UNIVERSITY OF CALIFORNIA

Los Angeles

Determining the Accuracy
# of Consumer's Review to
Guide Purchasing Decision through Sentiment Analysis

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Amanda Leonarto

2024

ABSTRACT OF THE THESIS

Determining the Accuracy
of Consumer's Review to
Guide Purchasing Decision through Sentiment Analysis

by

Amanda Leonarto
Master in Applied Statistics
University of California, Los Angeles, 2024
Professor Frederick Paik Schoenberg, Chair

This study investigates how individual sentiment towards product help another individual purchasing decision. It explores numerous sentiment analysis models such as Ordinal Regression, LSTM, Logistic Regression, Random Forest, and BERT, in predicting customer recommendations and ratings from Sephora product reviews. These findings demonstrate the models' ability to forecast customer recommendations consistently, offering significant insights that can enhance customer satisfaction, marketing initiatives, and inventory management. Despite the promising results, limitations, such as capturing the nuanced nature of evaluations and ensuring model generalizability remain, emphasizing opportunities for future research.

The thesis of Amanda Leonarto is approved.

Nicholas Christou

Ying Nian Wu

Frederick Paik Schoenberg, Committee Chair

University of California, Los Angeles

2024

This thesis is dedicated to my family
and to my best friends who have cheered and
supported me through all the ups and downs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The COVID-19 pandemic acted as an unexpected and powerful stimulus, even a shock for some people, also in the case of those who did not experience any health problems but only the social consequences resulting from lock-downs, social distancing rules, the need to wear masks, etc. The social isolation during the COVID-19 pandemic led to negative psychical states such as chronic stress, anxiety feelings, or mood disorders. In response, individuals turn to compensatory and compulsive buying as means of coping mechanisms to escape from negativity, such as undesired psychical states and problems of everyday life. The lack of healthier coping mechanisms, such as social interactions, physical exercise, and recreational activity, further exacerbated this trend during the pandemic [1].

The COVID-19 pandemic increases social media usage by 20 percent worldwide, enhancing connectivity and facilitating people all over the world to share their ideas, thoughts, and personal lives [3]. The psychological impact of the pandemic and the subsequent rise in social media usage have not only changed consumer behavior but also influenced marketing strategies. Brands now prioritize digital marketing and influencer collaborations to reach their target audiences effectively. The beauty industry, in particular, has adapted by focusing on personalized and relatable content that resonates with consumers' experiences and emotions during the pandemic. This surge has particularly impacted the beauty industry, where social media influencers play a pivotal role. Predominantly female influencers dominate this digital landscape, which attracts a largely female audience. Since most female influencers gravitate towards beauty content, female audiences engage heavily with content related to skincare, makeup, and fashion. As a result, In 2023, companies like Sephora, a leading French multinational retailer in the beauty industry, reported record-breaking revenue of €86.2 billion and profits of €22.8 billion , driven by strong organic growth across most business units [6].

The influence of social media extends to consumer behavior, leading beauty companies to release new products annually to keep up with trends and demands. With social media, influencers's reviews often prompt their followers to immediately purchase products, exemplifying the impact of para-social relationships, where followers feel a personal connection to influencers's and trust their endorsements. However, with vast technologies and a tremendous amount of information, consumers have become more discerning and critical and recognize that beauty products are not universal, especially skincare. To navigate product quality and suitability, consumers rely on detailed reviews from others. Often times, one of the characteristics that plays into the core of buying products is the skin type that an individual has. The shift towards online shopping and increased social media engagement have also led to greater scrutiny of product claims and marketing messages. Consumers demand transparency and authenticity, making it essential for brands to build trust through honest communication and genuine endorsements. Detailed reviews and user-generated content have become valuable assets for brands to showcase the effectiveness and suitability of their products, catering to the diverse needs and preferences of their audience.

Machine learning algorithms can analyze vast datasets to detect patterns and trends. Natural language processing, a subset of machine learning, offers a nuanced and sophisticated understanding of human language and sentiment, providing extensive benefits. This paper aims to evaluate the accuracy and impact of online reviews in shaping consumer purchasing decisions, using sentiment analysis as the core methodology. To reach the objective, several analysis are conducted such as using Logistic regression, Random Forest and Neural Network Model: LSTM, BERT. This approach is pivotal in understanding the influence of the consumer's reviews. By applying this technique, industries can gain nuanced insights to enhance product development, refine marketing strategies, and understand emerging trends.

ML and NLP equip beauty industry players with tools to make data-driven decisions, enhancing product development, marketing strategies, and customer engagement. By leveraging these technologies, businesses can stay ahead in a competitive market, continuously meeting consumer expectations and fostering brand loyalty.

CHAPTER 2

Data and Exploratory Data Analysis

2.1 Data Collection

the raw data for this study was collected from Sephora's official website, supplemented by user-scraped data from Github repositories. The data collection spanned from late 2017 to 2023, resulting in a comprehensive dataset segmented into several CSV files. Before combining all these files, we obtained dataset consisting approximately 1 million observations, and 18 variables. To ensure data quality and integrity, we addressed the missing values issue which were present in both categorical and numerical variables. Missing values can affect the data by skewing the results and leading to inaccurate analysis.

2.2 Variables

The dataset contains 18 variables, but in the analysis we are only using several variables in the model. However, this is the list of the

- 'author_id': The unique identifier for the author of the review on the website.
- 'rating': The rating given by the author for the product on a scale of 1 to 5.
- 'is_recommended': Indicates if the author recommends the product or not (1-true, 0-false).
- 'submission_time': Date the review was posted on the website in the 'yyyy-mm-dd'

format.

- ’review_text’: User’s review regarding product. It’s a string data type
- ’review_title’: The title of the review text. It’s a string data type
- skin_tone: Author’s skin tone (e.g. fair, tan, etc.)
- eye_color: Author’s eye color (e.g. brown, green, etc.)
- skin_type: Author’s skin type (e.g. combination, oily, etc.)
- hair_color: Author’s hair color (e.g. brown, auburn, etc.)
- ’helpfulness’: The ratio of all ratings to positive ratings for the review: helpfulness = total_pos_feedback_count / total_feedback_count
- ’total_feedback_count’: Total number of feedback (positive and negative ratings) left by users for the review.
- ’total_neg_feedback_count’:The number of users who gave a negative rating for the review.
- ’total_pos_feedback_count’: The number of users who gave a positive rating for the review.
- ’product_id’: The unique identifier for the product from the site
- ’product_name’ : The full name of the product.
- ’brand_name’: The full name of the product brand.
- ’price_usd’: Price of the product.

2.3 Imbalanced Dataset

There are 413906 observations after cleaning all the missing values. After doing exploratory analysis, there is an indication of imbalanced dataset. Class imbalance occurs when one class significantly outnumbers another in a classification problem, which can lead to biased model performance favoring the majority class. Due to the high number of the recommendation instead of the latter. To handle imbalance issue, performing downsampling is one the techniques that is performed on the majority class which is the 'recommended'. The down sampling is done without replacements and the number of samples drawn from the majority class should be equal to the number of instances in the minority class, 'not recommended'. After the downsampling, the number of the observations is reduced to 196542 observations. This approach is particularly useful and ensures that the machine learning models built using this dataset are more robust and generalizable.

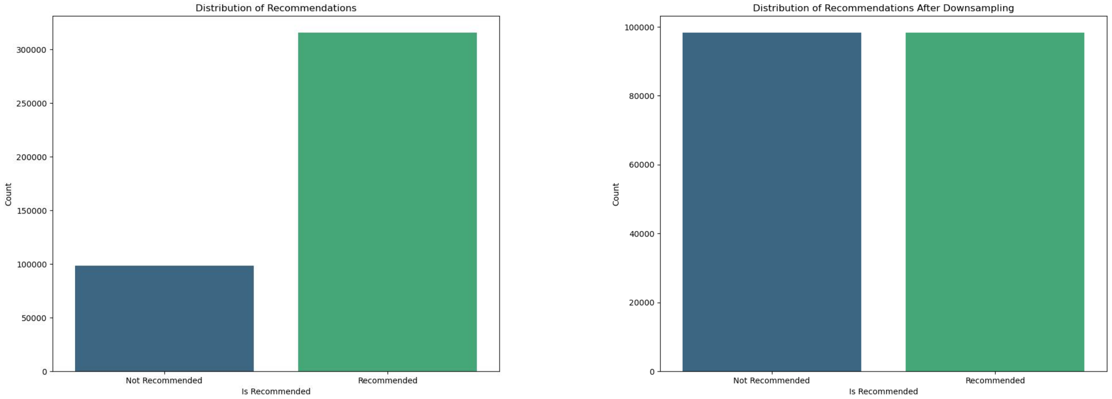

Figure 2.1: Before and after of the recommendation bar plot

2.4 EDA

To ensure the dataset's usability, we also conducted exploratory data analysis (EDA) to understand the underlying patterns and distributions. EDA helped us identify any anomalies or outliers that could affect the analysis. By visualizing the data, we could better comprehend the relationships between variables and the overall structure of the dataset.

2.4.1 Review Text

The variable 'review text' is one of the most important aspect in this analysis. Below are the density plot for the distribution of sephora reviews word count. As we can see from the figure below, the overlap between both positive and negative reviews does not differ significantly based on sentiment. Both have similar word count distributions. Additionally, the distribution of the Sephora reviews word count are right skewed.

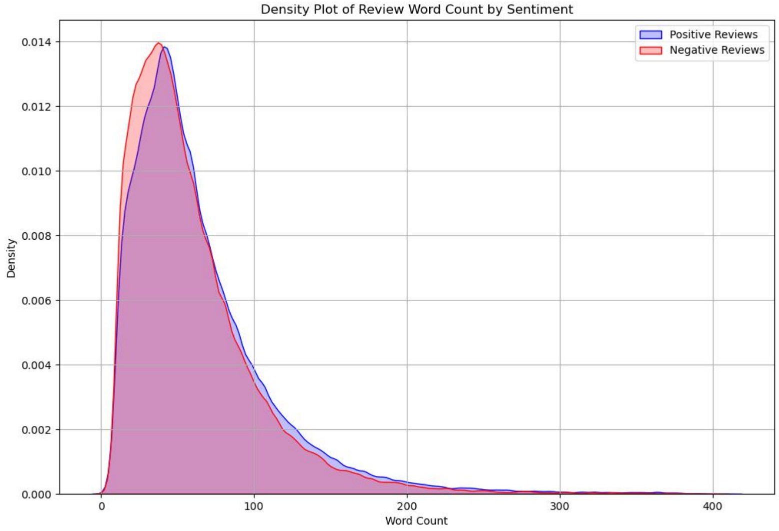

Figure 2.2: Distribution of the sephora reviews word count

2.4.2 Submission Time

The timeline of the review submission are throughout the year of late 2017 to early 2023. The graph below is a line graph that illustrates monthly review counts from the Sephora dataset over a period from late 2017 to early 2023. There are gradual increase in the number

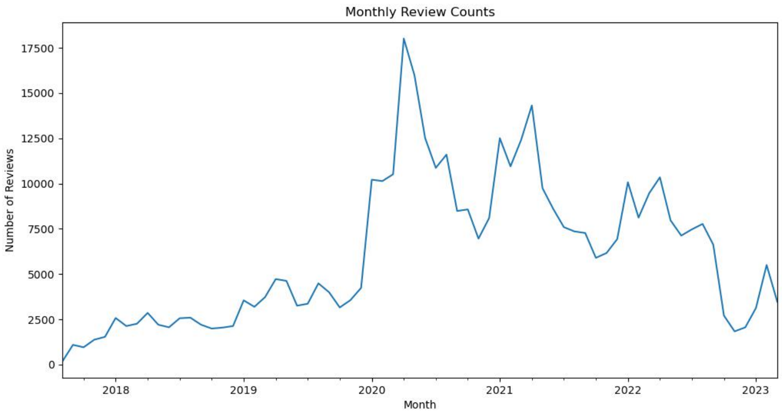

Figure 2.3: Number of Reviews across the Years

of reviews from early 2018 to early 2019 and then fluctuations with intermittent peaks and troughs until early 2020. The line shows significant peak around early to mid-2020, with the number of reviews reaching nearly 18,000 which represent the highest review activity in the dataset. It slowly declines post-2020, but with continued fluctuations, which can indicate changes in user engagement or external factors affecting review submission rates. Additionally, the highest review activity peak in the year of 2020 might also indicates the effect of pandemic that may have impacted the user behavior.

2.4.3 Rating

In the dataset, there consists of 5 ratings that indicates 5 as the highest rating to 1 as the lowest rating. Below are the distribution of the rating after down sampling the data. The distribution of the ratings shows that the majority of the reviews are polarized towards the extremes, with a large number of very dissatisfied (rating 1) and very satisfied (rating 5). In this graph, it can be indicated that consumers appears to have strong opinions about the products the review, either very positive or very negative, with few users feeling neutral or moderately satisfied.

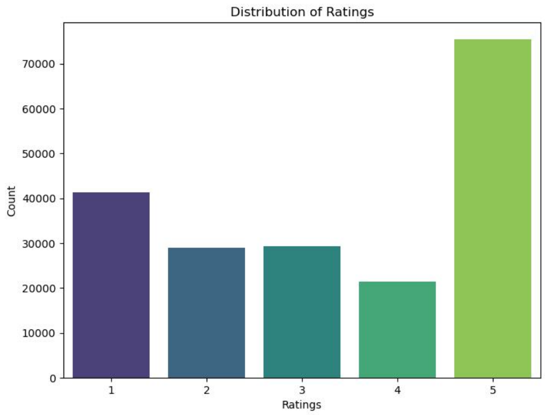

Figure 2.4: Ratings Distribution

2.4.4 Positive and Negative Feedback

The graph below illustrates the trends in both positive and negative feedback counts for product reviews from 2017 to 2023. For notes, the orange line indicates the positive feedback and blue line indicates negative feedback. Overall, both positive and negative feedback counts show significant variants over the year. The peak is dramatically increased for the positive feedback in the year of 2020, also a notable rise for the negative feedback although not as dramatic as positive feedback, which we can assumed is caused by COVID-19 pandemic due to lack of social activities and it leads to emotional spending.

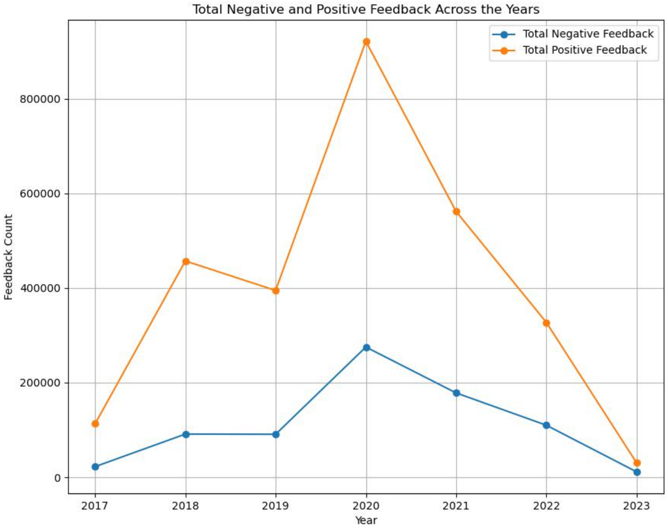

Figure 2.5: Positive and negative feedback across the year

2.4.5 Categorical Variable

In this dataset, there are several categorical variables which are the characteristics of the reviewer which are skin tone, skin type, hair color and eye color. We did a quick logistic regression model, the first step is to convert categorical data into numerical values by using one hot encoder and then perform the logistic regression to the converted categorical data. Below is the result of the logistic regression, but without the coefficients of the categorical variables.

Dep.Variable:	is_recommended	No of Observations:	196542
Converged:	False	Df Residuals:	196514
Covariance Type:	non-robust	Df Model:	27
Log-Likelihood:	-1.3601e+05	Pseudo R-Squ	0.00163
LLR p-value:	7.054e-79	LL-Null:	-1.3623e+05

Table 2.1: Logit Regression Table Result

The table shows that the Pseudo R-squared value is quite low in 0.001663, indicating that the model explains just a small fraction of the variation in the is_recommended variable. Convergence refers to the process by which the optimization technique used in logistic regression determines the best-fitting model parameters. In this case the result say false, meaning the algorithm did not converge which suggest difficulties with model specification or data. Overall, while the model provides some significant results in the LLR p-value, the low Pseudo R-squared and convergence issues suggest that the categorical variables included may not sufficiently capture the factors influencing product recommendations, and further model refinement or additional data. In this case, although this physical characteristics may attribute additional information for analyzing, it is not included in the sentiment analysis process.

CHAPTER 3

Methodology

3.1 Prep-processing Text

Review text will be going through the prepprocessing by creating function that perform several steps to clean and prepprocess a given text input. The purpose of this process is to prepare the text for further analysis which is the natural language processing by removing unwanted characters, tokenizing, filtering stop words, and lemmatizing the words.

3.1.1 Removing Punctuation

Firstly, removing punctuations is considered essential because punctuations do not usually contribute to help recognizing the topic or sentiment of the text sentences and in many computational tasks, they can be considered noise. It helps simplifying, and streamlining the text data.

3.1.2 Tokenization

Tokenization is the process of splitting the text into individual words or tokens. It involves using a tokenizer to segment unstructured data and natural language text into distinct chunks of information, treating them as different elements. Tokens within this data can be used as vector, transforming an unstructured text document into a numerical data structure suitable for the modelling process.

3.1.3 Removing Stop Words

In this process, removing stop words are necessary because these words might take up space in the database and take up processing time. By storing a list of words such as preposition, it can help to channel more attention towards words that truly convey the essence of the text.

3.1.4 Lemmatization

The process of grouping together different forms of a word so they can be analyzed as a single item. This step help to normalize the words, making the text more uniform and easier to analyze.

3.1.5 Adjectives

The objective of this section is to get a comparison between the two ratings which are rating 5 and rating 1. Positive adjectives are more expected from the highest rating while negative adjectives are expected from the lowest rating. The word cloud below shows that although the rating is high, there is still some negative adjectives such as ’bad’, ’cant’, and ’sticky’ in the word cloud. However, the lowest rating word cloud indeed have much more negative adjectives.

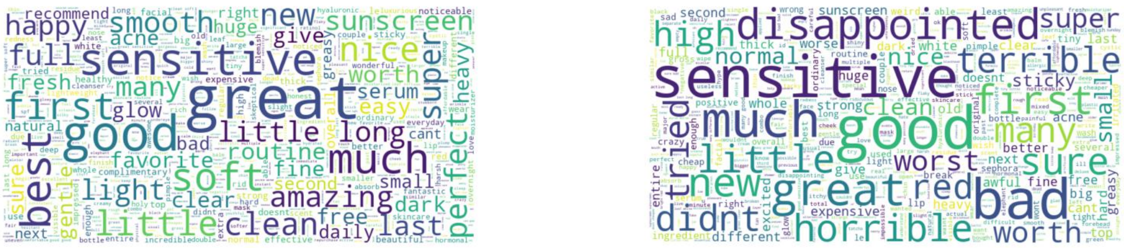

Figure 3.1: Rating 5 and Rating 1 Comparison

3.2 Logistic Regression

Logistic regression is a classification technique that is used to predict categorical outcome especially binary outcomes [0,1]. This technique is used to help classify the positive and negative sentiment using a logistic function.The logistic regression method will be applied to the two variables that we assumed have direct relationship with the review text. In this case, the model tries to learn whether the consumer reviews actually represented the binary outcome of the 'is_recommended' variable. Since the model is suitable for this particular task, it helps to provide probabilistic framework for classification, offering insights into the confidence of its prediction.

To achieve the goal to predicts whether a review recommends a product based on the sentiment expressed in the review text, we need to transform the review string into a numerical format suitable for the logistic regression. This is typically done using the text vectorization technique such as TF-IDF (Term Frequency-Inverse Document Frequency), which will be used a lot in other models too since it has the ability to convert strings into numerical values and reflect its importance in the context of the entire corpus.

The logistic regression model:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.1}$$

z is a linear combination of the input features and usually expressed as

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{3.2}$$

- β_0 : intercept
- $\beta_1, \beta_2, \ldots \beta_n$:Coefficients for the features $X_1, X_2, \ldots, X_n$

The logistic regression model for the binary classification can also be expressed as :

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}} \tag{3.3}$$

$$P(y=0|X) = \frac{e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}} \tag{3.4}$$

By setting threshold, we can classify the reviews into recommended (1) or not recommended (0). The training process involves optimizing the coefficients to minimize the log loss:

$$\text{Log-Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{3.5}$$

By minimizing the loss, it enables the model to learn to accurately predict our objective. Additionaly, the decision rule in logistic regression especially in binary classification If $P(Y = 1 \mid X) > P(Y = 0 \mid X)$, it means that $Y = 1$.

3.3 Ordinal Regression

Ordinal regression is a statistical technique for predicting an ordinal variable, which is categorical and has an inherent order. It extends logistic regression by relating the logit of a binary response to the predictors linearly [9]. A key assumption is proportional odds, indicating the consistent effect of an independent variable across response levels. The model generates an intercept for each response level, except one, and a single slope for each predictor. This makes ordinal regression particularly useful for modeling ordered response variables, such as a rating scale. The goal of ordinal regression is to predict the chance that a result will fall into each of the ordered categories. This is often achieved by establishing thresholds or cut-off points that divide the continuous latent variable into intervals corresponding to the ordinal categories. Cumulative link models (CLMs), such as the proportional odds model, are widely used for performing ordinal regression.

We are applying ordinal regression to predict the ratings of reviews in the Sephora dataset. Data is split into 80/20 split to evaluate the model's performance on unseen data, and to ensure the reproducibility of the results. The ratings are an ordinal variable with values ranging from 1 to 5. Similarly to the logistic regression method, we transformed the review text into numerical features using vectorization. Since the rating variable is ordinal, each threshold requires the creation of binary targets. This involves creating separate binary variables indicating whether the rating is greater than each possible threshold (1, 2, 3, and 4). This approach helps in capturing the ordinal nature of the ratings. For each threshold from 1 to 4, a binary target is created where the rating is converted to 1 if it is greater than the threshold, and 0 otherwise. For each threshold, a logistic regression is trained because its suitability for binary classification and can be extended to ordinal regression by training multiple classifiers. Predictions are made for each binary classifier, resulting in predicted probabilities for each threshold. So the probabilities are then used to predict the ratings, if the predicted probability is greater than 0.5 for each threshold, the corresponding rating is incremented.

3.4 Random Forest

Random Forest is an ensemble of decision trees that is used for prediction and classification. This technique works by creating a number of Decision Tress during the training phase to reach a single results. Each tree is constructed using by randomly selecting subset of the data and then measure it randomly. In prediction, the algorithm aggregating the result of the predictions through voting or averaging. The advantages of using this random forest is introducing variability, reducing the risk of over fitting and improving overall prediction [2].

Review texts are split using an 80/20, 80 % is used for training the data into the model and 20 %) in testing data to help evaluating the model's performance. By converting the review text into numerical features using Term Frequency-Inverse Document Frequency(TF-IDF) vectorization. It helps transforming the review text into a matrix of TF-IDF features and this reflects the importance of words in the context of the entire corpus. By limiting

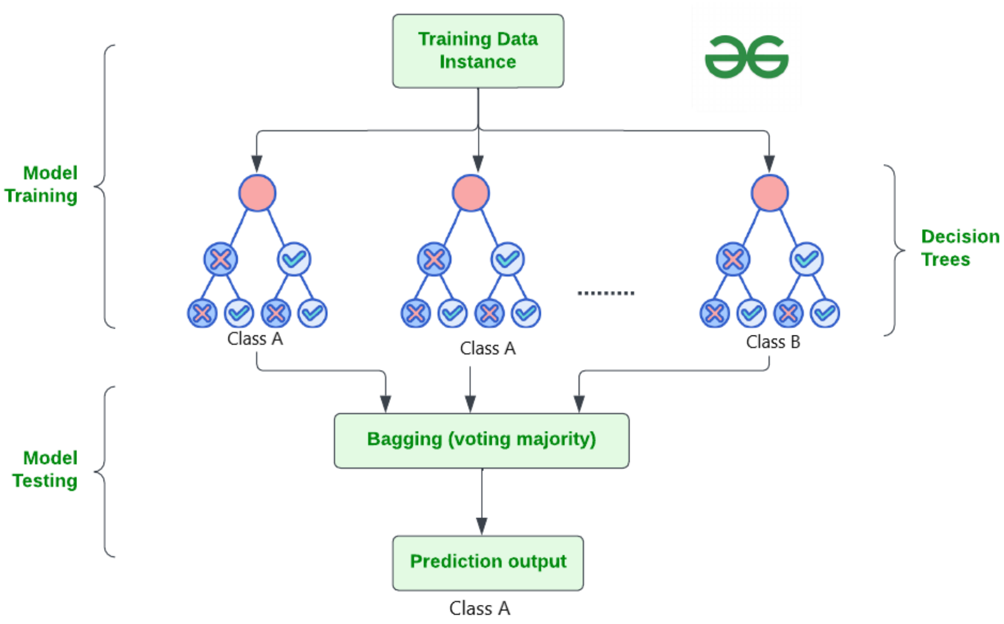

Figure 3.2: Random Forest Algorithm [2]

the number of features to 5000, it will help to maintain efficiency and reduce overfitting. In this method, the random forest classifier is instantiated with 100 tress and a fixed random state for reproducibility. For this method, we are doing it twice because we would like to see the difference between two response variables: is_recommended and rating. since we are interested in the comparison between these two variables, we are conducting it twice but with some differences in training aspect. For rating variable, it requires conversion into multiple binary classification problems for each threshold while is_recommended variable is very straightforward.

3.5 Neural Network

Before delving into LSTM and BERT model, it's important to understand neural networks, the foundational structures behind this model briefly. This machine learning program is usually used for supervised learning. Neural network model inspired by the structure and function of the biological function of the human brain. Every neural network consists of layer of nodes, or artificial neurons an input layer, one or more hidden layers, and output layer [7]. Each node connects to others with its associated weights. Neural networks are data driven, self-adaptive methods, and capable of adjusting to data without explicit specification of a functional or distributional form for the underlying model.

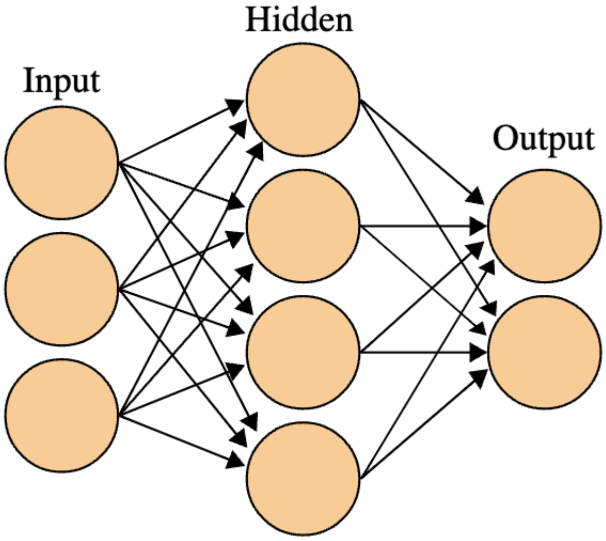

Figure 3.3: Simple neural network diagram [7]

In neural network, the input layer receives the raw data and then processed by one or more hidden layer where most computation occurs. Then each node inside the hidden layer applies a mathematical function to the input data, transforms it and sends the result to the next layer. The next layer, which is the output layer produces model's predictionTraining neural network includes weights within its node, this weights connections between neurons help to minimize error in prediction. The process typically done using the back-propagation method, the process of adjusting the weights of the neural network by analyzing the error from the previous iteration. Thus, to minimize the cost function during the training, it involves optimization algorithms, such as gradient descent or stochastic gradient descent. The network learn by iteratively processing the training data, comparing prediction to actual outcomes.

In terms of sentiment analysis, the architecture of the Neural Network typically:

- Input Layer: Converts text into numerical representations techniques like word embeddings or one-hot encoding
- Hidden Layers:
  - Dense Layers: Simple Neural Network
  - Convolutional Layers (CNN): Effective in capturing local patterns in text.
  - Recurrent Layers (RNN, LSTM, GRU): Excellent for processing sequences and capturing dependencies in text
- Output Layer: Produces the final sentiment classification.

3.6 Recurrent Neural Network (RNN)

Before focusing on LSTM, in short for Long short-term memory, we focus on Recurrent Neural Network. LSTM network is a type of an RNN and it's the simpler system of the network. In terms of definition, RNN is a deep learning model that is designed to train sequence data processing. Sequential data such as natural language processing, speech recognition, and time series data. On training the data, this model uses a back propagation algorithm, algorithm where it learns from the past context or error correction.

- X_t represents the input at time step t. In context of text, this could be a word or a feature vector representing a word.
- h_t represents the hidden state at time step t, which captures the information from previous steps to current step.

- A represents the unit that processes the input X_t and the previous hidden state h_{t-1} to produce the current hidden state h_t

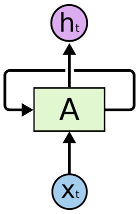

Figure 3.4: Recurrent Neural Network Loop Figure [8].

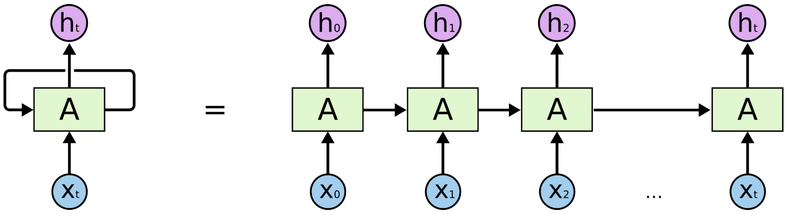

Figure 3.5: Recurrent Neural Network Loop Breakdown Figure [8].

The expanded version unrolls the recurrent neural network to show its operation over multiple time steps. The arrows show how the hidden state is passed from one time step to the next, allowing information to propagate through sequence.

3.7 Long short-term memory (LSTM)

The LSTM network was invented with the goal of addressing the limitations of traditional RNN which are the vanishing gradients problem and the inability to retain long-term dependencies data. It is particularly effective because they can capture and maintain context over long sequence of text, allowing them to understand the sentiment conveyed in a piece of writing more accurately. Unlike RNN, LSTM have more complex cell structure, including gates that control the flow of the information.

- X_t: The input at the current time step t

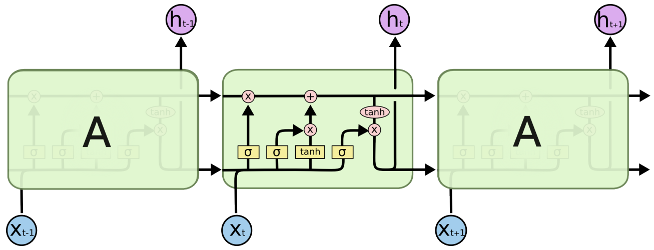

Figure 3.6: LSTM Cell Structure [8].

- h_{t-1}: The hidden state from the previous time step *t-1*

- h_t: The hidden state produced by the LSTM cell at the current time step t

- C_{t-1}: The cell state from the previous time step *t-1*

- C_t: The cell state updated by the LSTM cell at the current time step t

How it works:

- Forget Gate: Decides what portion of the cell state should be forgotten. It takes X_t and h_{t-1} as inputs and outputs a value between 0 and 1 through a sigmoid function.

- Candidate Cell State: Created by passing X_t and h_{t-1} through a tanh function, which is then scaled by the input gate.

- input Gate: Decides what new information should be added to the cell state. It also takes X_t and h_{t-1} as inputs and uses a sigmoid function.

- Cell State: The horizontal line running through the top of the cell represents the cell state, which carries information across time steps with minimal modifications. The cell state is updated by combining the old cell state that is filtered by the forget gate. This step ensures that relevant information is retained and updated over time.

- Output Gate: Determines what part of the cell state should be output. It uses a sigmoid function on concatenation of X_t and h_{t-1}

In short, the LSTM cell effectively manages the flow information through three gates: forget, input and output. It maintains a cell state that carries forward relevant information across time steps. By selectively forgetting, updating and outputting information, LSTMs can retain long-term dependencies in sequential data compared to RNN which is why it is a powerful tools to use for language modeling [8]. The LSTM model is well-suited for sentiment analysis for the Sephora's reviews because of its ability to capture sequential dependencies in the text data. By leveraging LSTM capability, the model can effectively predict whether the reviews are capable in handling product recommendation matter. To apply this technique we begin with text preprocessing and tokenizing our review text. We are transforming the text data into sequences and then padded them to ensure its uniformity. Similarly to other models, the data is split into training and testing sets using an 80/20 split. To build the LSTM model, we begin with creating an embedding layer to convert word indices into vectors of fixed size dimension. Two LSTM layers are stacked, with the first layer returning sequences to feed into the second layer, not to forget also adding dropout layer to prevent overfitting. A dense layer with sigmoid activation function outputs the probability of the reviews of products that is recommended. The model is then compiled with Adam optimizer and binary cross-entropy loss. With LSTM's capability to handle long term dependencies, the model can help to effectively predict whether a review recommends a product.

3.8 Transformers

Before delving into the BERT model, we are learning about some basic of Transformer. Previously, we discuss about RNN incapability in handling long term dependencies and problems and also vanishing gradient problem. So LSTM was created to handle this long term dependencies, but LSTMs process data sequentially which indicates that they cannot be fully parallelized, this means that their inherent sequential nature forces them to process one step at a time and dependent to the previous state. This leads to slower training times, especially for long sequences. It can be computationally intensive and require significant resource for training particularly with large datasets or long sequences. To address this issue, transformers model were created. This is a deep learning architecture developed by scientists working at google based on the paper ”Attention Is All You need”. This paper discuss about self-attention mechanism that allows the model to weigh the importance of different words in a sentence relative to each other, enabling the model to capture dependencies between words regardless of their distance in the sequence. Self attention mechanism is highly parallelizable and efficient architecture that captures long-range dependencies more effectively than traditional RNNs and LSTMs [10].

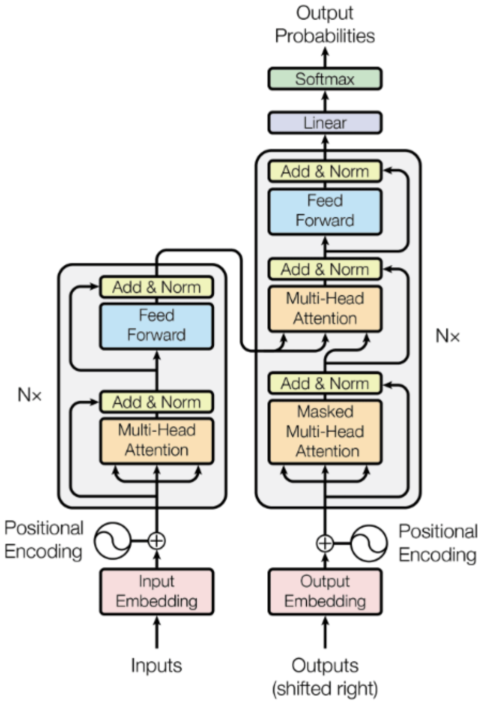

Figure 3.7: Transformer model archictecture [10].

The figure above is the transformer architecture, as introduced in the paper. The model consists of an encoder and decoder that utilize primarily self-attention mechanisms to process and generate sequences of data. The encoder processes the input sequence by transforming it into compact vectors using embeddings and incorporating positional encodings to capture information about the order of the sequence. It is composed of multiple identical layers, each having a multi-head self-attention mechanism and a feed-forward network, which are stabilized by residual connections and layer normalization. The layout of the decoder, which creates the output sequence, is identical to that of the encoder, but it also has a masked multi-head attention mechanism to make sure predictions rely only on outputs that are known. The decoder also processes the encoder's output, using context from the input sequence. This innovative use of self-attention mechanisms allows it to process and analyze all tokens in a sequence simultaneously, enabling parallelization and solving the problem with traditional RNN and LSTMs. Multi-head attention allows the model to focus on different sections of the sequence simultaneously, whereas feed-forward networks analyze each point independently. Positional encodings ensure that the model accurately represents the sequence's order. Finally, the output layer generates probabilities for each token using a linear transformation and a softmax function, making tasks such as machine translation and text synthesis more efficient and effective.

3.9 BERT Model

BERT, short for Bidirectional Encoder Representations, began from Transformer model in the previous section. BERT is designed to pre-trained deep bidirectonal representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, pre-trained BERT model can be fine-tuned with just one additional output layer to create models for wide range tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT process text sequentially, either left to right or right to left. Since its nature uses bi-directional instead of analyzing text sequentially, BERT looks at all words in sentence simultaneously. The BERT model was pre-trained with two objectives [4]:

- Masked Language Modeling (MLM): MLM take the sentence and then the model randomly masks 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words.
- Next Sentence Prediction (NSP): The models concatenates two masked sentences as input, it trains a model that understands sentence relationships.

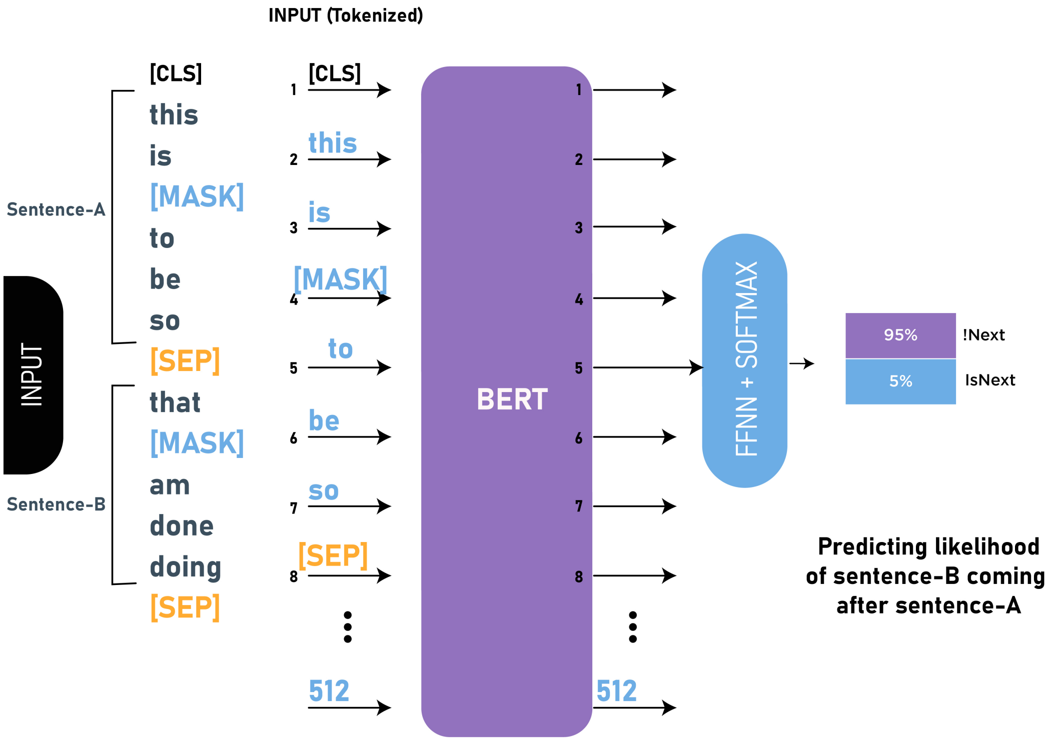

Figure 3.8: BERT Figure [5].

For our analysis, we are incorporating the pre-training knowledge of the BERT so that it could adapt for specific task of the sentiment analysis. A pre-trained BERT model 'bert-base-uncased' and its tokenizer are used to setup our model. Then created a class of custom dataset to help handling tokenization and batching of review texts for input into the BERT model. The model is trained for a specified number of epochs and during each epochs, it will processes batches of training data and compute the loss and updates the model parameters using back-propagation. Each batch consists of input layer and also attention mechanism to indicates which tokens are actual words and which are padding. The attention mechanism in BERT uses these input to weigh the importance of different words in sequence. During the forward pass, the input layer and the attention are fed into the model producing logits as outputs. Through softmax function, these logits are converted as probabilities to represent the predicted probabilities for each sentiment class as the output.

CHAPTER 4

Experiment

The purpose of this section is to present the results of the experiment that we conducted in the previous section. Some of the models were applied to the rating variables to examine how they compared to the is_recommended variable. The analysis was performed using Logistic Regression, Ordinal Regression, Random Forest, LSTM, and BERT. Each model's efficacy is evaluated using four metrics: accuracy, ROC AUC, precision, and recall. They provide a comprehensive understanding of a model's performance from different perspectives.

Before diving deep into the result, we need to understand how we get this evaluation metrics. For classification case in this paper, we need to calculate True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as the chosen metrics. True Positive stands for the % of incidents that are labeled positive accurately. True Negative stands for instances that were accurately identified as negative. False Positive as instances that were mistakenly labeled as positive. False Negative are instances that were mistakenly labeled as negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Accuracy is a metric used to assess how well a prediction model performs, and how often labels are correctly classified. High metric

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Precision, also known as Positive Predictive Value, is the ratio of correctly predicted positive instances to the total predicted positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

Recall, also known as Sensitivity or True Positive Rate, is the ratio of correctly predicted positive instances to all instances that are actually positive.

4.1 Model Result

4.1.1 Rating

Rating	Training Accuracy	Testing Accuracy	ROC AUC	Precision	Recall
Ordinal Regression	0.626	0.6019	0.3062	0.5324	0.5239
Random Forest	0.999	0.6010	0.936	0.8874	0.8888

Table 4.1: Rating Model Result Table

ROC AUC	1	2	3	4
Random Forest	0.923	0.933	0.952	0.935

Table 4.2: ROC AUC result for each threshold

If we review the findings in the tables above, we observe that the results for ordinal regression and random forest are quite close, particularly in terms of accuracy. Both imply that approximately 60% of the projections accurately predict the reviews. However, despite their identical accuracy, the ordinal regression model appears to perform poorly in terms of distinguishing rating classes compared to the random forest, according to the ROC AUC scores. The second table displays proof of random forest's high ROC AUC score, this suggests that random forest performs well at each threshold. Looking at the precision and recall scores, random forest has a substantially higher number than ordinal regression, indicating that it produces fewer false positive predictions and effectively detects all relevant positive cases, capturing the majority of positive ratings. Based on the results presented above, it is clear that the random forest model outperforms ordinal regression in predicting customer ratings for each Sephora product in the dataset. As a result of its capabilities, random forest is a preferred option for rating prediction analysis.

4.1.2 is_recommended

is_recommended	Training Accuracy	Testing Accuracy	ROC AUC	Precision	Recall
Logistic Regression	0.518	0.889	0.954	0.887	0.888
Random Forest	0.999	0.882	0.951	0.881	0.880
LSTM	0.943	0.904	0.957	0.911	0.896

Table 4.3: is_recommended Model Result Table

BERT	Training loss	Accuracy	Validation Loss	Accuracy
epoch 1	0.496	0.752	0.358	0.861
epoch 2	0.264	0.898	0.334	0.859
epoch 3	0.150	0.951	0.408	0.863

Table 4.4: BERT Training Result Table

is_recommended	Testing Accuracy	ROC AUC	Precision	Recall
BERT	0.874	0.938	0.871	0.878

Table 4.5: BERT Test Result Table

Based on the table 4.3 above, LSTM model appears to have the highest accuracy compared to the other models based on the numbers. However, all model shows approximately similar number around 88 89% and still performing well in predictions. Overall, LSTM shows better performance in predicting and distinguishing whether a product is recommended or not since most of their evaluation metrics approximately lies around 90.4% and above. The high value in ROC AUC in 0.957 indicates excellent model performance in terms of identifying true positive and true negatives, it also misses fewer positive recommendations. In short, LSTM is highly effective at predicting whether a product is recommended or not, capturing most of the true recommendations while make a very few errors. Its reliability indicates when LSTM predicts a product is recommended, it usually is. Logistic Regression and Random Forest also demonstrate strong performance, with accuracies of 88.9% and 88.2%, respectively, and high ROC AUC values (0.954 for Logistic Regression and 0.951 for Random Forest). These models maintain a good balance between precision and recall, making them reliable for predicting product recommendations.

Table 4.4 showcase the detail of the BERT model's performance across three epochs during the training. It shows improvement in terms of the accuracy training from 0.752 in the first epoch to 0.951 in the third epoch. Additionally validation accuracy remain consistently high around 0.861 to 0.863 which indicates that the BERT model is effectively learning and generalizing the training data. The training loss also decreases substantially which suggests that the model is optimizing well. However, when we look at the slight increase in the final validation loss, it could possibly indicate the beginning of overfitting, despite the consistent validation accuracy. Table 4.5 showcase the test result from the BERT model, the metrics shows a strong performance. BERT, while still performing very well, lags slightly behind with an accuracy of 0.874 and an ROC AUC of 0.938. Its precision in 0.871 and recall in 0.878 are slightly lower compared to the other models, indicating that while BERT is effective, it is not as strong as LSTM, Logistic Regression, or Random Forest in this specific task.

Overall, the strong performance metrics across all models, particularly LSTM, suggest that these models are well-suited to predicting whether a product is recommended based on customer reviews from the Sephora dataset. This shows that these models might provide useful insights into consumer happiness, allowing Sephora to better determine which products appear most inclined to be recommended by customers while also enabling more targeted marketing strategies and inventory management. When comparing both rating and is_recommended prediction results, it is evident that most of the models perform significantly better in predicting product recommendation. Based on the numbers on the table above, models like LSTM, logistic regression and random forest achieve high accuracy, along with substantial values of ROC AUC. These indicates that these models are highly effective in distinguishing recommended and non-recommended products. In contrast, rating shows much lower performance with both Ordinal Regression and Random Forest achieving an accuracy around 60% and low ROC value.

CHAPTER 5

Conclusion

5.1 Conclusion and Discussion

Overall, after conducting the sentiment analysis and also comparing some of the analysis, it is evident that the best model to determine the consumer's review accuracy is LSTM model. Additionally, it appears that recommendation variable is much more reliable in guiding other costumer in their purchasing decisions. When model reliably identifies the recommendation prediction, it helps to provide a clear signal to prospective buyers about the general satisfaction level of the previous customers. In addition, high accuracy in predicting the recommendation status means that new customers can trust the recommendation status displayed in the app or site, thereby influencing their purchase decision politely. Rating is not insignificant, but still less reliable compared to the recommendation aspect due to the complexity and nuances of the rating scale. From the business perspective itself, is_recommended has important implications that can help to enhance customer satisfaction, and guiding marketing strategies. Additionally, focusing on the the recommendation predictions also allows Sephora or brands that work with them to have more informed business decisions, enhancing customer experience and also drives sales growth.

5.2 Limitations & Future Work

Despite the high accuracy of models such as LSTM, logistic regression, random forest, and BERT in predicting the is_recommended variable, there are numerous limitations to the current sentiment analysis. For starters, models may fail to reflect the nuanced and subjective character of consumer reviews. This constraint is visible in the lower performance metrics for the "Rating" variable, which indicate that the models fail to distinguish between closely related rating levels. Furthermore, the models are trained on a Sephora-specific dataset, which may limit their applicability to other domains or product categories.

Future work could focus on addressing these constraints by investigating further advanced natural language processing techniques and using additional contextual data. For instance, we can use transformer-based models like BERT with further fine-tuning and adding elements such as customer profiles, product attributes, and historical review data that could enhance the model's performance and prediction. Additionally, if we have a larger and more diverse dataset, it could help increase the models' generalizability. Finally, implementing a real-time sentiment analysis system that continuously learns from new reviews and adapts to changing customer preferences would be a valuable extension of the current work, ensuring that the insights remain relevant and up-to-date.

Bibliography

[1] Grzegorz Adamczyk. Pathological buying on the rise? compensative and compulsive buying in poland in the pre- and (post-)pandemic times. *PLOS ONE*, 19:1, 3 2024. doi: 10.1371/journal.pone.0298856.

[2] Susmit Sekhar Bhakta. Random forest algorithm in machine learning, 2024. URL https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/.

[3] Hichang Cho, Pengxiang Li, Annabel Ngien, Marion Grace Tan, Anfan Chen, and Elmie Nekmat. The bright and dark sides of social media use during covid-19 lockdown: Contrasting social media effects through social liability vs. social support. *Computers in Human Behavior*, 146:2, 4 2023. doi: 10.1016/j.chb.2023.107795.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

[5] geeksforgeeks. Understandiing bert nlp, 2015. URL https://www.geeksforgeeks.org/understanding-bert-nlp/.

[6] Clara Ludmir. Sephora's record growth in 2023 reaffirms its beauty retail leadership, 2024. URL https://www.forbes.com/sites/claraludmir/2024/01/29/sephoras-record-growth-in-2023-reaffirms-its-beauty-retail-leadership/?sh=41d32f76e0f3. [Online; accessed 10-June-2024].

[7] Christopher Olah. Neural networks, manifolds, and topology, 2014. URL https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/.

[8] Christopher Olah. Understanding lstm networks, 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[9] Stephen Parry. Ordinal logistic regression models and statistical software: What you need to know, 6 2016.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.