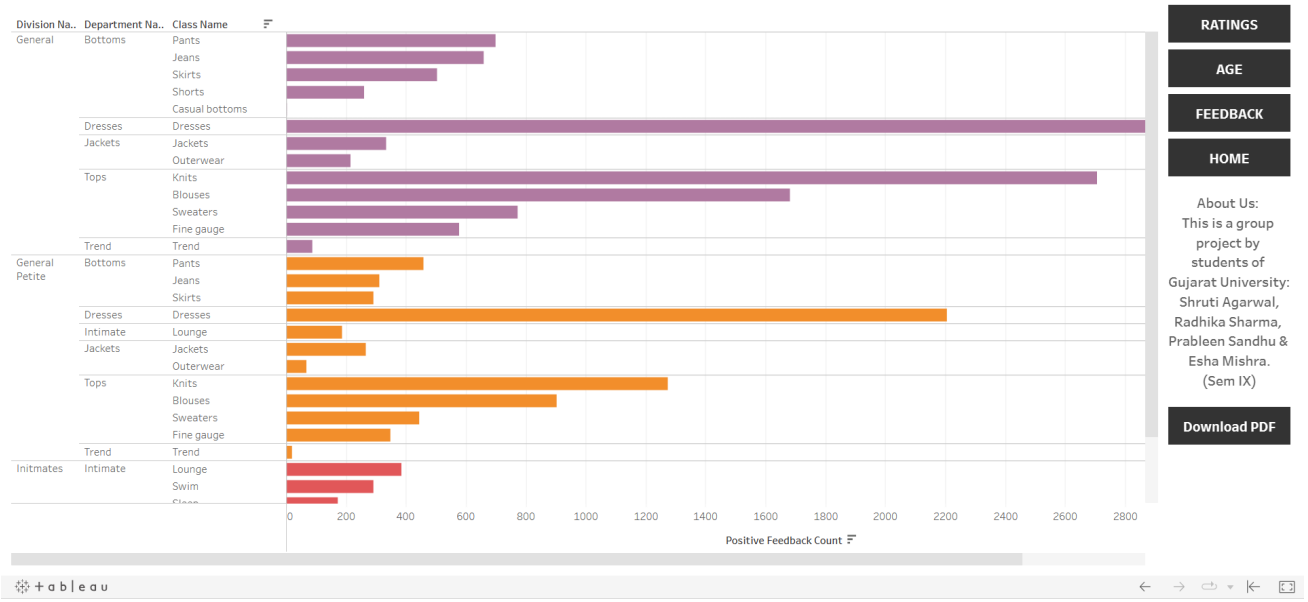
This graph signifies the distribution of ratings within the dataset. It provides a visual representation of how many times each rating level (1 to 5) appears in the data. It helps one understand the overall sentiment or feedback provided by customers, as one can see which ratings are more or less common in the dataset. The above graph shows that the dataset has more of 4 star and 5 star ratings comparatively.



The above chart helps one understand which products & product categories are receiving what ratings from customers. By looking at the heights of the bars, one can quickly identify which categories are most popular among customers, as they have the highest counts of ratings. Conversely, categories with shorter bars have received fewer ratings, suggesting that those products might have room for improvement or are less favored by customers. This information can be valuable for businesses and decision-makers to focus on product categories that are performing well and consider strategies for improving those with lower rating count.
Overall, this chart provides insights into customer satisfaction within different product categories and can inform business decisions related to product quality, marketing, and customer experience improvements.



A histogram is a graphical representation of the distribution of a numerical variable, in this case, "Age." It divides the data into discrete bins and counts the number of data points within each bin. The X-axis represents the age range intervals, and the Y-axis represents the count of data points falling into each interval. The different colors indicate the category of reviews. Where Purple color indicates General, Orange Color indicates General Petite and Red color indicates Intimates.



This shows the graph where positive feedback counts are shown with respect to division, department & class.

Conclusion

The data preprocessing and understanding takes most time when it comes to sentiment analysis like all other ML projects.

Random Forest and Logistic Regression models are practical choices for predicting customer sentiment, while Decision Trees are useful for revealing feature importance.

ML models give decent accuracy of 87%. They perform well when trained on all data including not just review text but also features like clothing type, recommended or not, department etc. But when trained only on review text data the model accuracy drops significantly to around 50%.

Deep learning model, RNN outshines all ML models as it gives accuracy of 84% when trained on only review text.

The insights gained from our analysis can be leveraged by the e-commerce business to prioritize product categories and areas of improvement. It can influence product selection, marketing strategies, and customer experience enhancements.

Platforms powered by ML to understand consumer feedback will have a larger share of the pie of the growing e-commerce market.

The distribution of customer ratings is highly skewed towards positive ratings, with a significant proportion of 5-star reviews.

The "Class Name" feature, representing product categories, plays a pivotal role in customer satisfaction. Some categories consistently receive more 5-star ratings than others, highlighting the importance of product diversity and quality within the e-commerce business.

In conclusion, this project sheds light on the potential of machine learning models in understanding customer behavior and satisfaction in the e-commerce domain.

The findings provide a valuable foundation for data-driven decision-making, allowing businesses to enhance their offerings and customer experiences.

Further research and refinement of models will undoubtedly yield even more actionable insights in the future.

Future Work

- NLP Enhancements:
  Apply advanced Natural Language Processing (NLP) techniques like Named Entity Recognition (NER) to extract specific entities mentioned in reviews (e.g., product names, brand mentions). This can help in identifying popular products and brands.

- Customer Segmentation:
  Segment customers based on their review content, demographics, and purchase history. This can help in tailoring marketing strategies to different customer segments.

- Data Enhancement:
  Gathering & collecting more data from different brands and websites with more columns for more detailed analysis.

- Time-Series Analysis:
  Includes a dataset which has timestamps and performs time-series analysis to understand how customer reviews, ratings, or sentiments change over time. This can help identify trends and seasonal patterns.

- Geospatial Analysis:
  Includes a dataset which has location information and performs geospatial analysis to understand regional variations in product preferences and customer behavior.

Bibliography

Agarap, A. F. (n.d.). *Women's E-Commerce Clothing Reviews*. Kaggle. Retrieved October 12, 2023, from https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

Buisnesswire owned by Berkshire Hathway. (2019, October 25). *Global $1182.9 Billion Clothing and Apparel Market Analysis, Opportunities and Strategies to 2022 - ResearchAndMarkets.com*. Business Wire. Retrieved October 12, 2023, from https://www.businesswire.com/news/home/20191025005178/en/Global-1182.9-Billion-Clothing-and-Apparel-Market-Analysis-Opportunities-and-Strategies-to-2022---ResearchAndMarkets.com

Eurostat. (2023, August 30). *E-commerce statistics for individuals - Statistics Explained*. European Commission. Retrieved October 12, 2023, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=E-commerce_statistics_for_individuals

yieldify. (2023, February 8). *Fashion eCommerce: Top Trends, Stats & Examples for 2023*. Yieldify. Retrieved October 12, 2023, from https://www.yieldify.com/free-guides/fashion-ecommerce-trends/

Jain, S. (2023, April 6). *What is Sentiment Analysis? GeeksforGeeks*. Retrieved October 12, 2023, from https://www.geeksforgeeks.org/what-is-sentiment-analysis/