E-Commerce Review Sentiment Analysis

A. Motivation

E-commerce refers to the process of conducting commercial transactions entirely over the internet. This digital marketplace enables companies to access a worldwide audience, offering consumers the convenience of shopping from anywhere and choosing a vast array of products. In such a digital ecosystem, customer feedback is critical to influencing the purchasing decisions of potential buyers. Therefore, there is significant business value in utilizing data mining techniques to extract meaningful insights from customer feedback.

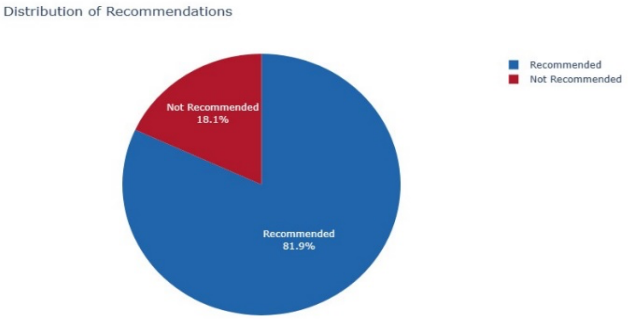
This project is based on the Women's E-Commerce Clothing Reviews dataset provided by Kaggle. We aim to predict the sentiment of reviews and identify potential issues through the insights given from review texts. LSTM will be used as our deep-learning model for sentiment prediction. We will also generate word clouds as visual representations, highlighting commonly used words in reviews, specifically for reviews with low ratings. Through this analysis, an E-commerce retailer can gain valuable insights into which areas require enhancement to boost customer satisfaction and overall ratings.

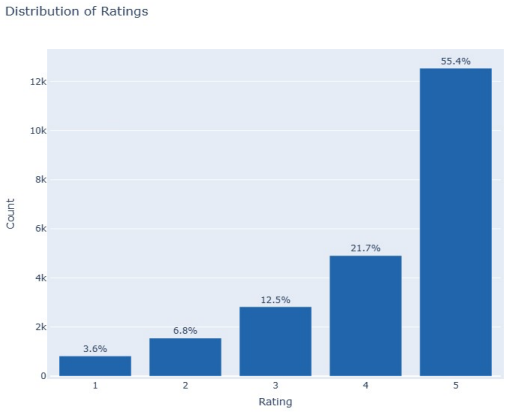
B. Data Understanding

The dataset comprises 23,846 entries and 10 distinct feature variables. The main feature of the dataset is *Review Text*, which encompasses customer reviews supported by additional attributes such as *'Clothing ID', 'Age', 'Title', 'Rating', 'Recommended IND', 'Department Name', and 'Class Name'*. Before data preparation and modeling, we developed visualizations to explore significant aspects of the data:

The average rating across all reviews is approximately 4.2, indicating predominantly positive feedback. This favorable outlook is bolstered by the fact that around 82% of the reviews include a recommendation for the product. The comparison of the two plots indicates that customers will recommend the item when the rating is around 4 to 5.

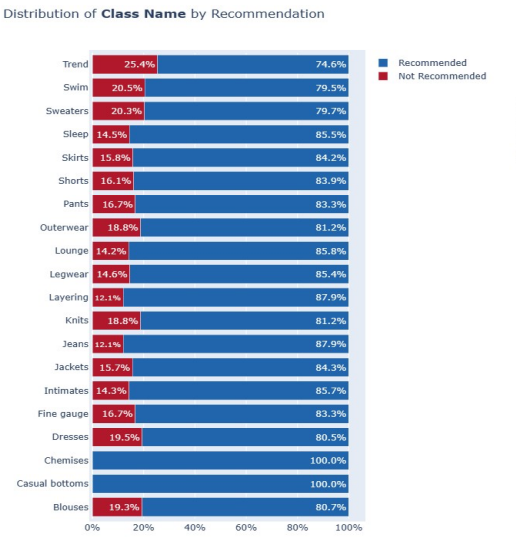
We also explore the distribution of class names by recommendation percentage. The

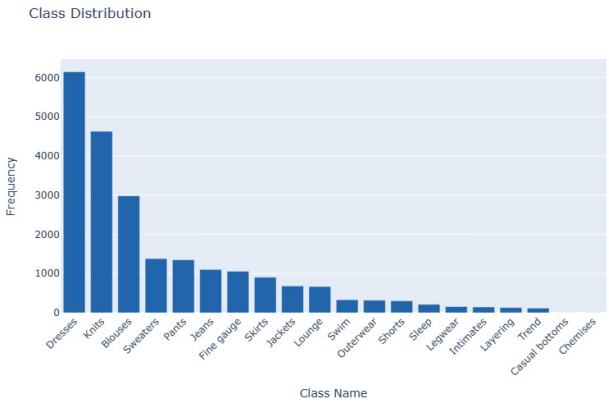




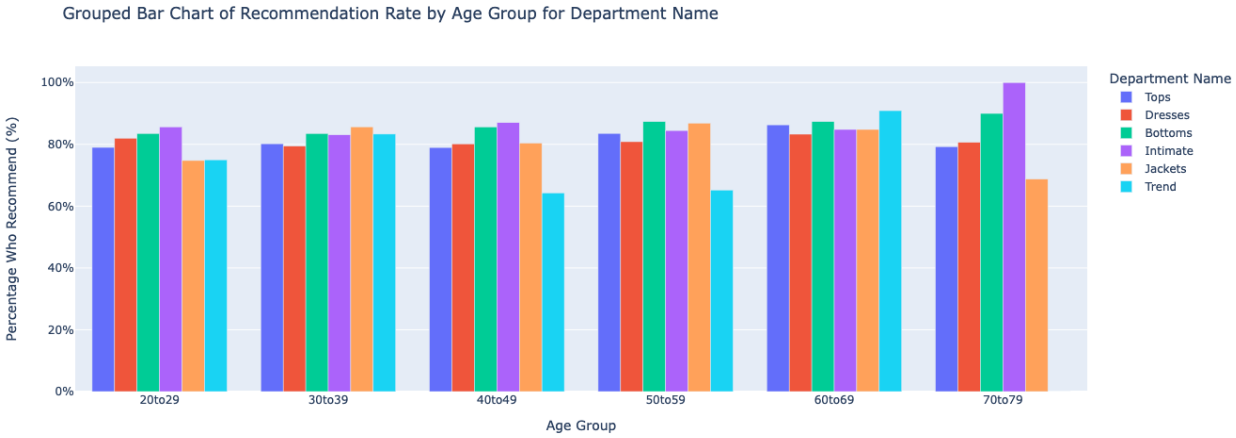
findings are consistent with the overall trend

except for the 'Trend' class, which has a slightly higher un-recommendation rate. Among all categories, 'Dresses', 'Knits', and 'Blouses' garnered the highest number of reviews.





In the last part of data exploration, we examine the recommendation percentage of different age groups among departments. A positive reception of the products is consistent across the range of demographics. However, the trend" category has the lowest rating among age groups 40 to 49 and 50 to 59.



C. Data Preparation

The data preparation can be divided into three parts: Data Cleaning, Text Tokening, and Length Padding.

a. Data Cleaning:

- Removing Duplicates: Remove 21 rows of duplicates to maintain the integrity of the dataset, ensuring that each row provides unique and valuable information.
- Drop Null Values: Drop rows which has empty review text. Since this analysis is largely based on text sentiment, these rows will not be useful for us.
- Add "Text Length" column: Add text length column as a reference for padding length

b. Text Tokenizing and Encoding

- Tokenize each review sentence into word tokens. Remove punctuation and encode the word into integers by sorting word frequency.
- Turn sentences into word vectors.

c. Padding Sequence

- We extend or cut sentences into specific lengths to deal with too short or long reviews. This measure is to ensure we have more standardized and consistent datasets to feed into deep learning networks. As observed in the "text length" visualization, there is a surge of around 100 words. Therefore, we use 100 as our padding sequence length.

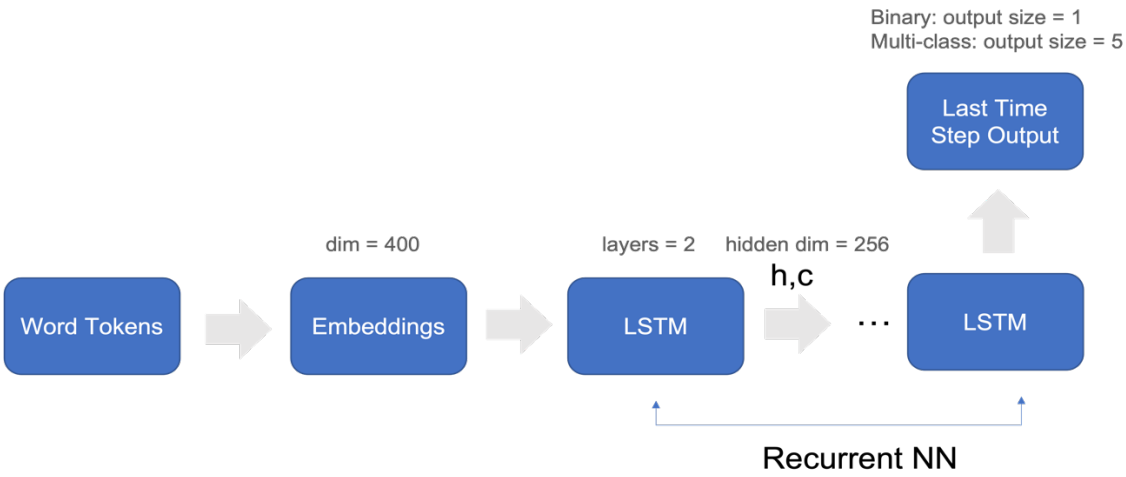
D. Modeling

In this analysis, the LSTM model will be used to make predictions about the sentiment of reviews. With its recurrent neural network, the LSTM model can consider the order of words and memorize the semantic meaning in the hidden state. We believe that the model is suitable for dealing with sentence sequences. Since there are both binary classification factors (Recommend/Unrecommend) and multi-classification factors (Rating from 1 to 5) in the dataset, we try to address both factors by identifying specific predicting outputs and loss functions. Below is the model architecture and hyperparameters:

a. Model Architecture

In the data preparation phase, the text has already been converted into integers. Instead of considering the entire vocabulary size, we can embed the text to condense it into lower dimensions. After that, we add LSTM layers as RNNs to memorize the long-term and short-term semantics of words. Finally, we output the last time step output to predict the

sentiment. For binary classification, since the loss function is BCE, we need to add another sigmoid layer; whereas for multi-class classification we use linear layer output because the cross-entropy loss itself includes the sigmoid transformation. Please refer to the below graph for the process:



b. Hyperparameters

Embeddings, number of layers, hidden dimension, and output size are the parameters used in the models:

1. **LSTM for Binary Classification**
    - Input: Review Text
    - Label: Recommend or Not (0 or 1)
    - Predicting Output: Sigmoid probability between 1 and 1
    - Loss Function: Binary Cross Entropy Loss

Embeddings	**Layers**	**Hidden Dim**	**Output Size**
400	2	256	1

2. **LSTM for Multi-Class Classification**

- Input: Review Text
- Label: Rating (1 to 5)
- Output: Linear layer output, output size = 5
- Loss Function: Cross Entropy Loss

Embeddings	Layers	Hidden Dim	Output Size
400	2	256	5

The alternative models that could also be used in this analysis are GRU networks and Transformer models. With the Transformer model, it can solve the issue of encoding bottleneck and parallelization problems in RNN. However, it requires a large amount of data and more computational resources for training.

E. Implementation

We first run the logistic regression as a baseline model and then build the LSTM model to see how well the model is performing. The implementation includes two parts: hyperparameter tuning and up-sampling:

a. **Hyperparameters Tuning:** One of the challenges we faced was model overfitting. We handled this issue by reducing the embedding dimension from 400 to 100, adding L2 regularization, and reducing epochs as early stops. After tuning, the validation loss curve performs a better-decreasing trend.

b. **Up-sampling:** Another challenge that we encountered during modeling was handling the proportion of positive sentiment. As we see from the visualization, over 80% of the reviews recommended the items, and the imbalanced distribution impacted the model's

ability to learn from diverse data. To address the issue, we tackled it by increasing the sample proportion of unrecommend reviews. After up-sampling, the percentage of recommended and unrecommend reviews is 54.4% to 45.6%, and the testing accuracy for binary classification also increased from 87% to 92%.

F. Results and Evaluation

This analysis is a classification problem, so we use **OOS Accuracy** to evaluate the result. The baseline model (logistic regression) accuracy for binary classification and multi-class classification is 87% and 55.43% respectively. With the implementation of the LSTM model in both cases, the results turned out to be better than the baseline model. Please see the plot below:

a. LSTM for Binary Prediction

- Test Loss: 0.163
- Test Accuracy: 93%

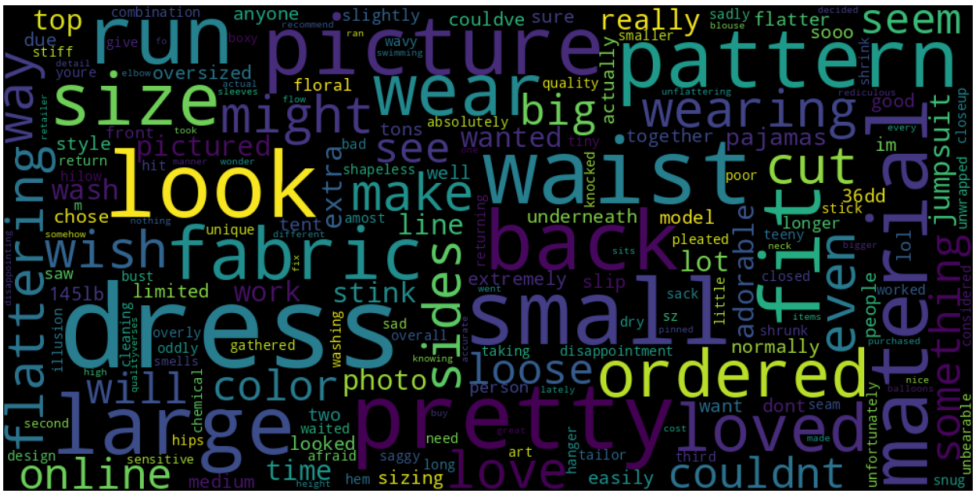
b. LSTM for Multi-Class Prediction

- Test Loss: 0.860

- Test Accuracy: 61.7%

It is clear that the LSTM model more accurately identifies whether customers recommend the item compared to classifying its rating. We noticed that the customer reviews often contain a blend of positive and negative aspects. For example, in a 2-star review, a customer mentioned, "it's soft and fits okay, but it has zero support or shape." We believe that such mixed reviews pose a challenge for the model to classify ratings.

In addition, from the visualization of the Recommendation Rate by Age Group for Department Name, we could see that the “trend” category has the lowest rating among age groups 40 to 49 and 50 to 59. To deep dive into this issue, we use word cloud to identify the most frequently used words in low-rating customer reviews on "trend” items in that group. This approach helps the business get insights from customer feedback and make improvements



accordingly.

"Look", "pattern", "fabric", "picture", "small", and "waist" are the words that appear more frequently than others, indicating that those customers who give lower ratings have concerns regarding size, material quality, and correspondence with product images. We suggest that the business could focus on those areas to enhance customer ratings.

E. Deployment

With data mining techniques such as the LSTM model, businesses can gain insights from customers by identifying the sentiment of their feedback. Word clouds also allow companies to delve deeper into specific categories such as age groups and departments.

From the positive reviews, a word cloud reveals key terms like "Color", "Look", "Fit", "Dress", "Perfect", and "Soft", suggesting that customers particularly value the design and sizing of our clothing, especially our dresses.

Conversely, the word cloud for negative reviews highlights words like "Fabric", "Material", "Ordered", "Retailer", and "Dress". This indicates a need for improvement in product quality and after-sales service. The frequent mention of "Dress" in negative reviews is

likely due to higher sales volumes in this category, leading to a proportional increase in negative feedback.

However, our current model has limitations, as some reviews contain both positive and negative comments (e.g., “I like the size, but the material is disappointing”). This can lead to misleading results in our word cloud. To address this, we propose that the e-commerce platform should refine its review system to include more specific feedback options. For instance, at the beginning of the review section, a question such as “Why did you not like our product?” could be accompanied by choices like “Quality”, “Size”, “Design”, etc. This would facilitate more accurate analysis and aid in the company’s ongoing development and improvement.

References

1. Winter, Dayna. What Is Ecommerce? A Comprehensive Guide (2024). Shopify.com. May 26,2023.
2. Women's E-Commerce Clothing Reviews. Kaggle.com