Sentiment Classification using Bidirectional RNN with LSTM

Tarandeep Singh Mandhiratta
Master of Applied Computing
Toronto, Canada
2tarandeep@gmail.com

***Abstract*— The benefits of e-commerce in marketing have grown more obvious with the fast expansion of the Internet. Consumers, on the other hand, find it difficult to pick among a wide range of similar items. Consumers provide input on the purchasing process in the form of remarks, which influences the purchasing choices of other users. Social media is quickly becoming a prominent and popular technology platform that enables individuals to express personal thoughts on topics of mutual interest. These perspectives are useful for making decisions. People want to hear what others have to say before making a choice, while businesses want to keep an eye on what people are saying about their goods and services on social media and take necessary action. Sentiment analysis (also known as opinion mining or emotion AI) is the systematic identification, extraction, quantification, and study of emotional states and subjective information using natural language processing, text analysis, computational linguistics, and biometrics. Sentiment analysis is commonly used in marketing, customer service, and clinical medicine to analyze opinion of the customer materials such as comments and survey replies, as well as online and social media and healthcare resources. More complex data domains, such as news stories, where writers often convey their opinion/sentiment less openly, may now be evaluated thanks to the emergence of deep language models. In today's marketing techniques, it's critical to understand client attitudes. It will not only provide firms with information about how their consumers perceive their goods and/or services, but it will also provide them with suggestions on how to enhance their offerings. This study aims to decipher the relationship between several characteristics in customer reviews on a women's clothes e-commerce site, as well as categorize each review as to whether it suggests the reviewed product or not, as well as if it contains positive, negative, or neutral attitude. To accomplish these objectives, we used univariate and multivariate analysis on dataset characteristics, as well as a bidirectional RNN with LSTM for sentiment classification and recommendation. A recommendation is a major predictor of a good sentiment score, and vice-versa, according to the findings. Ratings in product reviews, on the other hand, are shaky indications of sentiment scores. We also discovered that the bidirectional LSTM could get an F1-score of 0.88 for recommendation classification and 0.93 for sentiment classification using the bidirectional LSTM.**

Keywords—— e-commerce, sentiment analysis, long short term memory, deep learning, recurrent neural networks, artificial neural networks.

I. Introduction

With the ongoing growth of the e-commerce sector, consumption has increased, and business competition has risen, making it tough for customers to pick among a range of identical items. Consumers' evaluations of the purchasing process and use value of commodities are transferred from consumers to merchants and consumers to consumers through e-commerce online reviews. If customers are happy with their purchases, their reviews may assist other customers in making judgments. On the contrary, it has a negative impact on the product's sales and the merchant's reputation [1]. As a result, online text reviews are very important to the growth of the e-commerce business.

The practice of evaluating and interpreting the remarks, thoughts, and feelings made by persons with emotional tendencies is known as sentiment analysis. The capacity of algorithms to understand text has greatly increased as a result of recent developments in deep learning. Advanced artificial intelligence techniques used creatively can be an effective tool for conducting in-depth research. For mining internet reviews, sentiment analysis technology has been frequently deployed. The findings may aid firms in making changes to their future marketing strategy, such as examining the benefits and drawbacks of items in many areas in order to enhance product quality and customized suggestions [2].

Sentiment analysis is the process of gathering text data from a range of sources, identifying views, and categorizing the results as positive, neutral, or negative responses to a product, service, or brand using artificial intelligence (AI) [3]. While survey responses have always persisted in retail, the rise of e-commerce has propelled the study of sentiment analysis to new heights, requiring precision-targeted methods that delve deeper into client opinion. Companies are increasingly using social media monitoring as a technique for better understanding their consumers and improving their goods and/or services. Text analysis has emerged as an active topic of study in computational linguistics and natural language processing as a result of this shift.

Text classification, a job that aims to group texts into one or more categories that may be performed manually or digitally, is one of the most common issues in the mentioned topic. In this approach, there has been a lot of interest in categorizing opinions expressed in social media, review sites, and discussion forums in recent years [4]. Sentiment analysis is a computer method that use statistics and natural language processing techniques to discover and classify views stated in a text, with the goal of determining the writer's polarity of attitude (positive, negative, or neutral) toward a subject or a product. Companies are increasingly using this activity to better understand their customers via social media customer service or online review boards.

Long short-term memory (LSTM) is a deep learning architecture that uses an artificial recurrent neural network (RNN). LSTM features feedback connections, unlike normal feedforward neural networks [5]. It can handle not just individual data points (such as photos), but also complete data streams (such as speech or video). Because there might be delays of undetermined length between critical occurrences in a time series, LSTM networks are well-suited to categorizing, processing, and generating predictions based on time series data. LSTMs were created to solve the issue of vanishing gradients that may occur while training standard RNNs [6]. In many cases, LSTM has an advantage over RNNs, hidden Markov models, and other sequence learning approaches due to its relative insensitivity to gap length.

We use statistical analysis and sentiment classification to assess customer evaluations on women's clothes e-commerce in this article. We begin by looking at the non-text review variables discovered in the dataset (e.g., age, dress class bought, etc.) to see whether there is any link between these and consumer recommendation on the product. Then, for identifying whether a review text recommends the bought product or not, and for categorizing the user review attitude towards the product, we design a bidirectional recurrent neural network (RNN) with long-short term memory (LSTM) [7].

II. RELATED WORK

This section assesses prior work in the subject of text sentiment analysis that has been published. The majority of the work employs sentiment lexicons, machine learning, and deep learning. Table I shows the results of the comparison and literature review.

1) On the foundation of metrics accuracy, precision, and recall values, L Dey et al. [8] compared two supervised machine learning algorithms, namely K-NN (K- Nearest neighbor) and Nave Bayes (NB). A movie review dataset was obtained from www.imdb.com. It was discovered that the Nave Bayes classifier outperformed the K-NN classifier.

2) B. Shin, T. Lee, and J. D. Choi [9] devised an approach in which lexical embeddings and an attention mechanism are combined in CNNs. The tweets were utilized as the dataset. Calculating the F1 score is used to evaluate the procedure. The suggested approach outperforms the present ones. The attention method used aids in the reduction of noise for successful sentiment analysis.

3)Y. Fang, H. Tan, and J. Zhang [10] suggested a multi-strategy sentiment analysis approach that heavily relies on fuzzy set theory, machine learning theory, and a polarity lexicons-based method. Consumer evaluations were then analyzed using this hybrid method. It considers adversative conjunctions like "but," "while," "although," and others, as well as opinion operators like "say," "present," and "suggest," among others. If such terms appear in a sentence, it is considered neutral. SVM (Support Vector Machine) and NB (Nave Bayes) were the conventional machine learning methods employed in the experiment. This article uses a balanced dataset from website2 that includes 3000 positive and negative hotel customer reviews. A training set of 2500 good and 3000 negative customer evaluations was sampled, while the remainder was sampled as a test set. This improved SVM (hybrid approach integrating multi-strategy sentiment analysis with SVM) improved accuracy (i.e. 86.35 percent). In addition, when using the upgraded NB, the author noticed a 3.8 percent increase in accuracy.

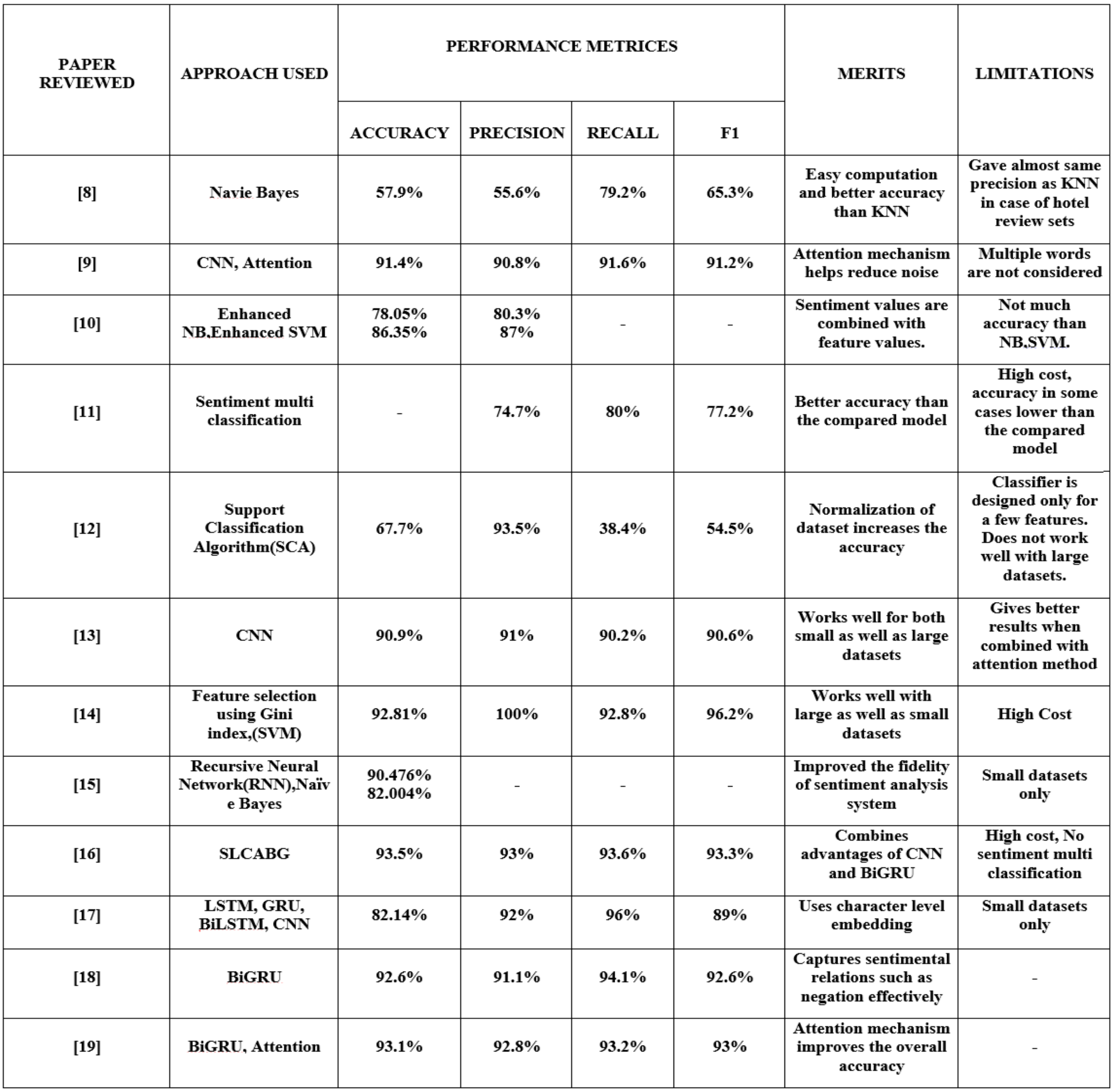
4) A sentiment multi classification approach was presented by S. Zhang et al. [11]. This approach combined a directed weighted model with sentiment analysis, extracting sentiment keywords as nodes from entities and their attributes, then framing a directed weighted correlation between two nodes to meet a set of requirements. Directed weighted linkages were used to depict node-to-node sentiment correlation. A directed weighted route is used to find similarities in feature nodes and to do sentiment classification analysis. Amazon Review Data was utilized as the source of the data (2018). The model's results were compared to the BERT model, and the experimental graphs showed that the suggested algorithm's CPU time was more efficient than the BERT model.

5) Two approaches for sentiment categorization were suggested by M. R. Huq, A. Ali, and A. Rahman [12]. They proposed the Sentiment Classification Algorithm, which employed the K-Nearest Neighbor approach, and the Support Vector Machine, which used the second technique. Real tweets are used to verify the performance. The results obtained by the proposed algorithm are superior than SVM in terms of experimental validation.

6) A combination of two deep learning approaches, CNN (Convolutional Neural Network) and K means clustering algorithm, was suggested by B. S. Lakshmi, P. S. Raj, and P. R. Vikram [13]. The performance of CNN and KNN was then compared to CNN with K means clustering in this work. On the basis of experimental validation, the first combination provided better results for smaller datasets, but CNN paired with the K means clustering approach offered better results for bigger datasets.

7) A. S. Manek et al. [14] proposed a technique for categorizing data that employed gini index-based feature selection and an SVM classifier. The data set for this project was a massive movie review dataset. The provided approach was shown to be less accurate in comparison to other methods based on the results of the experiments.

8) G. Preethi et al. [15] presented a recommendation system based on the cloud (RDSA). For sentiment analysis of reviews, this model combined the use of Recursive Neural Networks (RNN) with deep learning. Deep learning was used in this work to optimize suggestions based on sentiment analysis, which was performed on three separate reviews. The author looked into the datasets and looked into their statistical aspects before implementing the Nave Bayes baseline classifier and the RNN. The system's performance was then evaluated by comparing the Nave Bayes and



PAPER REVIEWED	APPROACH USED	PERFORMANCE METRICES				MERITS	LIMITATIONS
		ACCURACY	PRECISION	RECALL	F1		
[8]	Navie Bayes	57.9%	55.6%	79.2%	65.3%	Easy computation and better accuracy than KNN	Gave almost same precision as KNN in case of hotel review sets
[9]	CNN, Attention	91.4%	90.8%	91.6%	91.2%	Attention mechanism helps reduce noise	Multiple words are not considered
[10]	Enhanced NB,Enhanced SVM	78.05% 86.35%	80.3% 87%	-	-	Sentiment values are combined with feature values.	Not much accuracy than NB,SVM.
[11]	Sentiment multi classification	-	74.7%	80%	77.2%	Better accuracy than the compared model	High cost, accuracy in some cases lower than the compared model
[12]	Support Classification Algorithm(SCA)	67.7%	93.5%	38.4%	54.5%	Normalization of dataset increases the accuracy	Classifier is designed only for a few features. Does not work well with large datasets.
[13]	CNN	90.9%	91%	90.2%	90.6%	Works well for both small as well as large datasets	Gives better results when combined with attention method
[14]	Feature selection using Gini index,(SVM)	92.81%	100%	92.8%	96.2%	Works well with large as well as small datasets	High Cost
[15]	Recursive Neural Network(RNN),Naïve Bayes	90.476% 82.004%	-	-	-	Improved the fidelity of sentiment analysis system	Small datasets only
[16]	SLCABG	93.5%	93%	93.6%	93.3%	Combines advantages of CNN and BiGRU	High cost, No sentiment multi classification
[17]	LSTM, GRU, BiLSTM, CNN	82.14%	92%	96%	89%	Uses character level embedding	Small datasets only
[18]	BiGRU	92.6%	91.1%	94.1%	92.6%	Captures sentimental relations such as negation effectively	-
[19]	BiGRU, Attention	93.1%	92.8%	93.2%	93%	Attention mechanism improves the overall accuracy	-

Table 1: Comparison of related work

Recursive Neural Networks. The results of the trials showed that using a deep neural network based on RNN improved the fidelity of sentiment analysis, resulting in better recommendations for the user and aiding in the selection of a particular location depending on the user's requirements.

9) L. Yang et al. [16] introduced the SLCABG model, a sentiment analysis model that combines the Convolutional Neural Network (CNN) with the attention-based Bidirectional Gated Recurrent Unit (BiGRU) to support the sentiment lexicon. The advantages of sentiment lexicon and deep learning technologies are combined in this model. Because it combines the benefits of sentiment lexicon with learning technology, it solves the disadvantages of existing sentiment analysis models for product evaluations. Initially, sentiment lexicon is recruited in order to enhance the sentiment characteristics found in the reviews. Following that, the CNN and the Gated Recurrent Unit (GRU) network are used to extract main sentiment characteristics as well as contextual information. The data for this study came from the book reviews on the website dangdang.com. The accuracy, precision, recall, and F1 score were used as model assessment metrics in this research. The suggested model has a 93.5 percent accuracy rate in experimental validation, which was higher than the NB, SVM, and CNN models.

10) M. U. Salur and I. Aydin [17] combined character-level embedding with LSTM, GRU, BiLSTM, and CNN deep learning algorithms. The team proposed a hybrid technique for sentiment categorization. The major goal of this strategy was to improve classification performance by combining the capacity of several word representations with different deep learning algorithms. CNN could easily recognize feature extraction in the immediate neighborhood. LSTM can also extract valuable features from datasets with persistent dependencies, such as natural languages and signals. The datasets used were compiled from user-generated tweets about a GSM provider in Turkey. F1, kappa, and Recall were the performance matrices employed. When compared to the prior CNN model, this M-hybrid model had a higher accuracy rate of 82.14 percent during experimental validation.

11) C. Chen, R. Zhuo, and J. Ren [18] proposed a Gated Recurrent Neural Network with inter-opinion connections. The accuracy of this approach was 92.6 percent.

12) L. Zhou and X. Bian [19] presented a Bi Directional Gated Recurrent Unit (BiGRU) for categorization, together with an attention mechanism. This approach was shown to be useful for text categorization and produced better results than previously utilized methods, with a 93.1 percent accuracy rate.

III. DATA AND EXPERIMENTS

A. *The Dataset*

The dataset for this research was the Women's Clothing E-Commerce Reviews. Because this dataset contains evaluations made by actual consumers, it has been disguised by removing customer names and replacing references to the firm with "retailer." [20]

There are 10 feature variables and 23486 rows in this dataset. Overall frequency distribution across dataset characteristics and labels is shown in Table 2.

Feature	Unique Count
Clothing ID	1172
Age	77
Title	13984
Review Text	22621
Rating	5
Recommended IND	2
Positive Feedback Count	82
Division Name	3
Department Name	6
Class Name	20

Table 2: Dataset Feature Frequency Distribution.

B. *Data Analysis*

In this paper, we look at the dataset's properties and how they affect user recommendations and review sentiments. Four statistical studies are covered in this section. Table 3 summarizes the dataset's statistical description.

Feature	Mean	Standard Deviation	Type
Clothing ID	919.695908	201.683804	Integer
Age	43.282880	12.328176	Integer
Rating	4.183092	1.115911	Categorical
Recommended	0.818764	0.385222	Categorical
Positive Feedback	2.631784	5.787520	Integer

Table 3: An overview of Statistical Description of Dataset Features.

a) *Analysis on Univariate Distributions*

(1) Class Name: Figure 1 shows the frequency distribution of apparel classes most reviewed. The top three apparels are blouses, knits and dresses.

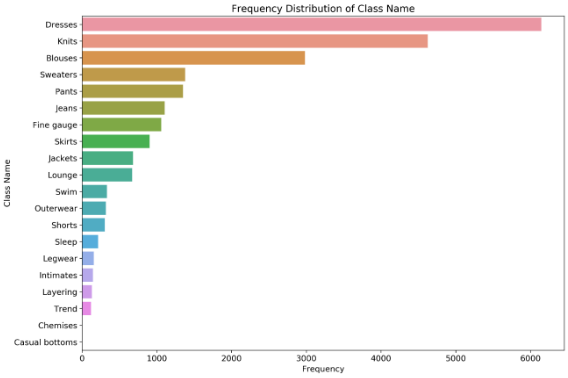


Figure 1: The distribution of apparels by frequency in each class.

(2) **Department Name and Division Name:** The frequency analysis of customer feedback by department and division is shown in Figure 2. This provides the e-commerce with information on the most popular consumer garment sizes and clothing kinds, i.e. general refers to clothing size and tops refers to fashion types.

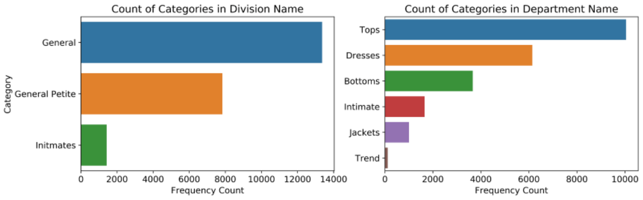


Figure 2: The frequency of apparel distribution per division and department.

(3) **Age and Positive Feedback Count:** Customers between the ages of 35 and 44 were the most involved in evaluating bought items, as seen in Figure 3. Furthermore, the data indicates that they have by far the most favorable feedback on the things they have bought. We may draw two conclusions from this: (1) Because the age group 35 to 44 is the most pleased in the range of consumers, the e-commerce entity in question should concentrate on sustaining this section, and (2) the e-commerce entity may investigate why other age groups are less happy than the 35 to 44 age group.

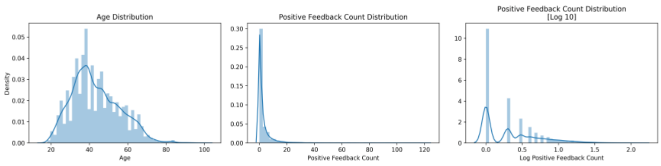


Figure 3: The frequency distribution of age and positive feedback

(4) **Word Length:** Figure 4 demonstrates that customers had essentially the same number of vocabulary in their evaluations independent of review quality, garment type, or suggestion.

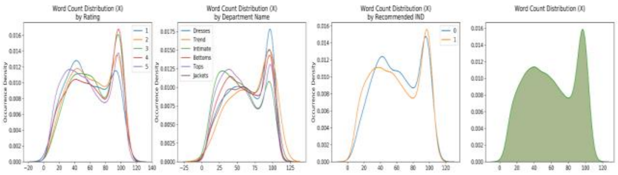


Figure 4: The frequency distribution of review texts per department, rating and recommendation

(5) **Rating, Recommendation, and Label:** The majority of assessments were favorable, as seen in Figure 5, implying that customers are fairly satisfied with e-commerce. A review with a suggestion may be seen as axiomatically implying a better rating and favorable attitude. However, the processing of feelings was based on a positive threshold of 3 and a negative threshold of 0 for the remainder.

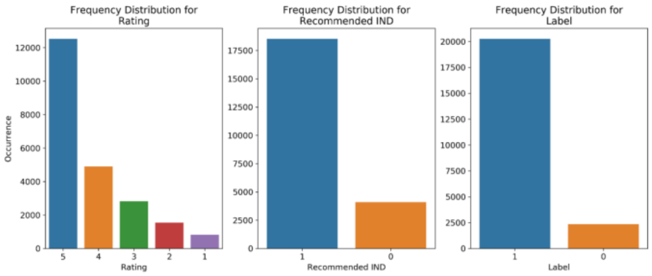


Figure 5: The frequency distribution of recommendation, review ratings, and labels.

(6) **Top 60 Clothing ID:** Figure 6 illustrates the IDs of the top 60 e-commerce apparel reviews. According to, the garments with clothing IDs 1078, 862, and 1094 are from the common subdivision and dresses apparently kind, and have a favorable title review of "Beautiful dress."

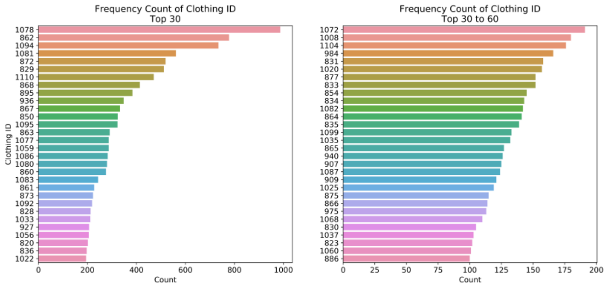


Figure 6: The frequency distribution of top 60 items per clothing ID

b) Analysis on Multivariate Distributions.

(1) **Class Name by Department Name:** Figure 7 shows the supremacy of dress among clothing kinds, which is backed up by Figure 8.

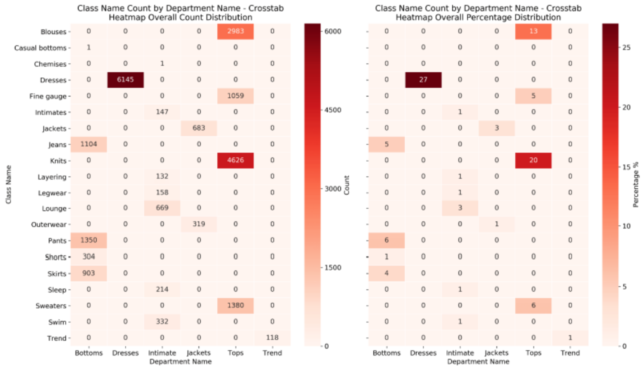


Figure 7: The cross tabulation for apparel per department and class.

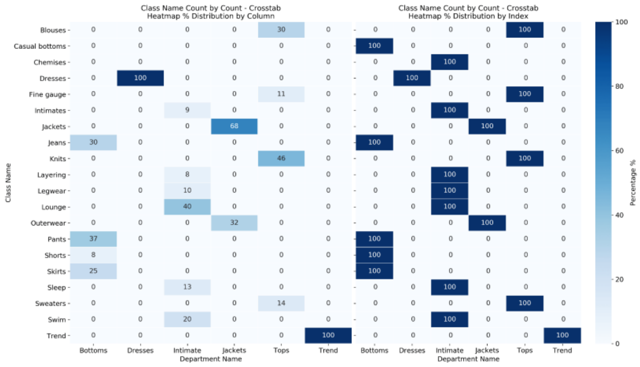


Figure 8: The normalized cross tabulation for apparel per department and class.

(2) **Class Name by Division Name:** Figure 9 shows that general-sized tops, dresses, and weaves are the most popular garment kinds. However, 10 reveals that the majority of dress ratings come from women in typical tiny sizes.

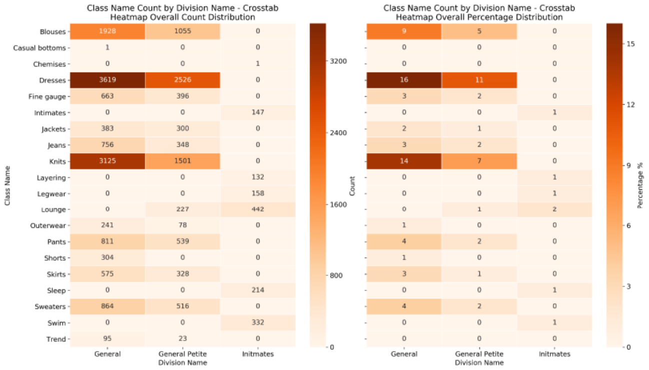


Figure 9: The cross tabulation for apparel per class and division

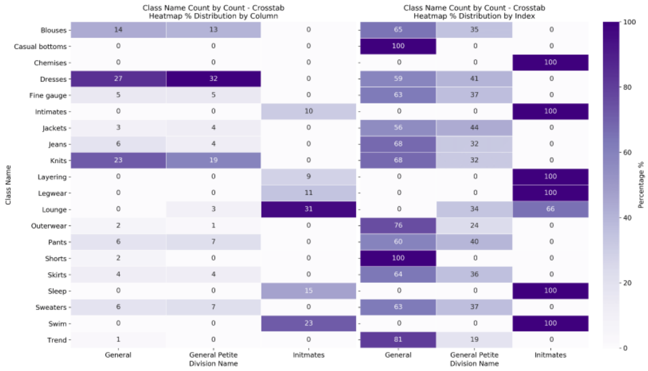


Figure 10: The normalized cross tabulation for apparel per class and division.

(3) **Division Name by Department Name:** The supremacy of general-sized tops is seen in Figure 11, and this conclusion is supported by Figure 12.

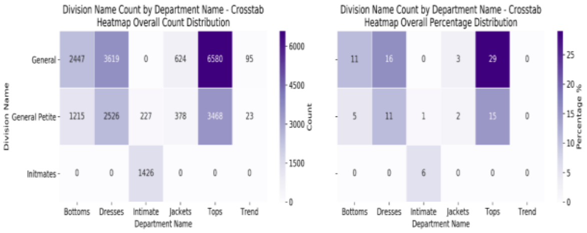


Figure 11: The cross tabulation per division and department.

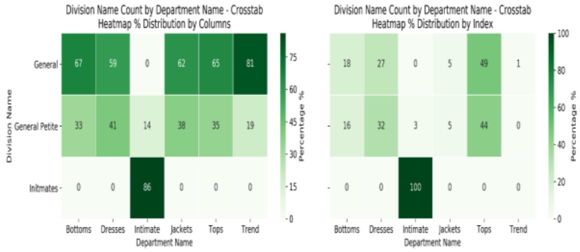


Figure 12: The normalized cross tabulation per division and department.

(4) **Age by Positive Feedback Count:** Figure 13 illustrates a minor relationship between age and good review feedback. According to the graph, the age range of 35 to 44 appears to be the one that provided the most favorable response.

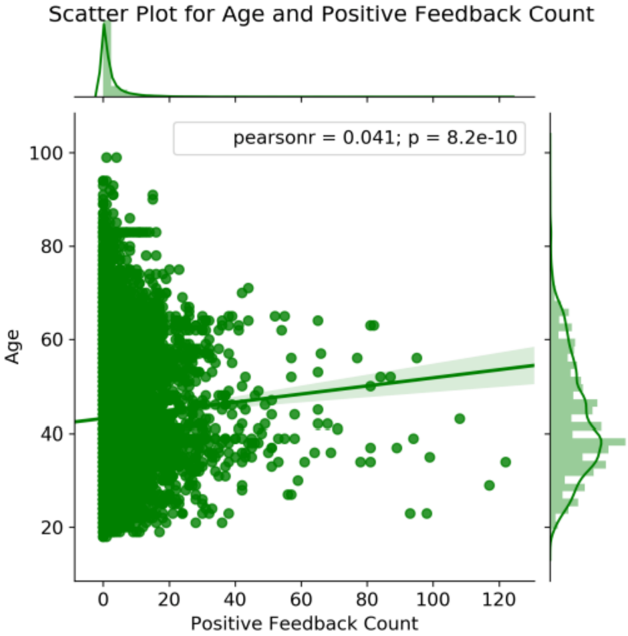


Figure 13: The scatter plot for age and positive feedback count.

(5) **Rating by Recommendation:** Figure 14 backs up the idea that a review rating reflects its recommendation status, with a higher rating indicating a recommendation and vice versa.

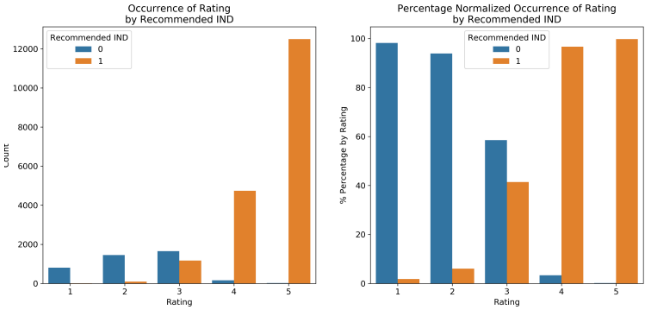


Figure 14: The frequency of rating by recommendation indicator.

(6) **Recommendation by Department Name and Division Name:** Figure 15 backs up what was found in Figure 11.

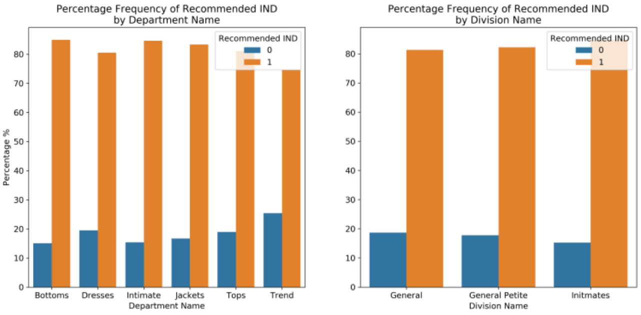


Figure 15: The percentage frequency of recommendation indicator.

(7) **Rating by Department Name and Division Name:** Figure 16 depicts the rating distribution's consistency.

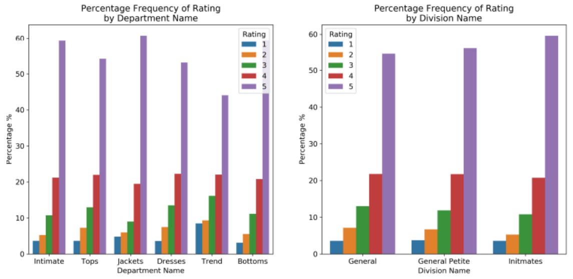
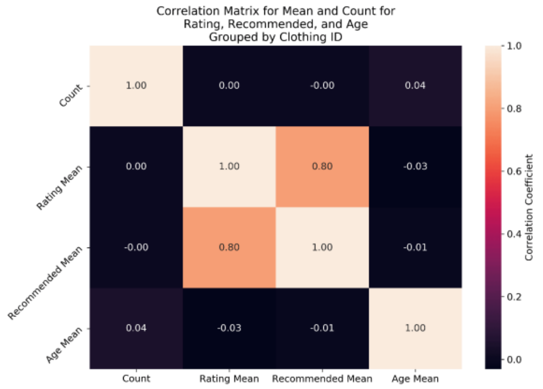


Figure 16: The percentage frequency of review rating by department name and division name

c) Multivariate Analysis and Descriptive Statistics.

(1) **Average Rating and Recommendation by Clothing ID Correlation:** Figure 17 examines the relationship, if any, between a product's average rating and the customer reviews for that product, as categorized by garment ID. The correlation matrix indicates that there is no such association between the variables studied, but it does show a significant 0.8 correlation between rating and suggestion. The previously indicated correlation coefficient supports the premise of a link between rating and suggestion.



	Count	Rating Mean	Recommended Mean	Age Mean
Count	1.00	0.00	-0.00	0.04
Rating Mean	0.00	1.00	0.80	-0.03
Recommended Mean	-0.00	0.80	1.00	-0.01
Age Mean	0.04	-0.03	-0.01	1.00

Figure 17: The correlation matrix grouped by clothing ID for average rating and recommendation.

(2) **Average Rating by Recommendation:** Figure 18 depicts consistency in rating and recommendation, i.e., when a review has a recommendation, the ranking is beneath the highest benefit of rating; when a review does not have a suggestion, the evaluation is halved.

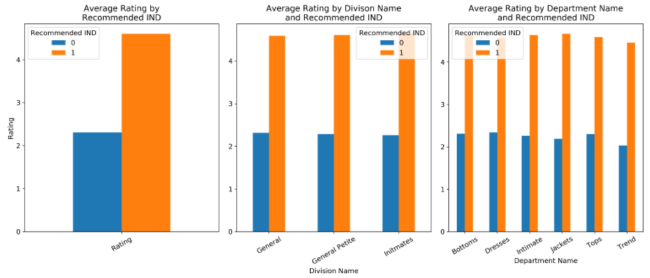


Figure 18: The average rating frequency per division department, and recommendation.

d) Word Frequency Distributions.

(1) **Most Frequent Words in Low-rated Comments:** Because Figure 19 is a term cluster for low-rated reviews, it's safe to infer that the terms in this figure correspond to what's written in the reviews.

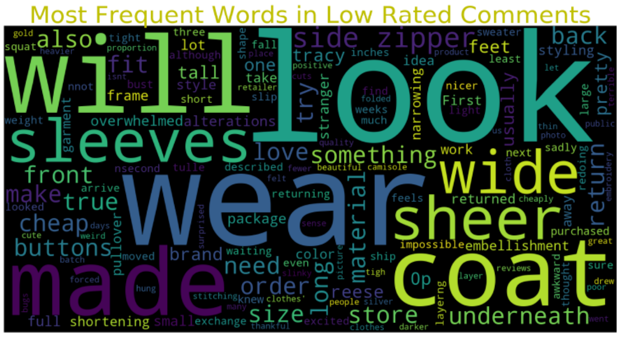


Figure 19: Most Frequent Words in Low-rated Comments

(2) Most Frequent Words in Highly-rated Comments: Because Figure 20 is a word cloud for high-rated reviews, it's safe to infer that the terms in this image correspond to what's stated in the comments.

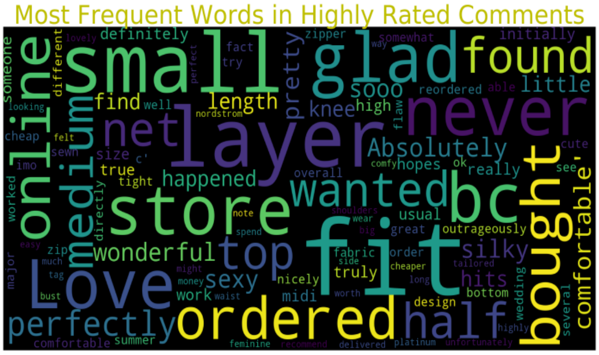


Figure 20: Most Frequent Words in Highly-rated Comments

(3) Titles: The most common terms in a reviewer title are shown in Figure 21. Only the phrase "flaws" seems to signify a poor review, but it does not mean the whole product evaluation is unfavorable. It's worth noting that this glossary only takes into consideration the frequency of words in titles, not phrases. In other words, negative word indicators may have counter-words, but they did not make it into the word cloud. The same may be stated of positive word signals in the word cloud, since it excludes any negators if any exist.

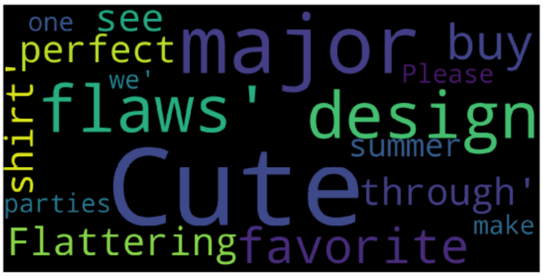


Figure 21: Most frequent words used for review title

(4) Word Clouds for Division Names: Figures 22 and show the most common terms in product reviews from the "intimates" division; Figure 23 shows the most common words in product reviews from the "general" division; and Figure 24 shows the most commonly used words in product information from the "general petite" segment. Further research into such word clouds may provide some important information about client acceptance by division.

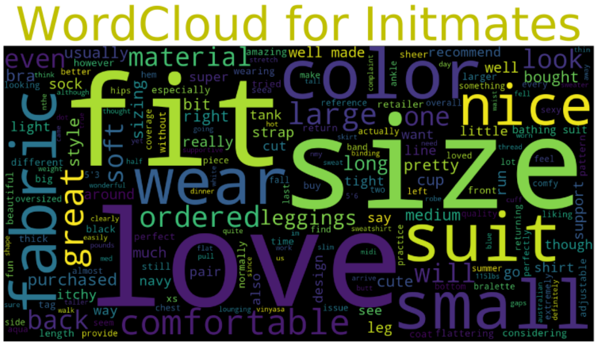


Figure 22: Most frequent words used in intimate apparels.

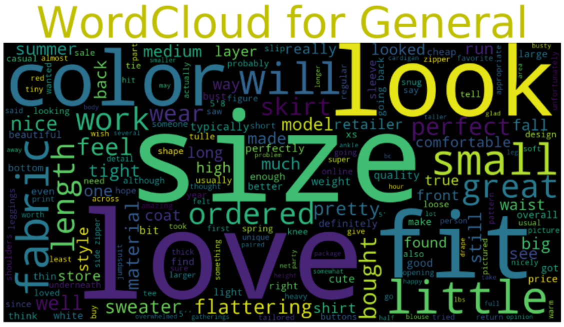


Figure 23: Most frequent words used in review texts in general-sized apparel

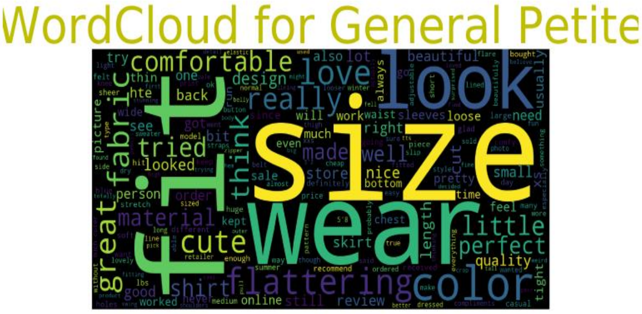


Figure 24: The most frequent words used in review texts in petite-sized apparels.

C. *Dataset Preprocessing*

(1) **Text Cleaning**: Delimiters like \r and \n were eliminated to clean the user review texts.

(2) **Word Embeddings**: The words in review texts were mapped to the vector space using GloVe word embeddings. GloVe (Global Vectors for Word Representation) is an unsupervised learning approach for obtaining word vector representations [20].

(3) **Sentiment Analysis**: NLTK's sentiment analyzer was utilized to automate the process of manually labeling the review texts. As a result, the intuitive

labeling of review texts that had a rating threshold of 3 (i.e., if a review rating is more than or equal to 3, it is deemed positive feedback; otherwise, it is considered negative feedback) has been abandoned. The above-mentioned manual, intuitive categorization has the drawback of overlooking certain neutral feelings. As a result, NLTK employs a sentiment analyzer [24]. The frequency distribution of attitudes per suggestion is shown in Figure 25.

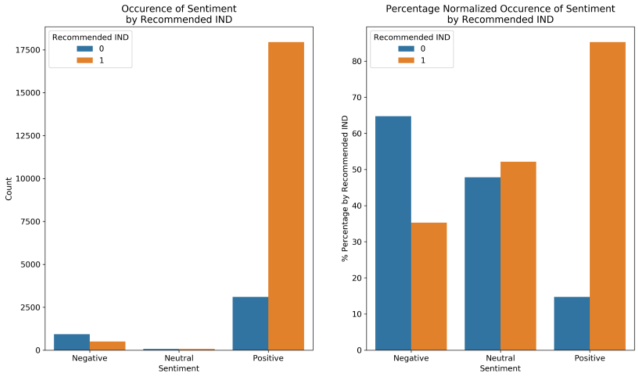


Figure 25: The frequency distribution of recommendation and sentiment

IV. METHODOLOGY

A. *Machine learning libraries used*

The bidirectional recurrent neural network (RNN) with long-short term memory (LSTM) [26] was implemented in this research using Keras [25] and Google TensorFlow. The numpy [27] and pandas [28] Python libraries were utilized for data preparation and handling. Finally, the matplotlib [29] and seaborn [30] Python libraries were utilized for data visualization.

B. *Machine Learning*

Given that the issue at hand is a sentiment classification job, the best machine learning technique to use is a recurrent neural network (RNN). However, we know from the literature that a vanilla RNN has diminishing gradients. As a result, we employed the RNN with long-short term memory (LSTM) units, which was created specifically to handle the situation at hand [31][32][38]. We also used a bidirectional RNN with LSTM to better capture the context of terms in the review articles (see Figure 26). That is, the model can learn the context of a text sequence from the "past" to the "future" and vice versa. As a result, the model gains a better understanding of each review text.

.

Bidirectional recurrent neural networks: Bidirectional recurrent neural networks (BRNN) are neural networks that link two hidden layers with opposing orientations to the same output. The output layer of this kind of generative deep learning may collect knowledge from both past (backwards) and future (forward) states at the same time. The BRNN idea is to divide the neurons of a standard RNN into two paths, one for positive time (ahead states) and the other for negative time (back states) (backward states). The outputs of the two states are not related to the inputs of the states in the opposite manner [33][34][35].

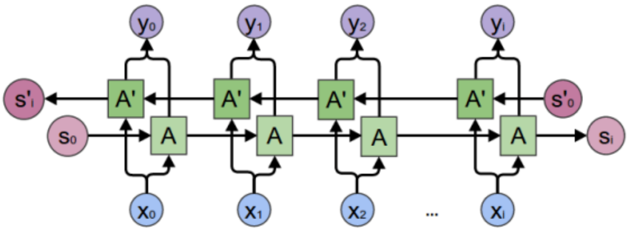


Figure 26: Bidirectional RNN

A standard Bidirectional RNN transfers input sequences x to target sequences y, with loss L(t) at each time step t, according to its computation. The RNN cells s propagate data forward in time (towards the right), while the RNN cells s ′ propagate data backward in time (towards the left) (towards the left) [6]. Thus, at each time step t, the output units o(t) may profit from a relevant summary of the past in its s(t) input, as well as a relevant summary of the future in its s'(t) input (before applying an activation function to obtain y).

Long short-term memory: LSTM is a deep learning architecture that uses an artificial recurrent neural network (RNN) [38]. LSTM features feedback connections, unlike normal feedforward neural networks. It can handle not just individual data points (such as photos), but also complete data streams (such as speech or video). A cell, an input gate, an output gate, and a forget gate make up a typical LSTM unit [39].

Why Bidirectional RNN with LSTM: Classic RNNs, in principle, may monitor arbitrary long-term relationships in input sequences. Because of the calculations involved in the process, which employ finite-precision numbers, the long-term gradients that are back-propagated may disappear, that is, they can trend to zero, or explode, that is, they can go to infinity, while training a vanilla RNN using back-propagation. Because LSTM units enable gradients to flow unmodified, RNNs utilizing LSTM units partly address the vanishing gradient issue [36][37].

The LSTM gate equations, which we built using Google TensorFlow, are shown below [6].

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * tanh(C_t) \quad (6)$$

Where, f is the forget gate, which "forgets" non-essential model information; I is the input gate, which receives fresh data input at a certain time step st; and C is the candidate cell state value of each LSTM cell; C is the cell state value to be passed on to the next RNNLSTM cell; o is the output gate that determines what the cell state will output; and h is the cell state output derived from the cell state value as well as the determined output.

On the dataset, we used this machine learning model to solve two text categorization problems:

(1) Recommendation categorization assesses whether or not a review text recommends the product under consideration.

(2) Sentiment categorization, which establishes the tone of the review text in relation to the acquired item.

V. Evaluation and Results

The data was divided in a 60/20/20 format, with 60% of the training dataset, 20% of the validation dataset, and 20% of the testing dataset.

Table 4 illustrates the hyper-parameters utilized in the experiments using the Bidirectional RNN-LSTM. These hyper-parameters were generated at random since hyper-parameter tweaking requires more processing resources. Table 5 demonstrates the Bidirectional RNN-test LSTM's accuracy and test loss in both recommendation and sentiment categorization studies.

Hyper-parameter	Value
Batch Size	256
Cell Size	256
Dropout Rate	0.50
Epochs	32
Learning Rate	1e-3

Table 4: Hyper-parameters

Task	Test Accuracy	Loss
Recommendation Classification	≈0.882678	≈0.572342
Sentiment Classification	≈0.928414	≈0.453205

Table 5: Test accuracy and Test loss using Bidirectional RNN-LSTM

Take notice, however, that the frequency distributions for classes in recommendation and sentiment are both skewed, i.e., there are more recommended courses than not recommended classes, and there are more positive feelings than negative and neutral sentiments combined. This is problematic because the model will create a biased classification in favor of the class with the largest frequency distribution. As a result, we look at Table 6 for a statistical report on suggestion categorization.

Class	Precision	Recall	F1-Score	Support
(0) Not Recommended	0.70	0.65	0.68	847
(1) Recommended	0.92	0.94	0.93	3679
Average / Total	0.88	0.88	0.88	4526

Table 6: Statistical Report on Recommendation Classification

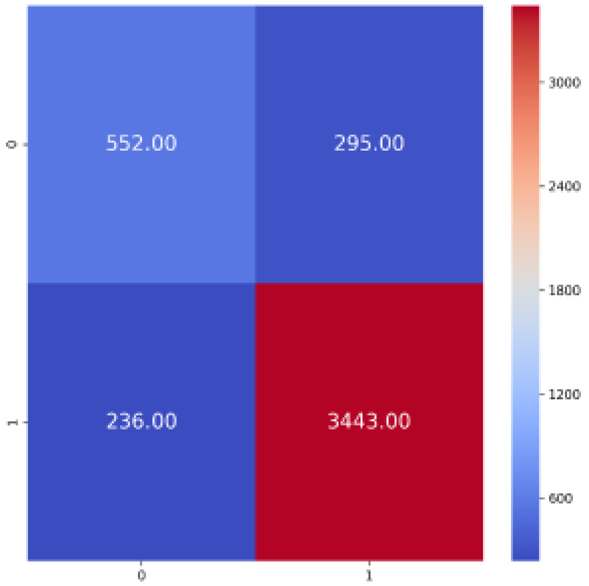


Figure 27: Recommendation classification confusion matrix

Table 6 reveals that negative class has a worse predictive performance in the recommendation classification issue, as seen in Figure 27's confusion matrix (where 0 represents not recommended class and 1 represents recommended class), confirming our conclusion. We examine the ROC curve for the outcome to assess how well the model performs on a pretty fair scheme (see Figure 28).

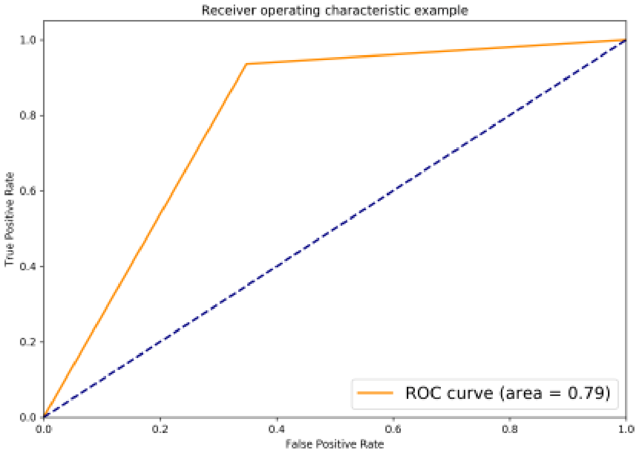
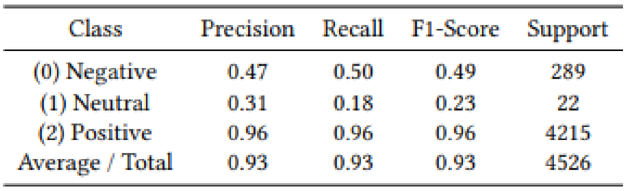


Figure 28: Recommendation indicator ROC Curve for binary classification



Class	Precision	Recall	F1-Score	Support
(0) Negative	0.47	0.50	0.49	289
(1) Neutral	0.31	0.18	0.23	22
(2) Positive	0.96	0.96	0.96	4215
Average / Total	0.93	0.93	0.93	4526

Table 7: Statistical Report on Sentiment Classification

Table 7 backs up our results on biased categorization in favor of the class with the largest frequency distribution, which is corroborated by Figure 29's confusion matrix (where 0 represents the negative class, 1 represents the neutral class, and 2 represents the positive class). In this report, we can observe that the model had a worse predicted performance for negative and neutral feelings.

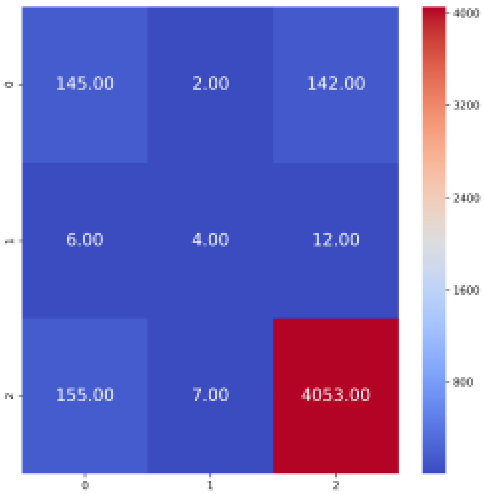


Figure 29: Sentiment classification confusion matrix

VI. CONCLUSION

Sentiment analysis, often known as opinion mining, is a branch of research that examines people's feelings, attitudes, and emotions about certain entities. On the Women's Clothing E-Commerce dataset, this study addresses a basic topic of sentiment analysis. In this project, we successfully constructed an autonomous sentiment and suggestion categorization system that learns from a huge collection of E-commerce women clothes reviews using machine learning methods. Sentiment analysis is a popular technique for extracting useful information from big volumes of data. Despite the dataset's skewed class frequency distribution, the empirical evidence given in this work suggests a reasonably high-performing prediction performance on both suggestion classification and sentiment classification. This finding backs up the assertion that Bidirectional RNN-LSTM catches the context of review texts better, resulting in superior prediction performance. However, for a fair comparison, we propose using a uni-directional RNN-LSTM on the identical classification tasks to back up this claim.

Hyper-parameter adjustment is required to enhance the model further. Due to computational constraints, this investigation was confined to a set of hyperparameters picked at random. Furthermore, kfold cross validation may provide us with a deeper understanding of the model's prediction ability.

Despite the constraints of this study's experiment, it can be concluded that the Bidirectional RNN-LSTM model performed well (with F1-score of 0.88 for recommendation classification, and 0.93 for sentiment classification). Furthermore, the categorization problem's statistical metrics may be regarded adequate.

REFERENCES

[1] Q. Sun, J.W. Niu, Z. Yao, H. Yan. “Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level”, Engineering Applications of Artificial Intelligence, vol. 81, pp. 68-78, 2019.

[2] J. Chen, G. Kou. “How Online Review Valance Affect on Consumer Opinion Evaluation?” Procedia Computer Science, vol. 81, pp. 635-641, 2016.

[3] Z.P. Fan, Y.J. Che, Z.Y. Chen. “Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis”, Journal of Business Research, vol. 74, pp. 90-100, 2017.

[4] L. Xiong. “Research on key technologies of e-commerce comment sentiment analysis”, Nanchang university, 2013.

[5] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

[6] Henderson M, Thomson B, Young S, et al. Word-Based Dialog State Tracking with Recurrent Neural

Networks[C]. annual meeting of the special interest group on discourse and dialogue, 2014: 292-299.

[7] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. arXiv: Computation and Language, 2013.

[8] Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using naive bayes and k-nn classifier. arXiv preprint arXiv:1610.09982.

[9] Shin, B., Lee, T., & Choi, J. D. (2016). Lexicon integrated cnn models with attention for sentiment analysis. arXiv preprint arXiv:1610.06272.

[10] Y. Fang, H. Tan and J. Zhang, "Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness," in IEEE Access, vol. 6, pp. 20625-20631, 2018, doi: 10.1109/ACCESS.2018.2820025.

[11] S. Zhang, D. Zhang, H. Zhong and G. Wang, "A Multiclassification Model of Sentiment for E-Commerce Reviews," in IEEE Access, vol. 8, pp. 189513-189526, 2020, doi: 10.1109/ACCESS.2020.3031588

[12] Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment analysis on Twitter data using KNN and SVM. Int J Adv Comput Sci Appl, 8(6), 19-25.

[13] Lakshmi, B. S., Raj, P. S., & Vikram, R. R. (2017). Sentiment analysis using deep learning technique CNN with KMeans. International journal of pure and applied mathematics, 114(11), 47-57

[14]A. S. Manek, P. D. Shenoy, M. C. Mohan, and K. Venugopal, ‘‘Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier,’’ World Wide Web, vol. 20, no. 2, pp. 135–154, Mar. 2017.

[15] G. Preethi, P. V. Krishna, M. S. Obaidat, V. Saritha and S. Yenduri, "Application of Deep Learning to Sentiment Analysis for recommender system on cloud," 2017 International Conference on Computer, Information and Telecommunication Systems (CITS), Dalian, 2017, pp. 93-97, doi: 10.1109/CITS.2017.8035341.

[16] L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for ECommerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in IEEE Access, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

[17] M. U. Salur and I. Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," in IEEE Access, vol. 8, pp. 58080-58093, 2020, doi: 10.1109/ACCESS.2020.2982538.

[18] Chen, C., Zhuo, R., & Ren, J. (2019). Gated recurrent neural network with sentimental relations for sentiment classification. Information Sciences, 502, 268-278.

[19] Zhou, L., & Bian, X. (2019, November). Improved text sentiment classification method based on BiGRU-Attention. In Journal of Physics: Conference Series (Vol. 1345, No. 3, p. 032097). IOP Publishing.

[20] Nick Brooks. 2018. Women’s E-Commerce Clothing Reviews.(2018). https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

[21] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.

[22] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Stéfan van der Walt and Jarrod Millman (Eds.). 51 – 56.

[23] Christopher Olah. 2015. Neural Networks, Types, and Functional Programming. (2015). http://colah.github.io/posts/2015-09-NN-Types-FP/

[24] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In Empirical Methods in Natural Language Processing (EMNLP). 1532–1543. http://www.aclweb.org/anthology/ D14-1162

[25] François Chollet et al. 2015. Keras. https://github.com/keras-team/keras. (2015).

[26] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

[27] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. 2011. The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering 13, 2 (2011), 22–30

[28] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Stéfan van der Walt and Jarrod Millman (Eds.). 51 – 56.

[29] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. Computing In Science & Engineering 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[30] Michael Waskom, Olga Botvinnik, Drew O’Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. 2017. mwaskom/seaborn: v0.8.1 (September 2017). (Sept. 2017). https://doi.org/ 10.5281/zenodo.883859

[31] J. O. Berger, Statistical Decision Theory and Bayesian Analysis. Berlin, Germany: Springer-Verlag, 1985.

[32] C. M. Bishop, Neural Networks for Pattern Recognition. Oxford, U.K.: Clarendon, 1995.

[33] H. Bourlard and C. Wellekens, “Links between Markov models and multilayer perceptrons,” IEEE Trans. Pattern Anal. Machine Intell., vol. 12, pp. 1167–1178, Dec. 1990.

[34] J. S. Bridle, “Probabilistic interpretation of feed-forward classification network outputs, with relationships to statistical pattern recognition,” in Neurocomputing: Algorithms, Architectures and Applications, F. Fougelman-Soulie and J. Herault, Eds. Berlin, Germany: SpringerVerlag, 1989, NATO ASI Series, vol. F68, pp. 227–236.

[35] C. L. Giles, G. M. Kuhn, and R. J. Williams, “Dynamic recurrent neural networks: Theory and

applications," IEEE Trans. Neural Networks, vol. 5, pp. 153–156, Apr. 1994.

[36] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 1990, pp. 1361–1364.

[37] R. A. Jacobs, "Methods for combining experts' probability assessments," Neural Comput., vol. 7, no. 5, pp. 867–888, 1995

[38] Hochreiter, Sepp, and Jrgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[39] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).