CUSTOMER REVIEW CLASSIFICATION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

By

Nikhil Sunil, BE Computer Science, Birla Institute of Technology Ranchi, 2022

A Major Research Project Report

presented to Ryerson University

in partial fulfilment towards the requirements for the degree of

Master of Science (M. Sc.)

In the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2022

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Nikhil Sunil

CUSTOMER REVIEW CLASSIFICATION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

Nikhil Sunil

Master of Science 2022

Data Science and Analytics

Ryerson University

ABSTRACT

In today's world, the online ecommerce industry has become very competitive and it continues to grow at a rapid pace. Companies generate a lot of data, that contain customer feedback data like reviews about their products and services. Customer online reviews play a very important role in helping the company improve their sales and increasing their customer base.

In this paper, we take different customer reviews on women's clothing and classify them as good or bad, that can help in making a decision on whether the product/service is doing well in the market or not. To achieve this, I will be using various classification models in deep learning and machine learning to classify the customer reviews and compare their accuracies and other parameters, to decide which model would be the best fit for this task. The machine learning models proposed here are Logistic Regression, AdaBoost, Decision Tree, Support Vector Machine, Random Forest and deep learning models like LSTM, Bi-LSTM and GRU.

The models are evaluated by using various metrics like confusion matrix, AUROC curve and classification report.

Keywords: Text Classification, Logistic Regression, AdaBoost, Decision Tree, Random Forest, LSTM, Bi-LSTM, GRU

ACKNOWLEDGEMENTS

I would like to thank my supervisor **Dr. Farid Shirazi**, who was instrumental and very helpful in giving suggestions, during the planning and completion of this research project. Dr. Farid was my supervisor for this MRP; he has been a great mentor, guiding me in the right direction and providing advice and insights whenever required.

Table of Contents

List Of Figures

List Of Tables

1. INTRODUCTION

Customer reviews are pieces of feedback that are given to businesses and Retailers based on the customer's experience, with their organization. They are used by companies to improve upon their existing service or the product they are selling.

In an e-commerce driven world, where people have no physical access to the goods they wish to purchase, many customers will turn to online reviews to get an opinion on what to buy.

E-Commerce is really changing the way in which, people buy products and services. To move up in the corporate world, many businesses are finding different ways to increase their customer count by looking at past data. Companies can use customer reviews on their products to promote their high selling promote their products and gain a competitive edge in the market. This paper focusses on classifying the sentiment of customer reviews with the use of modelling techniques, and the results vary with each technique.

We first start off by giving a brief explanation of the dataset, define the research question, followed by the literature review and an exploratory data analysis on the dataset, to gain some insights into the data we're working on.

The methodology section talks about the experimental design i.e., how the data is processed before the they are used to train the models for classification. We then go over the modelling results and compare them with appropriate metrics, to select the best performing model.

A. Research Question

The focus of this research paper is split into two sections. The first part comprises of the Exploratory data analysis on the dataset, to find some patterns that may help us understand the data better and answer questions about company is performing in terms of clothing sales in each department and ratings from customer on those products.

The second part involves training and testing machine learning and deep learning classifiers to correctly classify the sentiment of customers based on their reviews and evaluate their performances to select the best performing classifier.

B. Dataset

The selected dataset contains reviews written by customers purchasing different type of clothing. It has 9 important features and will instantiate a perfect way to classify the reviews for further analysis. The data has been anonymized and credit has been given to the company, since the data is commercial. The reference to the company in the review text and body has been replaced with the text 'retailer'

This Dataset has 23486 rows and 10 feature variables. Each row corresponds to a customer Review and includes the variables. As the data is filled with noise, I will perform intensive data cleaning to filter out the unwanted data to make it ready for model testing.

The below are the features of the dataset:

- **Clothing ID**: reference for the clothing being reviewed (categorical variable).
- **Age:** variable depicting the reviewer's age(Positive Integer variable).
- **Title:** the heading sentence for the review (String variable).
- **Review Text:** review of the customer (String variable).
- **Rating:** rating score on a scale of 1(worst) to 5(best) (Positive Ordinal Integer variable)
- **Recommended IND:** customer recommendation of a product between 1 and 0.(Binary variable) 1- recommended, 0- not recommended
- **Positive Feedback Count:** keeps count of the customers that found this review to be helpful
- **Division Name:** Description of the product high level division (Categorical variable).
- **Department Name:** product dept name (Categorical variable).
- **Class Name:** product class name (Categorical variable).

C. Variables

The variables that are the main focus of this experiment are

- Review Text - the customer reviews
- Recommended IND - the recommendation signal

2. LITERATURE REVIEW

To get a more detailed understanding of use of customer reviews, I started off, by reviewing articles on the impact of online reviews in the market.

A 2019 paper on impact of Online consumer reviews on hotel Booking, explained about the influence that reviews have on the decision of customers. The explained that online reviews are like an 'electronic word of mouth' that have changed the way people buy stuff with more internet usage. Customers are more inclined to buy products, if the advertisements and reviews of the product are realistic. The paper has stressed that businesses need understand the importance of online reviews use them to devise strategies, to improve sales [1]. Another paper in 2020, on Sentiment Analysis on an ecommerce product, using an Indonesian dataset explained how different techniques in Machine learning can be used to classify reviews. The paper concluded that, the best accuracy was achieved by applying the TF-IDF and Backward Elimination in SVM which performed well with a score of 85.97%, that goes up by 7.91% after applying feature selection [2]. Sudhakaran and Jaiganesh conducted sentimental analysis on a SNAP Dataset that contains reviews on Amazon. The goal of this paper was to use a popular Machine learning Algorithm called a Support Vector Classifier to classify data. Based on their experiments, they concluded that SVM classifier did perform better than naïve bayes and Random Forest Classifiers making them suitable for classifying large data [3]. This paper talks about a proposed amazon web analysis app that uses ml models to classify reviews as negative or positive Different models were used in combination with different feature extraction methods which were tested on data from the dataset. The researchers concluded from their experiments that, the logistic Regression model combined with count vectorizer, gave the best performance with an accuracy of 0.9339 [4].

Two papers on Exploratory and Sentiment Analysis on Netflix data [18] and [19] give a detailed overview of how review data from Netflix is used to get insights and also perform sentiment analysis on them. This 2021 paper on analysis of covid-19 tweets using sentimental analysis aims to do an analysis of tweets by Indians during the Covid-19 lockdown. The text from the tweets have been put into 4 different categories- fear, sad, anger, and joy. To predict the sentiment of these tweets, data analysis was conducted using 4 models (Bert, Logistic Regression, Support Vector Machine and LSTM). The Bert Model surpassed the other models in term of performance (89%). The researchers came to conclusion that the government needs to perform fact checks to avoid spread of false information. Using the findings from this research, the public authorities can work to overcome needless anxiety during pandemics [5]. The paper on 'Determinant Factors of E-commerce Adoption by SMEs in Developing Country' investigates factors that

influence SMEs in developing countries in adopting e -commerce. In this study, 11 variables namely, perceived benefits, compatibility, cost, technology readiness, Firm size, Customers/suppliers pressure, competitor pressure, external support, Innovativeness, IT ability, IT experience were studied to see if they are relevant to an SME's success in business. Finally, they were grouped into 4 groups: technological contexts, organizational context, environmental contexts and individual contexts, which were identified as factors that affect the Indonesian SMEs in their adoption of E-commerce. The paper concludes that various factors like technology readiness, innovative ability, IT experience and IT ability are substantial to the success of SMEs in developing countries like Indonesia [6]. This paper briefly talks about the various methods for sentiment analysis on tweets to give us a good overview of the field. The paper covers topics like sentiment monitoring, Twitter opinion retrieval over time, emotion recognition, irony recognition as well as other topics which have a relation with sentiment analysis of tweets. It also talks about the use of various supervised learning techniques like Maximum Entropy, Support Vector Machines, Random Forest, Naive Bayes, Logistic Regression, and Conditional Random Field [7].

Twitter is a platform that is widely used by people all over the world. Twitter data(tweets) is widely used to analyze sentiments of various tweets. The paper on study of Twitter sentimental analysis, gives a great overview of how tweets are collected and processed before they're used with machine learning models, for feeding data, in order to train them [20]. Another 2016 paper on "Techniques for Sentiment Analysis of Twitter" talks about the different techniques used to carry out sentiment analysis on twitter data. The key techniques to prepare the data are as follow. Firstly, the collected data is pre-processed to remove noise. After this, we extract important features from the data. The data is then labelled as positive or negative to prepare the dataset, which goes as input to the model classifier for training. Small part of the dataset is kept aside for testing purposes. The writers also applied various supervised machine learning based on identified parameters [10]. This paper implemented a sentiment classification approach using deep learning algorithms such as LSMT and CNN and hybrid CNN and LSTM models to predict the sentiment of reviews. Deep learning networks like CNN, LSTM and other hybrid models of CNN and LSTM were applied on data from the imdb dataset. The results have shown that, the hybrid CNN_LSTM model have outperformed the MLP and singular CNN and LSTM networks with a high accuracy rate of 89.2% [8]. Another paper on movie review classification written by gurshobit and ankit focused on using feature based opinion mining, speech tagging and supervised machine learning techniques to perform sentiment analysis of movie reviews [17]. The 2020 paper on Sentiment Analysis in E-Commerce-review of techniques and algorithms', tackles a comprehensive overview of sentiment analysis and relevant techniques in e-commerce industry, that is always keen to find out about the consumers' opinions of their

goods and services. The writers talk about how companies nowadays, are using social media as a tool to analyzing sentiments to find various trends, in order to achieve business value such as customers satisfaction, customer reputation while achieving high revenue and revenue. One of the challenges with sentiment analysis is that sometimes, exaggeration by users in reviews, cannot be easily picked by the models and they tend to classify them in one way, while in reality the sentiment is the opposite [9].

In order to explore other methods of sentiment analysis, I came across a technique known as Lexicon Sentimental analysis used by companies, which is another useful way of classifying the sentiment. This method makes use of predefined list of words, where each word is associated with a particular sentiment [11]. The paper on “Lexicon based methods for Sentiment Analysis” by Maite Taboada and Manfred Stede, proposed the Semantic Orientation Calculator to extract sentiment from text using dictionaries of words annotated with polarity and strength. The strength/intensity of sentiment word groups in a dictionary can be expressed as a number [12]. Another 2016 paper on lexicon feature extraction for emotion text classification had presented a unigram mixture model (UMM) based DSEL, through the usage of labeled & weakly-labeled emotion text to get important features for emotion classification [13]. A 2016 paper on Lexicon-enhanced Sentiment Analysis using rule based classification, came up with the idea to combine emoticons, modifiers and domain specific terms to find some trends in the reviews posted in online communities [14]. A 2001 paper on Sentiment parsing from small talk on Web, written by Sanjiv R Das and Mike Y Chen, proposes a technology for extracting investor sentiment from web sources. The method uses different classification algorithms for analyzing the sentiment of any message posted on the chat board [15].

The 2018 paper on Sentiment Analysis for Election Results followed a step-by-step approach to data collection, pre-processing of data and machine learning analysis to predict the result of the 2020 US presidential Election using Twitter emotional analysis. They use random forest classifier to predict the sentiment of users from tweets. They achieved an accuracy of 83.22% with trump's tweets and 85.73% with Biden's tweets [16].

3. EXPLORATORY DATA ANALYSIS

D. Data Acquisition

The data for this project was obtained from Kaggle and was prepared by user named nicapotato. The link for the dataset is [women's-ecommerce-clothing-reviews] and has a CC0: Public Domain license, that satisfies the MRP requirements.

E. Data Analysis and processing

The data is first loaded from the csv file into a dataframe using the pandas library. Before we perform EDA, we do some data cleaning, by removing duplicate and null values from the dataframe.

- Customer analysis

a. Proportion of customer recommendations

To get an idea about how the customers are giving recommendations, I created a pie chart to visualize this insight. Figure 1 shows that more than 80% of the customers are satisfied and have highly recommended the products they have purchased.

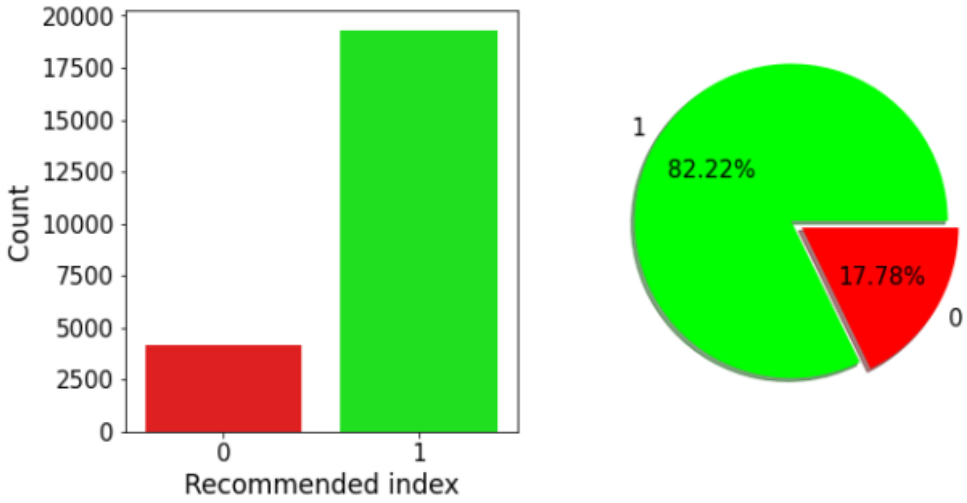


Figure 1– Proportion of customer recommendations before upsampling

While training our model, it is important that the proportion of 1(recommended) to 0(non-recommended) is very high. According to a blog written on data scaling [22], many of the machine learning models like logistic regression face issues with performance whenever the

difference in ratio between the number of data points is imbalanced. To resolve this, we go with the process of upscaling, to balance the data.

After upscaling, the data appears to be well balanced as seen in figure 2.

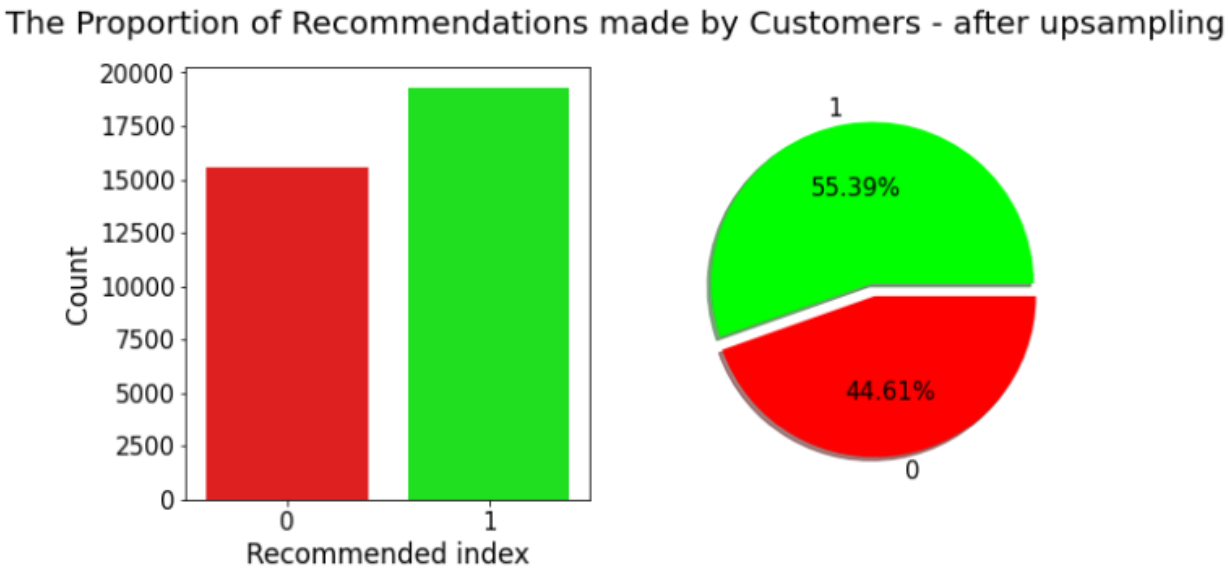


Figure 2 -Proportion of customer recommendations after upsampling

b. Proportion of customer Ratings on all products

Ratings provided by customers play a big role in predicting the sentiment of the customers. Based on this info, companies can provide good offers and discounts to their customers to promote more sales.

In the donut chart in figure 3, more than 50% of the customers have given a 5-star rating for all the products, up to 35% of the customers have given 3 and 4 star rating while the remaining 10% have given a lower rating.

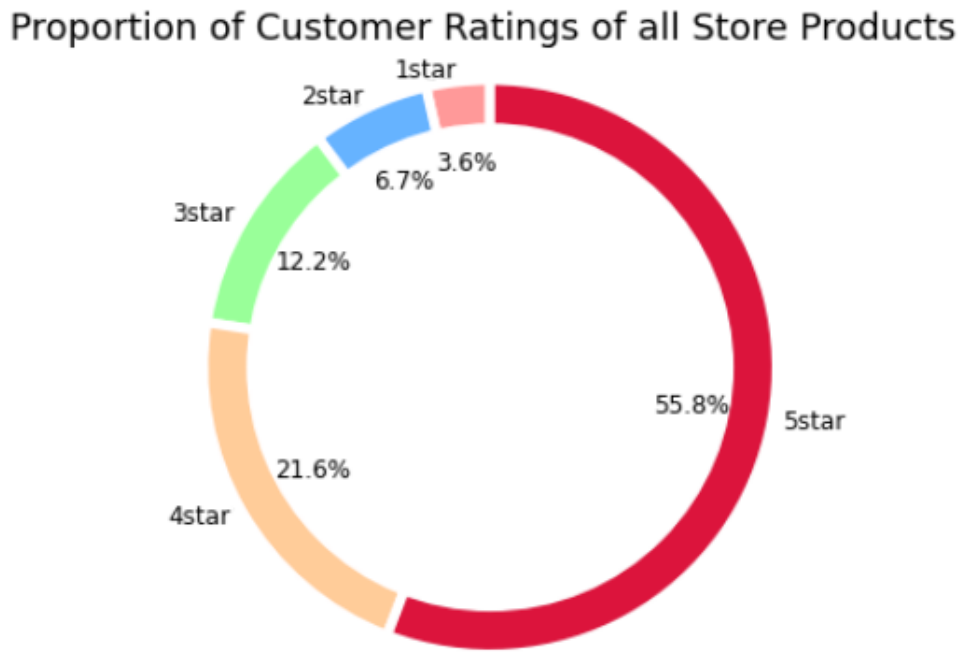


Figure 3- Proportion of customer ratings on all store bought products

c. Age distribution of customers

From the graphs in figure 4, we can identify the age group of the company's customers. A higher concentration of customers are in between the age group of 30 and 50. As the age value increases, the customer count keeps declining.

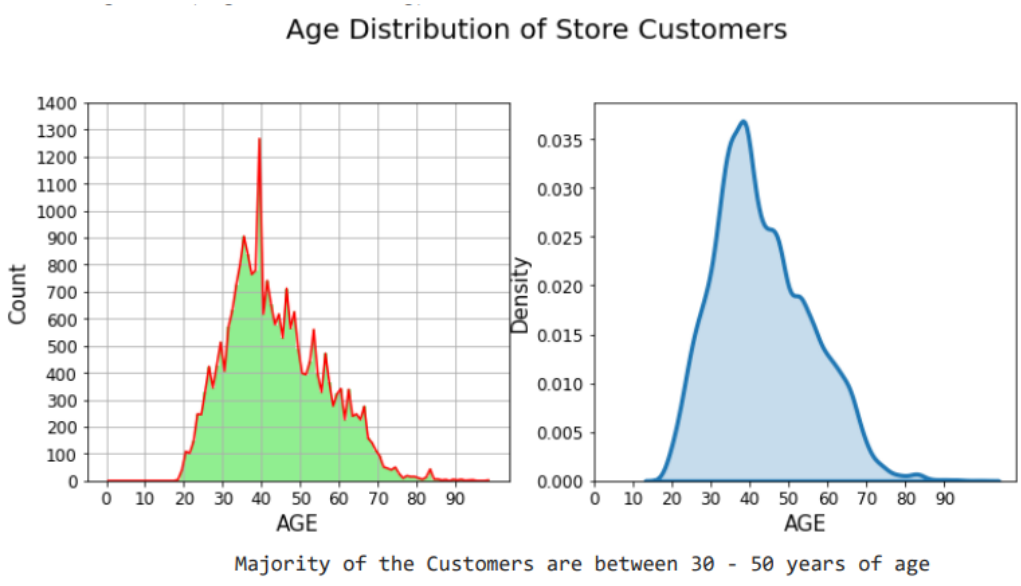


Figure 4- Age group of customers making purchases

- Product trend analysis

  a Products purchased from each department

  The clothing store has 3 departments- namely the general department, Intimate clothing department and the General petite department. According to the pie chart in figure 5, almost 60% of the store purchases have been made from the general department, making it the popular choice among the customers. The general petite department comes second in terms of clothes sold with a customer base of almost 35%, while the intimate department sells under 7%.

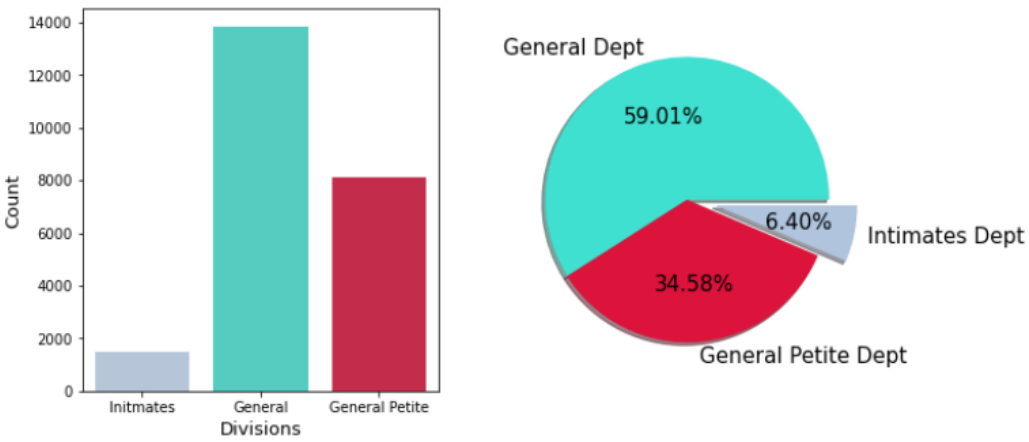


Figure 5- Proportion of purchases made in each product divisions in the store

b Popular choice of clothing among the customers

According to the visualization in figure 6, clothing like Tops and Dresses are the top choices among customers with just 70 % of clothes sales. A very low percentage of customers have opted to buy Trendy clothes and jackets, making them the least sought out clothes at just under 7%. Intimate clothing and Bottoms make up just about 25% of the total store sales.

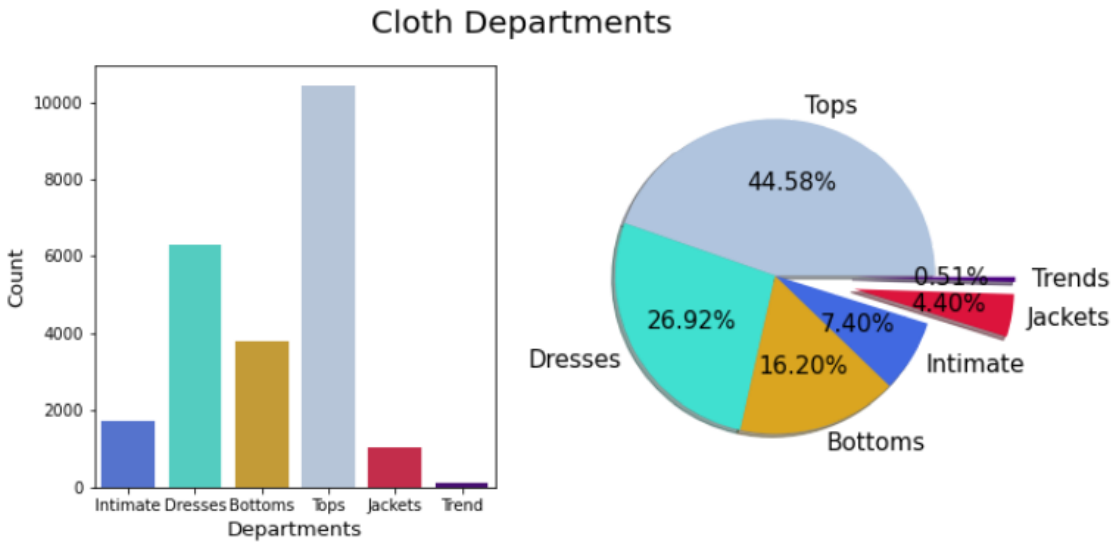


Figure 6- Different types of clothes purchased by customers

c Clothing recommendations made by customers

As per the scatterplot in figure 7, it is confirmed that almost all of the products have been highly recommended by the customers between the age of 20 and 70, the scatterplot shows a high concentration of 1's(good) and lower amount of 0's (bad).

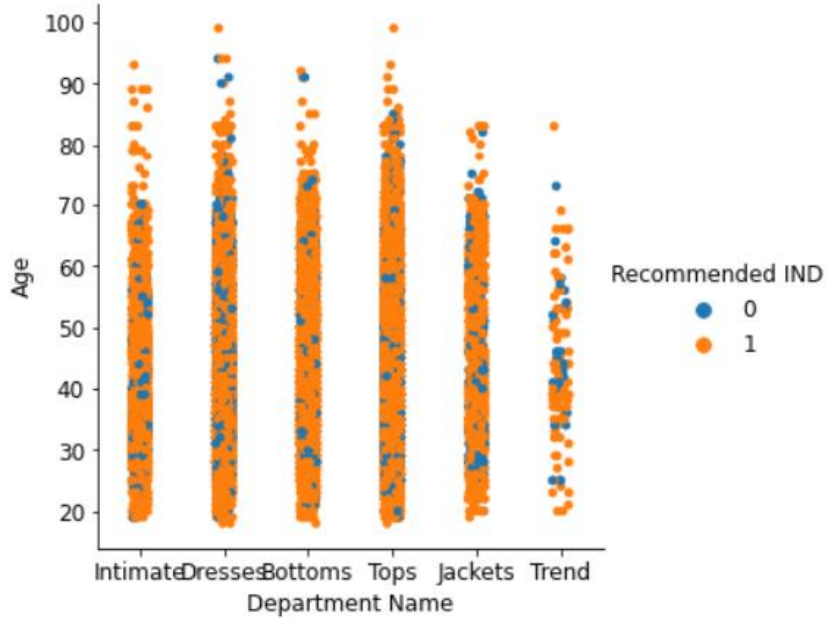


Figure 7- Recommendations made by customer of different age groups on different clothing

d. Ratings vs Recommendations

As per the visualization in figure 8, products of rating 4 and 5 have been highly recommeneded while products of rating 1-3 have comparitively lower positive recommendations. This plot shows us the relationship between the ratings and the recommendation data.

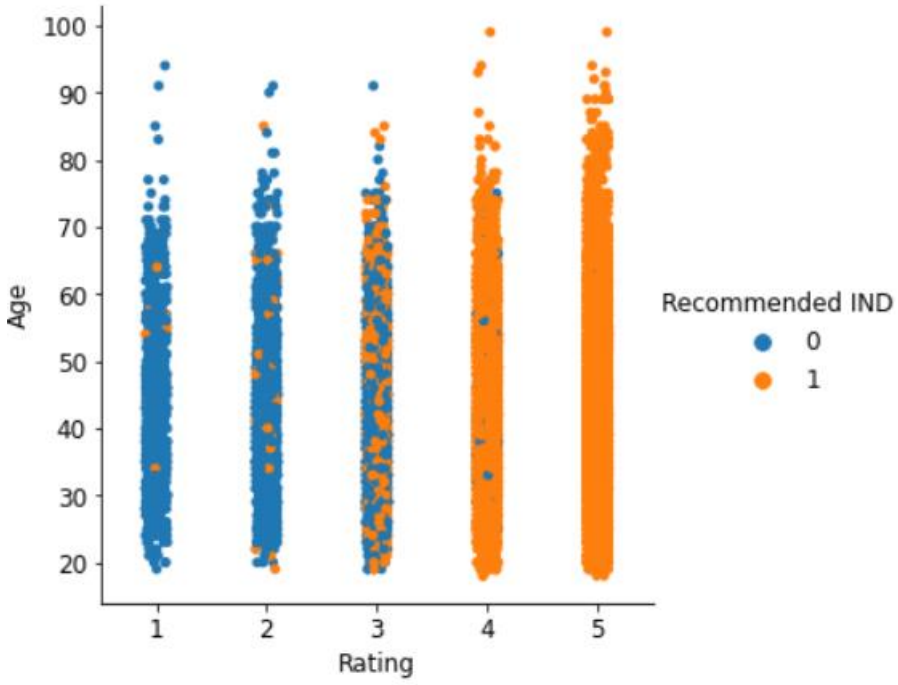


Figure 8- Ratings vs Recommended ID

4. METHODOLOGY AND EXPERIMENTS

F. Aim of Study

The aim of study is to perform a sentiment analysis, to predict whether the product is recommended by the customer based on the reviews. To find out the prediction, 5 machine learning models and 3 deep learning models were trained, tested and compared to determine the best fit for this process, which are namely Logistic regression, Linear Support Vector Machine, Decision Tree, Random Forest, AdaBoosting, Long short term memory(LSTM), Bidirectional LSTM and Gated Recurrent Unit (GRU).

G. Dataset

- Dataset details

  The dataset has 23,486 rows and 10 columns. It is based on the ecommerce marketplace that contains reviews written by the customers and is supported by 9 other extra features that give us a little more insight into the pattern of reviews.

```
[4]  df = pd.read_csv("Womens Clothing E-Commerce Reviews.csv")

[5]  df.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 23486 entries, 0 to 23485
     Data columns (total 11 columns):
      #   Column                   Non-Null Count  Dtype
     ---  ------                   --------------  -----
      0   Unnamed: 0               23486 non-null  int64
      1   Clothing ID              23486 non-null  int64
      2   Age                      23486 non-null  int64
      3   Title                    19676 non-null  object
      4   Review Text              22641 non-null  object
      5   Rating                   23486 non-null  int64
      6   Recommended IND          23486 non-null  int64
      7   Positive Feedback Count  23486 non-null  int64
      8   Division Name            23472 non-null  object
      9   Department Name          23472 non-null  object
      10  Class Name               23472 non-null  object
     dtypes: int64(6), object(5)
     memory usage: 2.0+ MB
```

Figure 9- Dataset Features

- Duplicate and Null Values
  - Duplicates: - the dataset contains duplicate rows which are dropped from the data frame

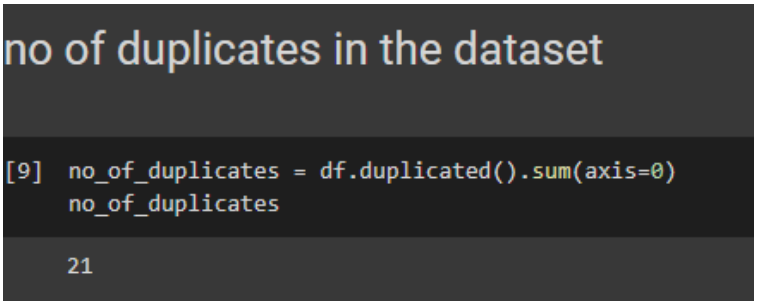


Figure 10- Duplicates

- Missing Values: - The values in the dataframe like null, N/A are referred to as null values. This is done by using the dropna () function [25].

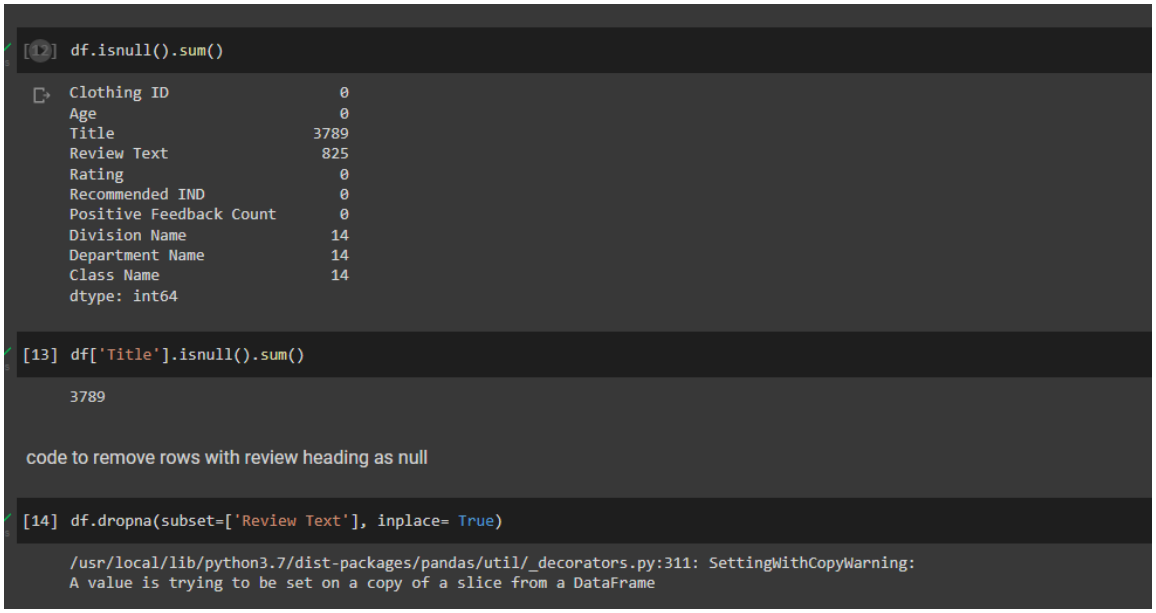


Figure 11- Removing null values - Experimental Design

H. Text Processing

To remove unnecessary symbols and other noise in the reviews, we need to perform the following techniques to attain the best model performance.

- **Removing Punctuations**: Characters that exist in the text apart from alphabets and whitespaces are removed. The “n’t” in words like “wouldn’t” are removed as well. We use the regex function substitute(re.sub) to remove the punctuation marks from the comments [23].
- **Tokenization**- It is the process of splitting a large text sample into a number of words or substrings. This is a common practice in Natural Language processing to for classifying a particular sentiment.

- **Removing numbers**- We use a common function in python called 'isalpha()' to differentiate the numbers from the text. Removal of numbers can help models focus on more important words.
    - **Filtering stop words**- While analyzing texts or performing nlp operations, stopwords like 'the', 'is', 'in', 'for'… etc may not add much meaning to the reviews. Some of the key advantages of doing this are as follows
        - Dataset size decreases
        - Accuracy of results is better
    - **Lemmatization** – It is the process of obtaining the root words from a particular word [30]. This method is applied to reduce the number of unique words in the Reviews, which reduces the model training time. We use the WordNetLemmatizer() to perform the lemma operation in our processing.

I. Randomization and Stratification

The dataset was randomly divided into two sets. 80% of the data is used for training the model while the remaining 20% is used for testing. We use the stratify parameter to ensure that the data is evenly split between the training and test set. It ensures that data is split evenly between each class in training and testing data [26].

J. Feature Selection

The dataset contains 11 columns. For this experiment, we will be using the 'Reviews Text' and 'Recommended IND' columns to train our models, to predict the recommendation index using the review.

K. Modelling and metrics used

a. Machine Learning Models

For predicting the sentiment of the reviews, we will be using 5 machine learning models, namely Logistic Regression, Support vector machine classifier, Decision Tree classifier

b. Deep Learning Models

We will be using mostly Recurrent neural networks for predicting the sentiment of customer reviews, as these types of models are the appropriate for text or sentence learning. The models that will be used in this research are LSTM, GRU and BI-LSTM.

c. Metrics used

To compare the performances of the models, the best metrics proposed are as follows

- **Confusion Matrix**

  It is a table that is used to visualize the performance of a model on a set of test data. The confusion matrix provides 4 different combinations of Predicted and Actual values [27]. We will evaluate 4 different parameters to compare the model performance

  - **True positive**

    It is the number of times that the model predicts the positive class correctly.
  - **True negative**

    It is the number of times that the model predicts the negative class correctly.
  - **False positive**

    It is the number of times that the model predicts the positive class wrongly.
  - **False negative**

  It is the number of times the model predicts the negative class wrongly

- **AUROC curve**

  Area under the Receiver operating characteristics tells us how much the model is capable of distinguishing between classes. Higher the area, the better the model is at predicting the sentiment.

  - **<u>Roc- probability curve</u>**

  It is calculated by using the actual results and the predicted results.

  We calculate the true positive rate values and the false positive rate values from them to plot the curve. The formula for them is as below

  True positive rate = True positive/ (True positive + False negative)

False positive rate = False positive/ (False Positive + True Negative)

- **Auc- degree of measure of separability**

  It is a perfect metric for measuring the performance of any model. Higher the value of auc, means that the model has done well in terms of classification [28].

- **Classification report**

  The classification report gives us a detailed report of the model scores such as accuracy, precision, recall, f1-score and support [29].

L. Experimental process/implementation

a. Data – preprocessing

To fit the data into the model it needs to be cleaned to remove noise and other anomalies to get accurate modeling results. After some research, I found the needed methods to complete this process. We remove punctuations and numbers that are irrelevant to the review comment, and then use tokenization to get the list of tokens(words) from a comment and further simplify them by lemmatizing them(converting words to their root words based on the dictionary).

b. Training the models

Initially, I focused on training the models on data that has not been upscaled to check the difference in performance. We have divided the data into 80-20 train-test ratio to allow the model to learn well.

c. Model Building

- **Using TF-IDF with the models**

Once the data is prepared, we convert the data into a vector form using the TF-IDF vectorizer, which transforms the words in the comment into a vector matrix.

TF-IDF(term frequency - inverse document frequency) is a vectorizer that converts words into vector matrix, where each row represent the frequency of words in all the comments [24]. The formula for TF-IDF is as below.

$$w_{i,j} = tf_{i,j} \times idf_i$$

Figure 12- Formula for TF-IDF

Wi,j -> TF-IDF score for a word 'i' in a doc 'j'. Words with a higher TF-IDF score have higher occurrence, as opposed to those with lower score.

d. Measuring Classifier Performance

We use the various model performance measuring methods like confusion matrix, AU-ROC curve and classification table.

Any model that has a higher true positive and true negative probability count, will have performed the best out of the rest. Model with a larger area of the auroc curve has a higher accuracy rate.

We'll be using the classification table to look at the accuracy, precision and recall values that will give us better insight into how the models differ in performance.

5. RESULTS AND DISCUSSION

M. Exploratory Analysis Results

The exploratory data analysis on the customer review data provided some impactful insights on how the company is doing in terms of clothes sales. For instance, more than 50% of the customer have given a 5-star rating for their products while only less than 15% have given 1 or 2 stars. This confirms that most of the company's products are very popular with their customers and are in line with the current fashion trends, that are popular these days.

An analysis conducted on the customer data confirmed that most of the buyers were in the age group of 30-50 and their preferred choice of clothing were Tops and dresses. Under 25% of the purchases included bottoms and intimate clothing and very few customers purchased jackets and trendy clothing.
The company needs to work on providing more options, to improve their sales in this category.

The company sells clothing from 5 different categories namely Tops, Dresses, intimate, Bottoms, Jackets and trendy clothing. As per the catplot in the ED analysis, almost all the products have high recommendations from customers of all age groups. This is a sign that company is doing a good job overall with product selection by following the current trends in the design/fashion industry.

Since the recommendation count is unevenly balanced, we have tested our models with the original and the up-sampled data to show the difference in performance. Modelling results with up-sampled data seems to perform slightly better than the original data.

N. Modelling Experiment Results

Before predicting outcomes using modelling techniques, the data to be used needed to be cleaned in proper way to fit the model. We performed the following experiments as below.

a) Experiment 1

While performing EDA, I noticed that there was an imbalance in the number of recommendations made. The number of good recommendations was significantly more than the number of negative recommendations. It is known that up sampling the data would be the right move to improve the accuracy of the model [21]. To test this, I performed modelling with both types of datasets and I found

an increase in accuracy when working with up sampled data. In the below tables, I have highlighted the scores of model working with up sampled data.

	Train_Score	Test_Score	Precision_Score	Recall_Score	F1_Score	accuracy
logisitc regression	0.918018	0.893633	0.908071	0.895092	0.901535	0.893633
lr for data without upscaling	0.910280	0.891564	0.904451	0.970065	0.936109	0.891564

	Train_Score	Test_Score	Precision_Score	Recall_Score	F1_Score	accuracy
Support Vector Machine	0.94076	0.913146	0.938626	0.899137	0.918457	0.913146

	Train_Score	Test_Score	Precision_Score	Recall_Score	F1_Score	accuracy
SVM on data without upscaling	0.844247	0.833922	0.831986	0.998921	0.907843	0.833922

	Train_Score	Test_Score	Precision_Score	Recall_Score	F1_Score	accuracy
Decision Tree	1.0	0.910065	0.981038	0.851133	0.911480	0.910065
Decision Tree for data without upscaling	1.0	0.812279	0.884760	0.886192	0.885476	0.812279

	Train_Score	Test_Score	Precision_Score	Recall_Score	F1_Score	accuracy
Random Forest Tree	1.0	0.970217	0.978172	0.966828	0.972467	0.970217
random forest for data without upscaling	1.0	0.847836	0.846932	0.993797	0.914506	0.847836

	Train_Score	Test_Score	Precision_Score	Recall_Score	F1_Score	accuracy
AdaBoost	0.901181	0.870745	0.876632	0.887271	0.881919	0.870745
Adaboost for data without upscaling	0.864013	0.858436	0.884818	0.950917	0.916677	0.858436

Figure 13- Modelling scores after working with and without upscaled data

b) Experiment 2

To get optimal results with modelling, we need to make sure that our data is cleaned in the right way. After some conducting some research, I found the most suitable methods for performing text cleaning. Many reviews are often filled with unnecessary numbers, punctuation marks and other common stop words, which would only make the modelling processes much longer. Therefore, the following methods in methodology section were implemented to achieve this.

After implementing the proposed models, we came to the conclusion that random forest model has outperformed most of the other models in certain aspects. The data was evenly passed to the models using the stratify method which ensures that there is an equal ratio of imbalance in the model error if any.

c) Model hyperparameters

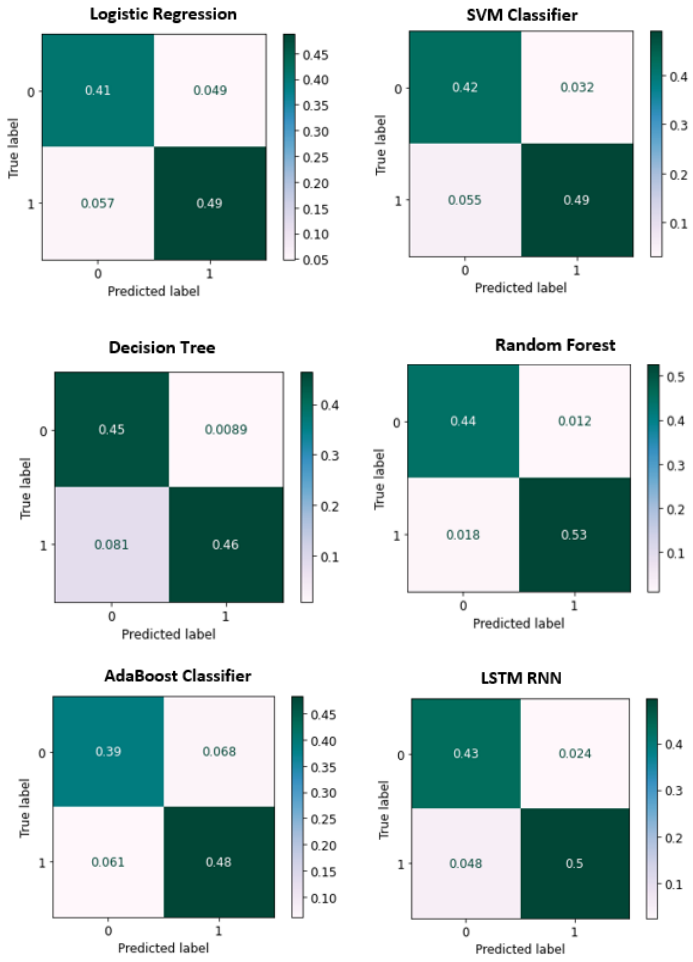
Hyperparameter testing has been done for certain models to reach the expected accuracy rates. For SVM model, we set the kernel to 'radial basis function'(rbf), random states are set to 0, the regularization parameter 'C' is set to 0.2 and the probability parameter is set to true to enable

probability estimates. In the AdaBoost classifier, the hyperparameter n_estimates is set to 100 initially and was increased gradually to see change in performance. The best result was achieved with n_estimators set to 500 with random_state of 0.

d) Performance Metrics

We will be using the confusion matrix to review the number of correct and wrong classifications made by all the models. The classification table will be reviewed to check the model scores on test and training data and the auroc curve will be checked to compare the area under curve score and the roc curve for each and every model.

- Confusion Matrix



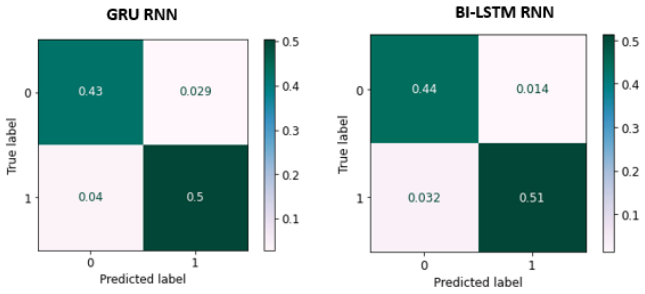


Figure 14- Confusion Matrix for all the modelling techniques

In the confusion matrix here, the True positive (TP) is the number of bad review recommendations that have been correctly classified by the model and the True Negative is the number of good review recommendations that were correctly classified by the model. On the other hand, false positive and false negative are the opposite, where the first is the number of bad review recommendations that were wrongly classified and the latter is the number of wrongly classified good review classifications.
We have normalized the values between 0 and 1, for comparing the values of all models.

Although all the models have done reasonably well, Random Forest has edged out the models in terms of parameter values. it has the second highest value of True positive (0.44) and highest value of True negative (0.53) while also having the lowest value of False negative (0.012) and False positive (0.018). Models like AdaBoost, Decision Tree and Logistic Regression on the other hand has fared low in classification, having marginally low True positive values, but have a good True negative value. Though decision Tree has a higher value of True positive than the Random Forest model, it still has better values of other parameters.

The Recurrent neural networks have also done reasonably well in classifying classes as they are just below random forest in terms of confusion matrix parameter scores.

- Model Classification Table

	Train_Score	Test_Score	Precision_Score	Recall_Score	F1_Score	accuracy
Random Forest Tree	1.000000	0.971097	0.979514	0.967098	0.973266	0.971097
BI-LSTM	0.990023	0.947330	0.964494	0.937702	0.950909	0.947330
LSTM	0.994645	0.941755	0.976129	0.915318	0.944746	0.941755
GRU	0.970435	0.933979	0.946057	0.931769	0.938859	0.933979
Support Vector Machine	0.940760	0.913146	0.938626	0.899137	0.918457	0.913146
Decision Tree	1.000000	0.907424	0.981534	0.845739	0.908590	0.907424
logisitc regression	0.918018	0.893633	0.908071	0.895092	0.901535	0.893633
AdaBoost	0.901181	0.870745	0.876632	0.887271	0.881919	0.870745

Table 1- Classification scores of all the modelling techniques

Based on the above scores, we can see that Random Forest Tree has just marginally edged out other models. The accuracy and the test data score (0.9710) are much higher than that of the remaining models. The Recurrent neural networks (LSTM, GRU, BI-LSTM) are second best in the list and have done reasonably well in classifying the data. The AdaBoost ml classifier is placed last as it has the lowest scores compared to the rest.

- **AUC-ROC curve**

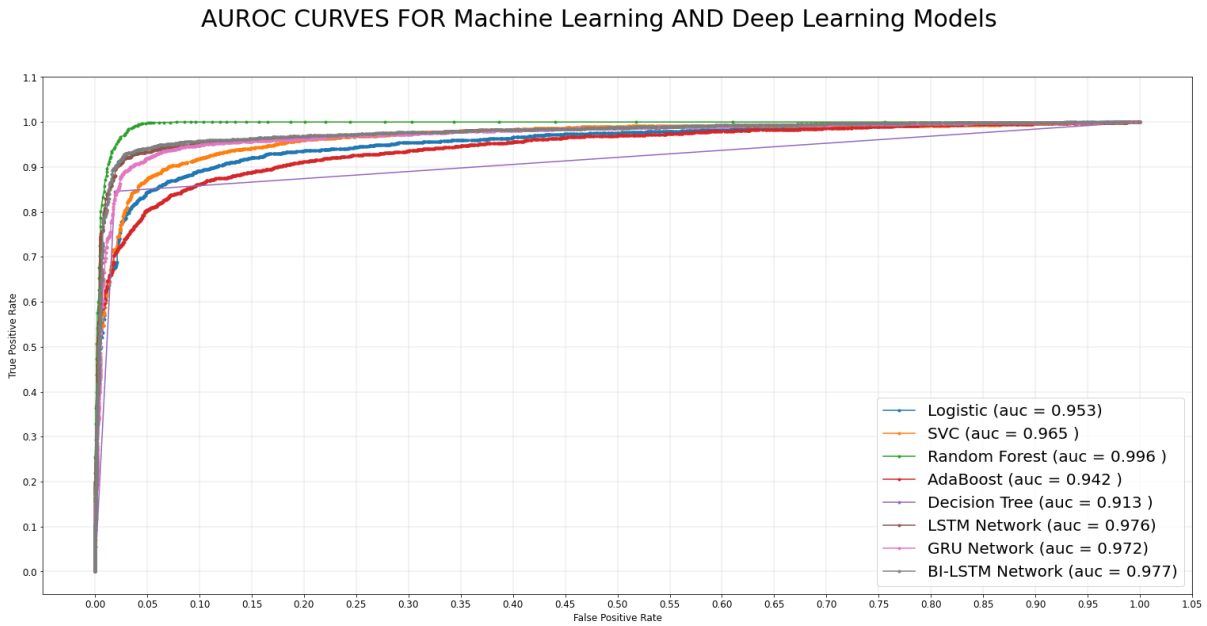


Figure 15- AUROC curve for all the modelling techniques

The AUC-ROC graph is the best graph for measuring the performance of any model. In the graph above. Random Forest classifier has once again achieved the best performance as it has the highest AUC value of 99.6% and it's ROC curve is higher than that of other models. LSTM and BI-LSTM are very close in terms of AUC values (LSTM AUC= 0.976, BI-LSTM AUC = 0.977) and their ROC curves are almost on a similar trajectory. The GRU model sees a slight dip with a AUC value of 0.978 and similar RUC curve trajectory. The 3 bottom ROC curves are produced by the remaining ML models (SVC, Logistic Regression, AdaBoost and Decision Tree) which fared worse than the RNN models with low AUC scores and no upward incline of ROC curves.

6. CONCLUSION AND FUTURE WORKS

In this research paper, we predict sentiment of customer reviews using 5 machine learning models and 3 deep learning models. For the ML models, we use train them with help of a TF-IDF vectorizer. The experiments we carried out, helped us study the performance of all the models based on their accuracy while predicting testing data, comparing their confusion matrix and their AU-ROC curves.

Among all the models, The Random Forest Classifier with TF-IDF vectorizer seemed to predict sentiments with the highest accuracy and better AU-ROC curve than the rest. Based on overall performance, it can be concluded that the Deep learning models have a better overall score than the machine learning models.

For future works, we could consider the option of adding emojis to the dataset, which could help the models predict sentiments better. Reviews may contain many stop words that may remain even after processing, which can be classified as positive sentiment, even though the overall sentiment is a negative one. We can also look to further fine tune the hyperparameters of the deep learning networks to perform better, predicting sentiments.

7. Appendix

Link to my github repository for EDA- https://github.com/Nikhil2103/MRP-Womens-clothes-review-analysis

8. REFERENCES

[1] Shahid, Ms & Nadeem, Kashif & Hafeez, Shahid. (2019). Impact of Online Consumer Reviews on Hotel Booking Intentions: The Case of Pakistan. European Scientific Journal. 15. 1857-7881. 10.19044/esj.2019.v15n7p144.

[2] Willianto, Tommy, and A. Wibowo. "Sentiment analysis on E-commerce product using machine learning and combination of TF-IDF and backward elimination." International Journal of Electrical and Computer Engineering (IJECE) 8.6 (2020): 2862-2867.

[3] Sudhakaran, P., & Jaiganesh, M. (2020). Sentiment Analysis Based Product Selection for Enhancing E-Commerce.

[4] Awale, Rohan & Mane, Vijay & pisal, vipul & Patil, Sanmit. (2021). An Efficient Sentiment Analysis Based on Product Reviews. 10.18090/samriddhi.v13spli02.17.

[5] Chintalapudi, N.; Battineni, G.; Amenta, F. Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models. Infect. Dis. Rep. 2021, 13, 329-339. https://doi.org/10.3390/idr13020032

[6] Rita Rahayu, John Day, Determinant Factors of E-commerce Adoption by SMEs in Developing Country: Evidence from Indonesia, Procedia - Social and Behavioral Sciences, Volume 195, 2015, Pages 142-150, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2015.06.423.

[7] Giachanou, Anastasia & Crestani, Fabio. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Computing Surveys. 49. 1-41. 10.1145/2938640

[8] Mohamed Ali, Nehal et al. “SENTIMENT ANALYSIS FOR MOVIES REVIEWS DATASET USING DEEP LEARNING MODELS.” International Journal of Data Mining & Knowledge Management Process (2019): n.pag.

[9] Marong, Muhammad & Batcha, Nowshath & Raheem, Mafas. (2020). Sentiment Analysis in E-Commerce: A Review on The Techniques and Algorithms. 6.

[10] Desai, Mitali & Mehta, Mayuri. (2016). Techniques for sentiment analysis of Twitter data: A comprehensive survey. 149-154. 10.1109/CCAA.2016.7813707.

[11] Sattam Almatarneh, Pablo Gamallo, 2018. A lexicon based method to search for extreme opinions https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0197816

[12] Taboada, Maite & Brooke, Julian & Tofiloski, Milan & Voll, Kimberly & Stede, Manfred. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics. 37. 267-307. 10.1162/COLI_a_00049.

[13] Bandhakavi, Anil et al. “Lexicon based feature extraction for emotion text classification.” *Pattern Recognit. Lett.* 93 (2017): 133-142.

[14] Asghar, Dr. Muhammad & Khan, Aurangzeb & Ahmad, Shakeel & Qasim, Maria & Khan, Imran. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. PLoS ONE. e0171649. 1-22. 10.1371/journal.pone.0171649.

[15] Das, Sanjiv Ranjan, and Mike Y. Chen. "Yahoo! for Amazon: Sentiment parsing from small talk on the web." For Amazon: Sentiment Parsing from Small Talk on the Web (August 5, 2001). EFA (2001)

[16] Bhaskar, Nuthanakanti & Pudi, Sandeep & Amarnath, Nethi & Thankachan, Joel. (2021). A STUDY: SENTIMENTAL ANALYSIS FOR ELECTION RESULTS BY USING TWITTER DATA. 33. 6804-6812.

[17] Brar, G. S. and Ankit Sharma. “Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques.” (2018).

[18] Ashish Kumar Singh , Sonal Singh, 2022, Exploratory and Sentiment Analysis – Netflix, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 01 (January 2022)

[19] Vadloori, Karthik & Sanghishetty, Shriya. (2021). Exploratory and Sentiment Analysis of Netflix Data.

[20] Gupta, Bhumika & Negi, Monika & Vishwakarma, Kanika & Rawat, Goldi & Badhani, Priyanka. (2017). Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python. International Journal of Computer Applications. 165. 29-34. 10.5120/ijca2017914022.

[21] https://towardsdatascience.com/5-techniques-to-work-with-imbalanced-data-in-machine-learning-80836d45d30c

[22] https://analyticsindiamag.com/why-data-scaling-is-important-in-machine-learning-how-to-effectively-do-it/

[23] Library containing all the regex functions https://docs.python.org/3/library/re.html

[24] A documentation for TF-IDF vectorizer https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[25] Documentation for the dropna() method for removing rows with null values https://www.w3schools.com/python/pandas/ref_df_dropna.asp#:~:text=The%20dropna()%20method%20removes,in%20the%20original%20DataFrame%20instead.

[26] Benefits of using random and stratified sampling https://stats.stackexchange.com/questions/250273/benefits-of-stratified-vs-random-sampling-for-generating-training-data-in-classi

[27] Documentation for confusion matrix https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[28] A paper on How to compare the AU-ROC curves of models https://stats.stackexchange.com/questions/304478/comparison-of-two-models-when-the-roc-curves-cross-each-other#:~:text=One%20common%20measure%20used%20to,model%20with%20a%20smaller%20AUC.

[29] A documentation on how to understand the classification report and its attributes https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

[30] Introduction to lemmatization https://en.wikipedia.org/wiki/Lemmatisation