International Journal of Computer Science and Mathematical Theory (IJCSMT) E-ISSN 2545-5699
P-ISSN 2695-1924 Vol 11. No.2 2025 www.iiardjournals.org Online Version

A Comparative Study of Hybrid Recommendation Systems for E-commerce Based on Sentiment Analysis and Star Ratings using the weighted Hybrid Approach

Amaechi, Ikenna Victor 1[1], Naveed Anwar 2[1] Honglei, Li 3[1]
[1]Department of Computer and Information Sciences Northumbria University,
Newcastle, United Kingdom

DOI: 10.56201/ijcsmt.v11.no2.2025.pg1.25

Abstract

In recent times, the use of machine learning and deep learning models in sentiment analysis and star rating prediction has been a focal point of several studies. This research embarked on an in-depth examination and comparison of traditional and hybrid models in sentiment analysis and star rating prediction. The traditional models showcased strong capabilities with Random Forest illustrating high accuracy and exceptional classification abilities in sentiment analysis, and Logistic Regression surfacing as a dominant contender in rating predictions. In contrast, hybrid models, characterized by heightened flexibility and tuning capabilities, manifested optimal performance within specific sentiment weight ranges, thus indicating a promising approach to enhancing prediction accuracy and reliability. The project also highlighted significant learning and opportunities for innovation, despite encountering limitations such as computational resource constraints and dataset imbalances. The research implies that a careful amalgamation of various model strengths and feature combinations could potentially pave the way for achieving peak performance in sentiment analysis and rating predictions. Moving forward, future research could delve into exploring new hybrid combinations and further optimization of model parameters to enhance performance. This study sets a promising precedent for the development of more advanced and efficient predictive analytics solutions, thus fostering growth and innovation in the fields of machine learning and data science.

Keywords: *Sentiment Analysis, Deep Learning, Star Rating Prediction, Traditional Models, Hybrid Models, Random Forest, Logistic regression, Predictive Analysis, Innovation in Machine learning*

1. Introduction

In recent years, the rapid expansion of e-commerce has significantly transformed consumer shopping experiences. With an overwhelming number of products available online, recommendation systems have become essential tools for enhancing customer engagement, improving sales, and fostering brand loyalty (Wang et al., 2020; Suresh & MJ, 2020). These systems assist customers in navigating vast product catalogs by providing personalized suggestions based on their preferences and behaviors (Li et al., 2022; Kulkarni & Rodd, 2020).

Traditional recommendation approaches, such as collaborative filtering (CF) and content-based filtering (CBF), have been widely employed to suggest products based on user preferences. However, these methods often face limitations, such as the cold start problem, where new users and items lack sufficient interaction history to generate reliable recommendations (Elahi et al., 2023; Channarong et al., 2022). Additionally, sparsity issues in collaborative filtering arise when there are insufficient interactions between users and items, making it difficult to produce meaningful recommendations (Yi & Liu, 2020; García-Sánchez et al., 2020). Conversely, content-based filtering, which recommends products based on item attributes, often leads to recommendation homogeneity, where users are continuously shown similar products, reducing diversity in suggestions (Javed et al., 2021; Osman, Noah, & Darwich, 2019).

To address these challenges, hybrid recommendation systems have emerged as a viable alternative by integrating multiple recommendation techniques to improve accuracy, diversity, and personalization (Asthana, 2022; Walek & Fajmon, 2023). These models leverage both collaborative filtering and content-based filtering, offering better performance in mitigating the cold start problem, reducing sparsity, and increasing recommendation relevance (Kulkarni & Rodd, 2020; García-Sánchez et al., 2020). Recent research has demonstrated that hybrid models outperform single-method recommendation systems by providing a broader variety of recommendations and enhanced user satisfaction (Roy & Dutta, 2022; Logesh et al., 2019).

Among various hybrid approaches, the weighted hybrid model, which combines sentiment analysis with star rating systems, has gained significant attention. This model integrates both explicit numerical feedback (star ratings) and implicit textual sentiment (user reviews) to generate more reliable and personalized recommendations (Dang et al., 2021; Noor, Bakhtyar, & Baber, 2019). Sentiment analysis helps capture customer opinions beyond numerical ratings, addressing the discrepancies that often arise when users leave ratings that do not fully align with their textual reviews (Alqaryouti et al., 2020; Jabbar et al., 2019).

Studies indicate that star ratings alone may not fully reflect customer satisfaction, as they provide only a generalized evaluation of a product, while sentiment analysis uncovers deeper patterns in user emotions (Yang et al., 2020; Alamoudi & Alghamdi, 2021). For instance, some customers may leave a high star rating but express dissatisfaction in their review text, making sentiment analysis crucial for accurate recommendation generation (Handhika et al., 2019). By combining both sentiment polarity scores and star ratings, weighted hybrid models can enhance recommendation accuracy, capture user sentiment more effectively, and provide a more holistic understanding of consumer preferences (Shrestha & Nasoz, 2019; Rajeswari et al., 2020).

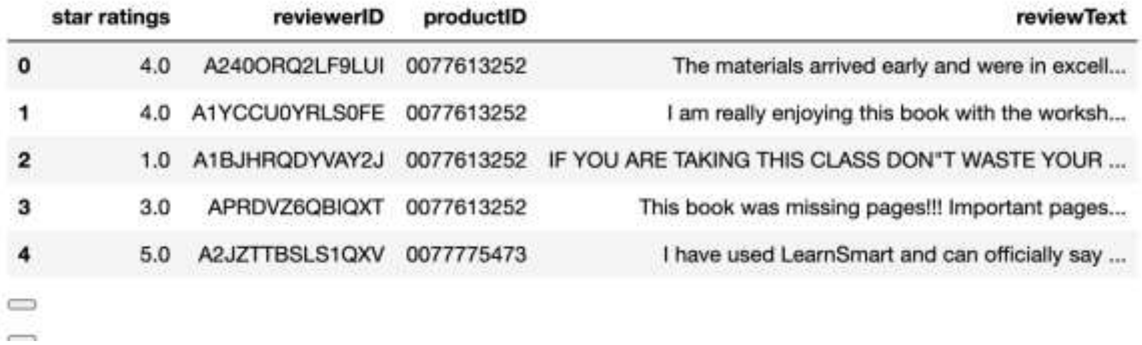
Business companies are facing increasing competition in the market for client acquisition because of the rapid deployment of technology innovation in e-commerce platforms. This increased level of competition is a direct result of the growth of e-commerce. It is necessary to reach greater levels of performance in hybrid recommendation systems by making efficient use of marketing strategy to aid consumers in making snap selections on the things they wish to buy. This will help customers make purchases more quickly. Systems that award stars also have a good relationship with sentiment analysis, which helps users build confidence in one another and the platform. This is because users are rewarded for positive comments rather than negative ones. In this instance, awarding five stars indicates an amazing customer experience (CX), but awarding lesser grades of stars indicates inadequate levels of consumer satisfaction with the product. Before making any sort of purchase, customers in today's world perform comprehensive research on previously published reviews. This is one of the most essential variables that decide the development of sales for an e-commerce company. As a direct result of this, the selection of components is of the highest significance when it comes to the classification of feelings. This, in turn, allows forms to describe the behaviour and attitudes of consumers when they are providing evaluations. In the current climate, it is more beneficial to adopt a customized recommendation system to simplify operation management and boost customer experiences (CX). This is because these two goals are directly related to each other.

2 Methodology

This chapter outlines the methodology underpinning the development of a hybrid recommendation system tailored to suggest products by discerning customer preferences and interactions. Central to this research is the fusion of sentiment polarity scores and star ratings, integrated in a weighted fashion to cultivate a hybrid model, enhancing prediction accuracy. In pursuit of a comprehensive grasp on customer behaviour, this research meticulously analyses their reviews using the TF-IDF method. Leveraging contemporary software tools and a uniform set of evaluation metrics, the methodology ensures transparency and reliability in its findings.

2.1 Data Preparation

Given the computational needs of deep learning models, especially the BERT model (Mutinda et al., 2023) used in this research, the initial steps were performed on Google Colab, a cloud-based platform that offers GPU runtime capabilities. Utilizing the Python pandas library, the data was imported from a CSV file stored on Google Drive. The dataset presents an array of attributes, with the key columns being 'overall' (representing star ratings), 'reviewerID', 'asin' (product ID), and 'reviewText' (the textual content of the review). A snapshot inspection of the dataset was carried out to gain an initial understanding of its structure and the nature of the stored data.



	star ratings	reviewerID	productID	reviewText
0	4.0	A240ORQ2LF9LUI	0077613252	The materials arrived early and were in excell...
1	4.0	A1YCCU0YRLS0FE	0077613252	I am really enjoying this book with the worksh...
2	1.0	A1BJHRQDYVAY2J	0077613252	IF YOU ARE TAKING THIS CLASS DON"T WASTE YOUR ...
3	3.0	APRDVZ6QBIQXT	0077613252	This book was missing pages!!! Important pages...
4	5.0	A2JZTTBSLS1QXV	0077775473	I have used LearnSmart and can officially say ...

Figure 2.1: Selected unprocessed dataframe.

2.1.1 Data Cleaning and Preprocessing

As part of an efforts to uphold these standards, column names were realigned for clarity. Specifically, 'overall' was renamed to 'star ratings' and 'asin' was changed to 'productID' to enhance comprehensibility. Furthermore, to concentrate the data around the research's requirements, superfluous columns were removed, centering the dataset on key elements such as 'star ratings', 'reviewerID', 'productID', and 'reviewText'. Lastly, for the sake of ensuring consistent analysis, any rows containing null entries in the 'reviewText' and 'star ratings' columns were diligently excluded.

Text preprocessing plays a crucial role in enhancing the quality and consistency of data. To achieve this, several measures were implemented on the review texts. Firstly, the entirety of the review texts was converted to lowercase, ensuring uniformity, and eliminating potential discrepancies arising from variant text cases(Jin et al., n.d.). Subsequently, any non-alphabetic characters were purged from the text, which allows for more precise tokenization and sentiment analysis. The tokenization process then came into play, where the text was systematically divided into individual words or tokens, a fundamental procedure in the Natural Language Processing (NLP) continuum as it segments the text into analyzable fragments. Once tokenized, lemmatization was executed on each word, converting them to their base or root forms. This action ensures that diverse word forms with an identical root meaning, exemplified by words like 'running', 'runs', and 'ran', are uniformly recognized as 'run'. During this step, a specialized function, named get_wordnet_pos, was employed to map the part-of-speech tags from NLTK's pos_tag function to those identified by WordNet, thereby refining the lemmatization process. As a result of these combined measures, the reviewText column in the dataframe underwent comprehensive preprocessing, priming it for subsequent feature extraction. Lastly, the Term Frequency-Inverse Document Frequency (TF-IDF) technique was utilized to metamorphose the preprocessed text into a matrix of features. This method assigns values to words based on their significance in a document in relation to their prevalence in the broader corpus. To optimize this, words with common occurrences across documents, often labeled as "stop words" due to their lack of distinctive insights, were intentionally left out.

2.2 Exploratory Data Analysis

During the exploratory analysis phase, descriptive statistics were leveraged to succinctly represent the central tendency, dispersion, and shape of both the 'star ratings' and subsequently introduced 'sentiment scores' distributions. Histograms provided a visual representation of these distributions, making it straightforward to identify any patterns or anomalies. Additionally, to ensure the integrity and accuracy of the analysis, outlier detection was meticulously executed on the 'star ratings' and 'sentiment' columns. Using specific criteria, rows that fell outside the expected range (1 to 5) for both 'star ratings' and 'sentiment' were isolated. However, after thorough examination, it was determined that neither 'star ratings' nor 'sentiment' columns had any outliers, ensuring the robustness of the dataset. Building on these insights, a correlation heatmap was crafted using Pearson correlation coefficients. This heatmap served as a powerful tool to visually depict and numerically quantify the degree of linear relationship between the star ratings and the calculated sentiment scores. Finally, a chi-square statistical test was performed to determine the strength of correlation by first discretizing both the sentiment and star ratings into categories.

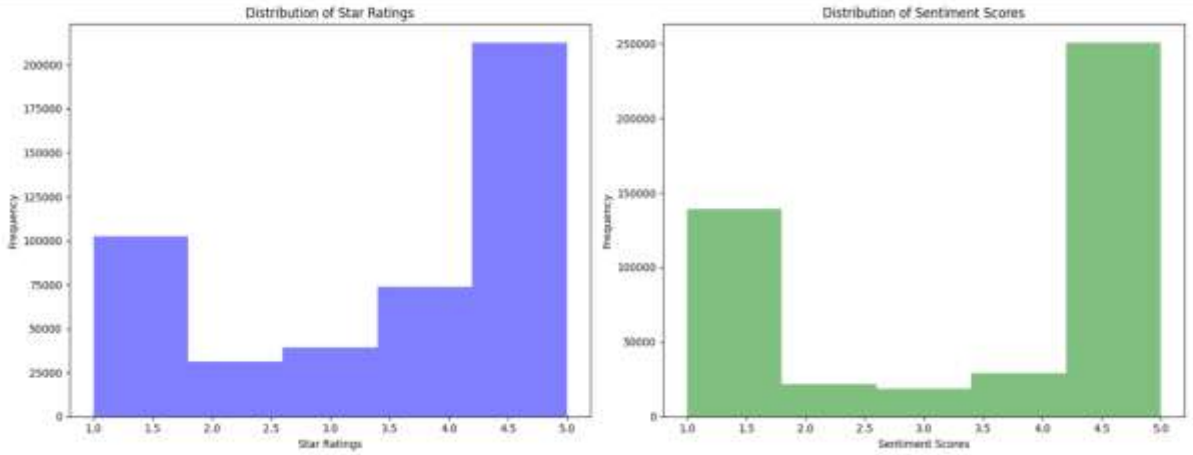


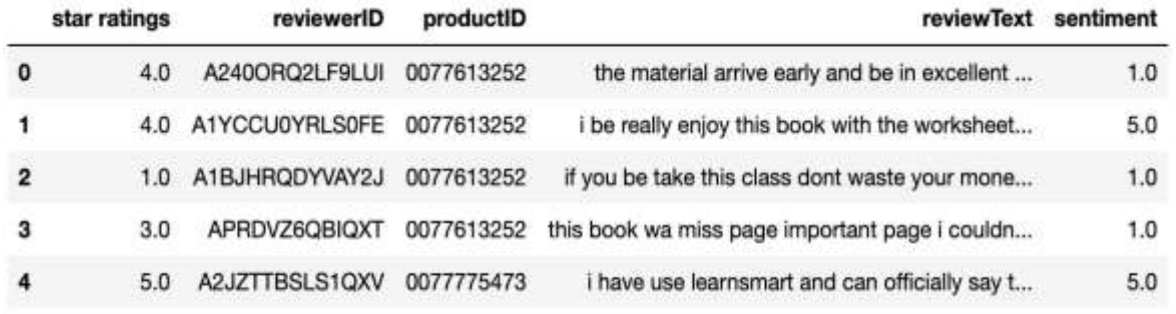
Figure 2.2: Star Rating and Sentiment Polarity score distribution.

2.3 Sentiment Analysis Approach

The sentiment analysis was primarily driven by a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model which have been previously used by other researchers (Hao et al., 2023; Lengkeek et al., 2023) due to its advantages, including its benchmark-setting performance, its ability to grasp contextual word subtleties because of its bidirectional nature, and its transfer learning advantages from pre-training on large text corpora. Specifically, a fine-tuned variant of the BERT model intended for sentiment classification on the IMDB dataset was employed. Sourced from the transformers library, this model stands at the forefront of natural language processing, renowned for its prowess in sentiment extraction from textual reviews. For the sentiment analysis, each review undergoes a series of transformations. Firstly, the review text is tokenized and formatted to be compatible with the BERT model. This involves segmenting the text into chunks or tokens and encoding them into a format that the BERT model can understand.

Given the vastness of certain reviews, they might exceed BERT's maximum input length, thus necessitating truncation to fit within the limit. To maximize efficiency and reduce computational time, the model's operations were executed on the GPU runtime provided by Google Colab, a platform that allows for robust cloud-based computations. Once the input is prepared, it's transferred to the GPU for fast processing. Utilizing the GPU over a traditional CPU ensures that the vast amount of matrix operations inherent in deep learning models like BERT are executed swiftly, leading to quicker sentiment score derivations for each review. The model then processes these tokenized reviews and provides an output, which consists of logits or raw prediction scores for each sentiment category. These logits are then passed through a softmax function to transform them into probabilities. The sentiment score for each review is computed by taking the difference between the probabilities of the positive and negative classes. This sentiment score essentially captures the polarity of the sentiment, with positive values indicating a positive sentiment and negative values indicating a negative sentiment.

Given that the original star ratings lie within a range of 1 to 5, it's crucial to align the sentiment scores with this scale. To achieve this alignment, the MinMaxScaler from the sklearn library was employed. This scaler transformed the sentiment scores to fit within the range of 1 to 5, mirroring the original star ratings' range. However, the nature of star ratings is inherently discrete, meaning they often exist as whole numbers. To emulate this and facilitate straightforward comparisons, the scaled sentiment scores were then rounded off to the nearest whole number. By the end of this process, each review in the dataset had a corresponding sentiment score, offering a machine-driven perspective on its sentiment, tailored to mirror the original star ratings.



	star ratings	reviewerID	productID	reviewText	sentiment
0	4.0	A240ORQ2LF9LUI	0077613252	the material arrive early and be in excellent ...	1.0
1	4.0	A1YCCU0YRLS0FE	0077613252	i be really enjoy this book with the worksheet...	5.0
2	1.0	A1BJHRQDYVAY2J	0077613252	if you be take this class dont waste your mone...	1.0
3	3.0	APRDVZ6QBIQXT	0077613252	this book wa miss page important page i couldn...	1.0
4	5.0	A2JZTTBSLS1QXV	0077775473	i have use learnsmart and can officially say t...	5.0

Figure 3.2: DataFrame after pre-processing and sentiment polarity scores rescaling.

2.3.1 Customer behaviour Analysis

To better understand customer behavior, reviews were delved into to extract key insights. Using the TF-IDF method, which measures a word's importance in a document relative to a collection of documents (Gomes et al., 2023), the reviews were converted into an organized DataFrame which allowed for easier analysis. Reviews were then categorized by their star ratings: those with a rating of 4 or higher were considered positive, highlighting what customers liked, while those with a rating of 2 or lower pointed out areas for improvement.

This classification made it possible to gather TF-IDF scores specific to each sentiment. The core of this analysis identified the top 10 recurring words in both positive and negative reviews. Bar charts vividly displayed the main topics in both positive and negative feedback. Additionally, word clouds were used for each sentiment category. These clouds emphasized the most important words visually: the more significant the word, the larger its display. Words frequently linked with customer satisfaction stood out in the positive cloud, while the negative cloud highlighted common customer concerns. These visuals offer a quick yet deep dive into customer opinions, aiding in informed future decisions.

2.4 Model Development

A subset of 70,000 samples was chosen from the main dataset. This decision was made to manage computational limitations, ensuring that the model development went smoothly without system hiccups. The data was partitioned into an 80-20 split, with 80% allocated for training to allow the model to adequately learn and recognize patterns, and 20% reserved for testing to verify the model's predictive accuracy and prevent overfitting, thereby ensuring a well-rounded evaluation of its performance (Najafabadi et al., 2015). To make sure these steps would not be repeated in the future, the processed data and the training-test sets were saved as CSV files and uploaded to one drive. The Random Forest model was chosen for its renowned accuracy and classification abilities, especially in discerning nuanced sentiment categories (Al Amrani et al., 2018). Logistic Regression was selected due to its unparalleled accuracy and balanced classification outcome in star rating predictions, a critical feature in achieving a holistic understanding of user sentiments (Shakhovska et al., 2020). Additionally, the Recurrent Neural Network (RNN) was chosen for this project because of its proficiency in processing sequential data, making it exceptionally suited for analyzing patterns and sequences in text data - a vital component in sentiment analysis and star rating prediction tasks (An et al., 2018).

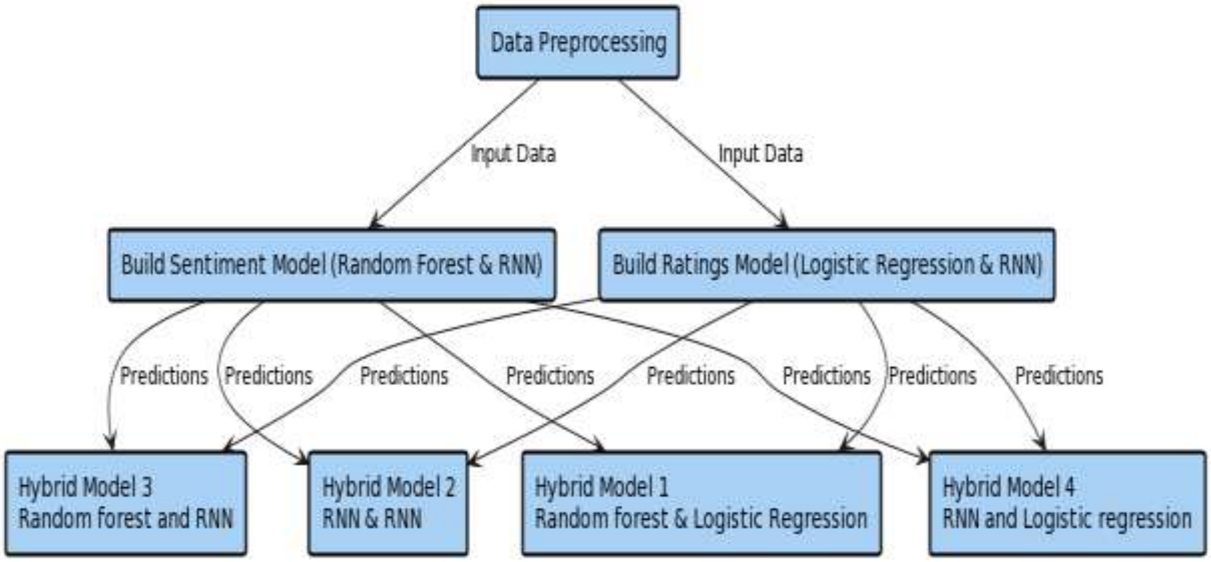


Figure 3.3: Hybrid recommendation system machine learning model pipeline

2.4.1 Random Forest for Sentiment Based model.

For the sentiment model, a Random Forest classifier was employed. The ratings were binarized using a threshold of 4, marking ratings of 4 and above as positive (encoded as 1) and the rest as negative (encoded as 0). The reviews underwent vectorization through the TF-IDF vectorizer, focusing on the top 5,000 terms by frequency. To ensure unbiased learning, especially in scenarios with significant class imbalances, the Random Forest was set with balanced class weights. Post-training, the model furnished both class probabilities and straightforward class predictions for the test dataset.

2.4.2 Recurrent Neural Network for Sentiment Based model.

In the development of sentiment models using RNN, TensorFlow played a pivotal role. Reviews were tokenized, focusing on the most frequent 5,000 words from the training data, and sequences were standardized to a consistent length of 200. Using the earlier stipulated threshold, star ratings underwent a transformation into a binary format. The RNN's structural blueprint encompassed an embedding layer set to a 200-length input and 64 dimensions. This was followed by a 64-unit RNN layer with L2 regularization and sequence returns. To mitigate the risk of overfitting, a dropout layer with a 0.5 rate was integrated. Subsequent layers included a second RNN layer with 32 units and L2 regularization, culminating in a dense output layer utilizing a sigmoid activation function tailored for binary classification. Training was fine-tuned using the Adam optimizer, configured with a 0.001 learning rate, and to further fortify against overfitting, early stopping was implemented with a patience threshold of three.

2.4.3 Logistic Regression for Star Rating Based Model

For the star rating model, a Logistic Regression approach was adopted. Initially, star ratings were transformed into binary format using a threshold of 4. This was followed by the application of TF-IDF vectorization to the review texts, emphasizing the top 5,000 words. With this preprocessed data, the logistic regression model was trained, setting the maximum iterations at 1000 to ensure optimal convergence. Upon completion, the model's performance was gauged, and subsequent predictions were executed on the test set.

2.4.4 Recurrent Neural Network for Star Rating Based Model

For the star rating prediction, an RNN model was brought into play, taking cues from the RNN framework previously employed for sentiment analysis. Star ratings underwent a binary conversion. Subsequently, the reviews were tokenized, spotlighting the top 5,000 words, and then tailored to a consistent length of 50 through padding. Structurally, the RNN was designed with an embedding layer featuring 50 dimensions. This segued into a simple RNN layer, equipped with 512 units, and endowed with L2 regularization on kernel, recurrent connections, and bias. A dropout layer, set at a rate of 0.5, was introduced to counter overfitting. The architecture was completed with a dense output layer, employing a sigmoid activation function for binary classification. Training of the model hinged on the RMSProp optimizer, with the added measure of early stopping to preempt overfitting.

2.4.5 Hybrid model

Some hybrid models, which combined many distinct models, were created in the search for a better prediction model. The key to creating hybrid models is using one model's advantages to partially compensate for another's possible weaknesses, aiming for a balanced and effective prediction process. This fusion intended to improve performance on unexplored data by leveraging the distinct predicting abilities of each algorithm, rather than merely combining them. A weighted sum of the probability scores from two models served as the foundation of the hybrid strategy. The weighted score of a particular data point can be expressed as follows:

$$Weighted\ score = w\ \times\ Probability\ Model\ A\ + (1 - w) \times Probability\ Model\ B$$

where 'w' signifies the weight factor, oscillating between 0 and 1. To pin down the quintessential weight for the best synergy, we embarked on an exhaustive search across varied weight magnitudes.

Model 1: Hybrid of Random Forest (Sentiment) & Logistic Regression (Ratings)

In Model 1, the sentiment predictions from a Random Forest model are merged with ratings predictions from a Logistic Regression model. The idea is that a tree-based model like Random Forest might capture nonlinear interactions in the sentiment data, whereas the Logistic Regression model is likely capturing linear relationships in the ratings data. The hybrid scores are obtained by taking a weighted average of the predicted probabilities from both models. By iterating over different weights, we can find the optimal balance between the two models' predictions to maximize the ROC-AUC.

Model 2: Hybrid of RNN (Sentiment model) & RNN (Ratings model)

Model 2 combines the power of two RNNs - one trained on sentiment and another on ratings. RNNs, being sequential models, are particularly suited for tasks where order matters, like time series data or text. In this setup, both the sentiment and ratings might have sequence-based patterns that the RNNs can capture. By adjusting the weightings of these two RNN outputs, the hybrid model tries to find the best combination of sentiment-based sequential patterns with rating-based sequential patterns.

Model 3: Hybrid of Sentiment model (Random Forest) and Ratings model (RNN)

Model 3 is an interesting mix of a tree-based model (Random Forest) for sentiment and a sequential model (RNN) for ratings. The idea here is to combine the non-linear pattern capturing ability of Random Forests with the sequential pattern capturing ability of RNNs. This can be particularly powerful if, for instance, sentiment has complex interactions between features, and ratings have sequential dependencies.

Model 4: Hybrid of Sentiment model (RNN) and Ratings model (Logistic Regression)

In Model 4, the RNN captures sequential patterns in sentiment data while the Logistic Regression captures linear patterns in the ratings data. This hybrid model believes in the strength of RNNs in capturing sequences in sentiments and the straightforward relationships in ratings which a logistic regression can efficiently tackle. For each of these models, different weights are tried out for the outputs of the two underlying models to find the combination that gives the highest ROC-AUC. This optimization helps in finding the best balance between the two models being combined.

2.5 Evaluation Metrics

In this research, a rigorous evaluation process using standard metrics to gauge the efficacy of each model was employed. By setting a rating threshold of 4, with reviews scoring 4 or above classified as 'Recommended'. This distinction clarified the underlying sentiment of each review. Precision gauged the proportion of accurately labelled 'Recommended' reviews, showcasing our model's reliability. Recall, meanwhile, reflected the model's skill in capturing genuine 'Recommended' reviews. The F1 Score, harmonizing Precision and Recall, became especially vital in cases of uneven datasets, with a score near 1 indicating model excellence. Accuracy offered a snapshot of the model's overall capability in identifying correct reviews, while the ROC Curve and AUC provided a visual and holistic assessment of the model's performance, respectively. With an AUC close to 1 denoting superior predictive prowess. By leveraging these metrics, the research elucidates each model's performance nuances, enhancing the credibility of our primary conclusions.

2.6 Software and Tools

The computational rigor and complexity of the tasks undertaken in this dissertation necessitated the use of an array of cutting-edge software, tools, and computational resources. This section provides a comprehensive overview of the software and tools employed, along with a brief rationale for each choice.

2.6.1 Development Environment and Computational resources

The primary tool for code development, execution, and documentation was the Jupyter Notebook. This interactive computing environment was selected for its seamless blend of interactivity, visualization, and comprehensive documentation. For tasks necessitating GPU capabilities, especially during the deep learning model training phase, Google Colab was employed. All computations pertinent to this dissertation were carried out on a MacBook Pro equipped with the M1 chip, a recent innovation from Apple, known for its speed and efficiency, thus making it apt for the research's data processing requirements.

2.6.2 Data Storage and Backups

All research materials, encompassing datasets, Word documents, Python notebooks, and other relevant documentation, were meticulously stored on the OneDrive platform provided by Northumbria University. Utilizing this platform conferred dual advantages: it ensured data integrity with seamless access, and it offered a layer of redundancy. This systematic approach safeguarded against potential information or code loss throughout the research process.

2.6.3 Programming, Data Processing, and Visualization

During the research, a wide range of Python libraries were harnessed to accomplish a variety of tasks. The Natural Language Toolkit (NLTK) played a central role in text processing, aiding in tokenization, lemmatization, and the removal of stopwords. For adept data manipulation and thorough analysis, Pandas stood out as an invaluable resource. Scikit-learn, rich in its assortment of machine learning algorithms and tools, catered to various needs, from text vectorization to the training and evaluation of models. Deep learning models, with an emphasis on LSTM models, were deftly crafted using TensorFlow's Keras API. For visual insight, Matplotlib was indispensable, aiding in the creation of intricate charts and plots, notably the ROC curve. Additionally, to offer a detailed comparison of the models and their performance metrics in a visually engaging manner, the interactive visualization capabilities of Power BI were employed.

2.7 Ethical Issues

Within academic research, a variety of ethical concerns emerge due to the varied roles of individuals within these institutions (Drolet et al., 2023). In the methodology chapter of this research, ethical considerations stand at the forefront, emphasizing the responsibility and integrity of the research process. One of the primary areas of focus was data privacy and protection. Given the sensitive nature of using real customer reviews or data, the onus was on ensuring the absolute privacy and anonymity of the contributors. This was meticulously achieved by anonymizing all customer and product identifiers. Each data entry was coded uniquely, eliminating any possibility of retracing to the original contributor or product. Moreover, strict adherence to terms of service and privacy policies was maintained whenever third-party data platforms were utilized, reinforcing a staunch commitment to data privacy.

Environmental sustainability was another pivotal concern. The computationally intensive nature of deep learning models inherently poses a risk of amplifying the carbon footprint due to the immense computational power they necessitate (Heguerte et al., 2023). In cognizance of this, the research was strategized to minimize its environmental impact. By limiting the scope to a subset of the dataset, the computational demand was considerably curtailed. Additionally, to further optimize computational efficiency and reduce potential environmental repercussions, all computationally rigorous deep learning models were run using Google Colab's GPU, ensuring not only swifter processing but also a judicious use of resources.

Lastly, the potential for biases in machine learning models, especially when predicated on skewed datasets, is a matter of significant concern. Such biases can inadvertently perpetuate societal prejudices, leading to potential misrepresentation or discrimination against certain groups or opinions (Pagano et al., 2023). To navigate this, rigorous measures were employed. The dataset chosen was a random subset, making it a balanced and unbiased representation. Further attestation to the unbiased nature of the study came from the evaluation metrics, which demonstrated the model's balanced and non-discriminatory treatment across diverse data points. Potential challenges were not just recognized, but proactive measures were implemented to address them. This approach ensured that the research upheld the highest standards of credibility and responsibility.

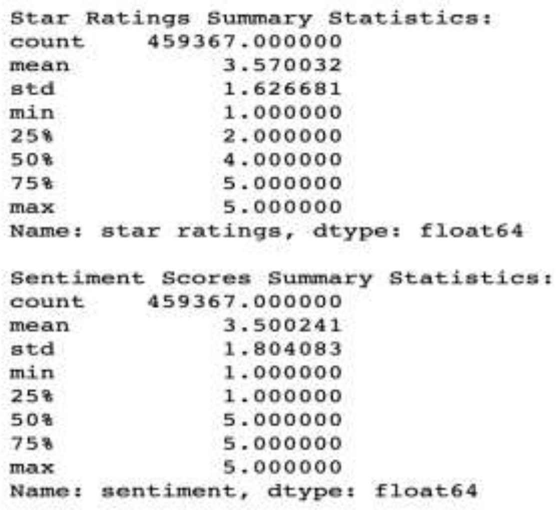
3 Result and Findings

3.1 Introduction

An in-depth analysis of customer evaluations, star ratings, and sentiment scores is performed in this crucial section of the study. Using a large dataset of software reviews, the research reveals the intricacies of customer sentiments and their effects on star ratings. Through keyword analysis, reoccurring topics in customer evaluations are identified, and the performance metrics of various analytical and predictive models are evaluated. This method is intended to promote a deeper understanding of consumer behaviours and the factors influencing their contentment or discontent with the items or services under consideration.

3.2 Summary statistics of Ratings and sentiment score

In this study, both the ratings and sentiment scores were statistically analysed. The results are as follows:



```
Star Ratings Summary Statistics:
count    459367.000000
mean          3.570032
std           1.626681
min           1.000000
25%           2.000000
50%           4.000000
75%           5.000000
max           5.000000
Name: star ratings, dtype: float64

Sentiment Scores Summary Statistics:
count    459367.000000
mean          3.500241
std           1.804083
min           1.000000
25%           1.000000
50%           5.000000
75%           5.000000
max           5.000000
Name: sentiment, dtype: float64
```

Figure 3.1: Summary statistics of Star Ratings and Sentiment polarity scores.

Star Ratings Summary Statistics: The dataset comprises a substantial count of 459,367 reviews, providing a robust foundation for any analytical interpretations. Such a voluminous sample size affirms the dependability of the derived statistical measures. Delving into the details, the average star rating is pegged at 3.57, revealing a sentiment that skews mildly positive, especially when considering that this average is closer to the pinnacle score of 5 than to the floor value of 1. However, with a standard deviation of 1.63, there's a clear indication of a broad dispersion in the star ratings. This suggests varied experiences and perceptions among the users regarding the product or service. A closer look reveals that 25% of the users rated the product at 2 stars or below, hinting at some level of dissatisfaction. Yet, in a redeeming light, the median rating stands at 4, signifying that half of the users found the product commendable. Further testament to the positive inclination is that three-quarters of all ratings touch the maximum, with many users deeming the product or service impeccable.

Sentiment Scores Summary Statistics: Aligning perfectly with the star ratings, the sentiment scores, too, register a count of 459,367, facilitating an exact one-to-one comparison. The average sentiment score echoes closely with the mean star rating, settling at 3.50, buttressing the reliability and precision of the sentiment analysis process. This score translates to a balanced interplay of positive and negative sentiments, though there's a subtle tilt towards the positive spectrum. The sentiment scores have a standard deviation of 1.80, which surpasses that of the star ratings. This indicates a more extensive variability, potentially because sentiments can encapsulate more intricate emotions than a mere numerical grade. It's noteworthy that, akin to the star ratings, the lowest sentiment scores depict sheer dissatisfaction. The 25th percentile showcases that sentiments lean negative for a quarter of the reviews, resonating with the star ratings. Yet, a significant revelation is the median sentiment score, which peaks at 5, underscoring an overwhelmingly positive sentiment in the textual content of most reviews. This buoyancy in sentiment is further highlighted by the fact that the 75th percentile also hits the zenith, confirming a dominantly positive sentiment amongst the reviews.

3.3 Star Ratings Correlation with Sentiment Analysis

The correlation between sentiment scores and star ratings was notably high, with a Pearson's correlation coefficient of 0.7501. This suggests a strong positive relationship between the sentiment of the review and the given star rating. In simpler terms, reviews with positive sentiments tend to have higher star ratings and vice versa. Further statistical verification was conducted using the Chi-squared test, resulting in a value of 254,699.58. The associated P-value was found to be 0.0, indicating this correlation is statistically significant.

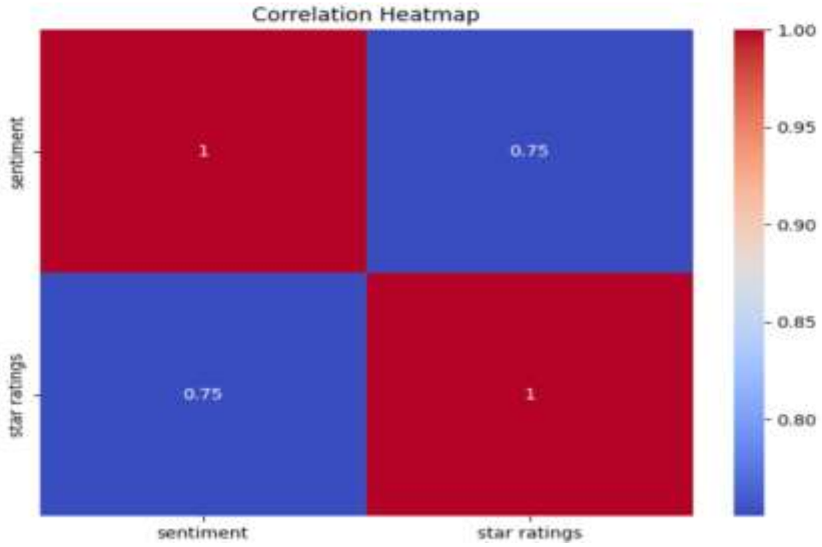


Figure 3.2: Heatmap representation of the correlation coefficient of Ratings and Sentiment polarity score.

3.4 Keywords Analysis and Relationship with Customer Behaviours

The word frequency analysis illuminated the specific terminologies recurrent in customer reviews. For those expressing satisfaction, keywords like 'Great', 'use', 'work', 'good', 'easy', 'product', ,'year', 'love', 'program' and ‘software’ were top 10, reflecting contentment, practicality, and a generally affirmative experience with the product or service. In contrast, reviews tinged with dissatisfaction frequently featured terms such as ‘work', 'product', 'software', 'use', 'program', 'buy', 'version', 'computer', 'try' and ‘time’. From these, it's evident that many negative sentiments stemmed from issues related to functionality, challenges with specific software versions, and overarching concerns about usability.

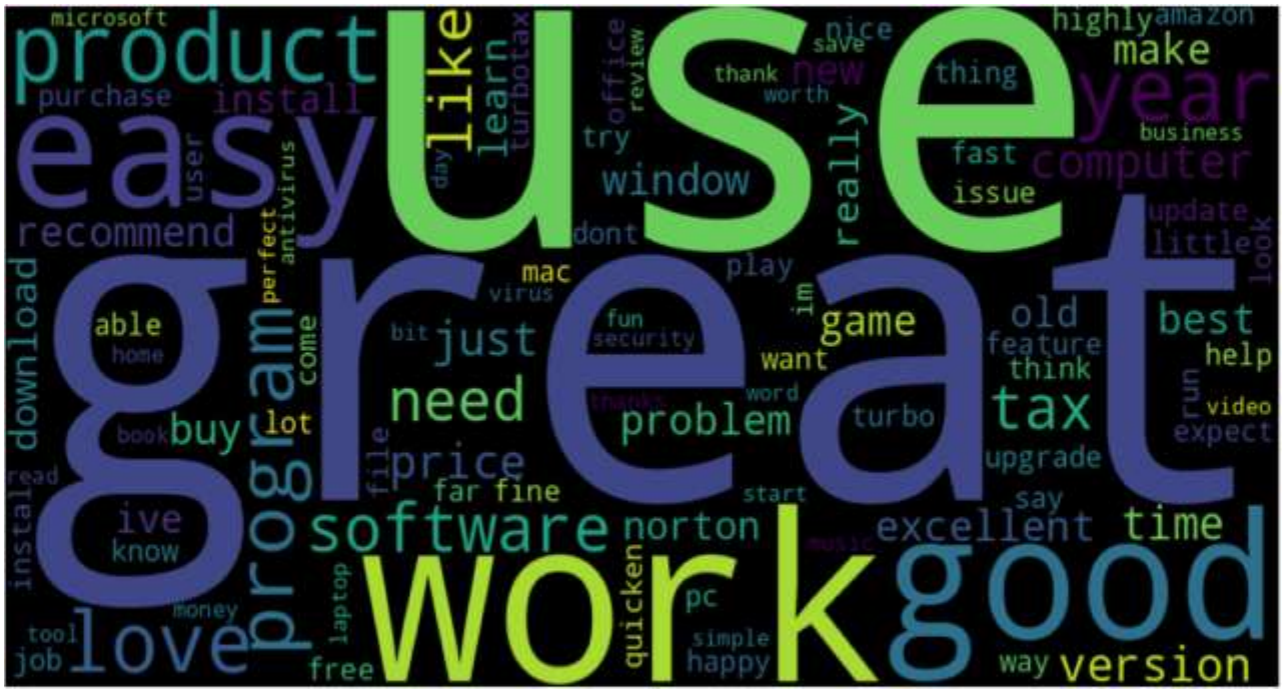


Figure 3.3: Word cloud representation of the most frequent words in positive reviews.

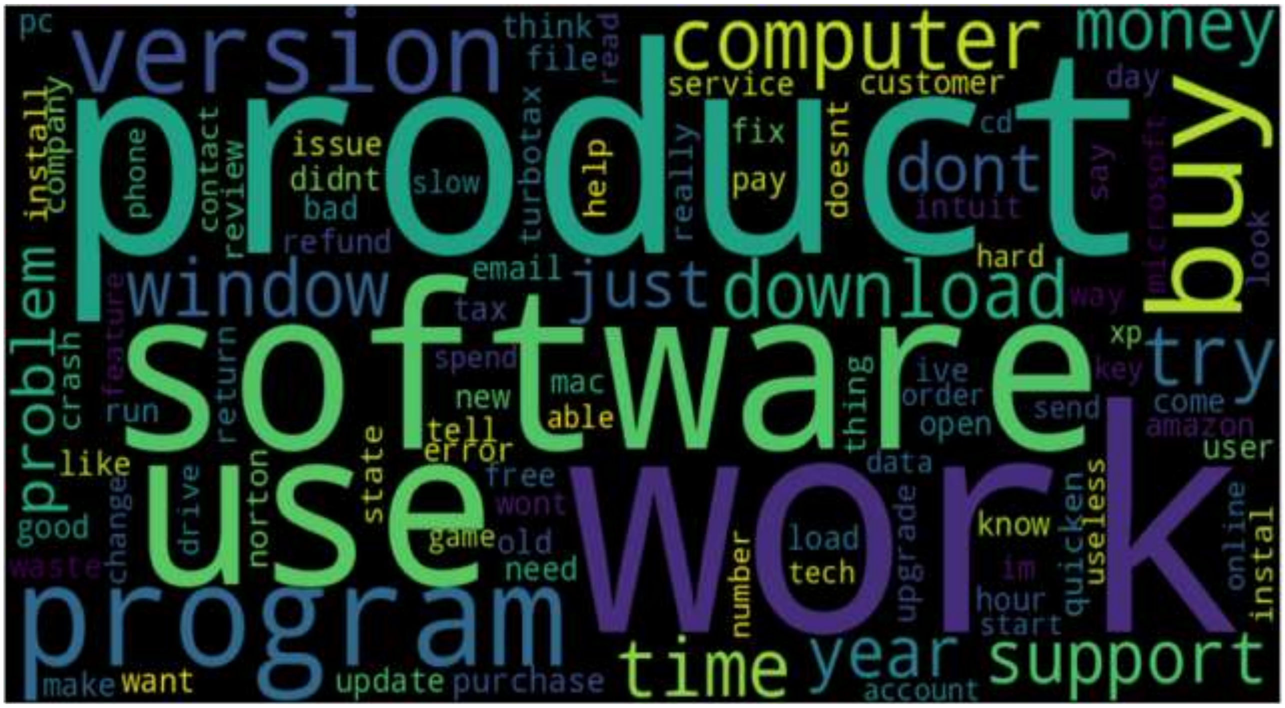
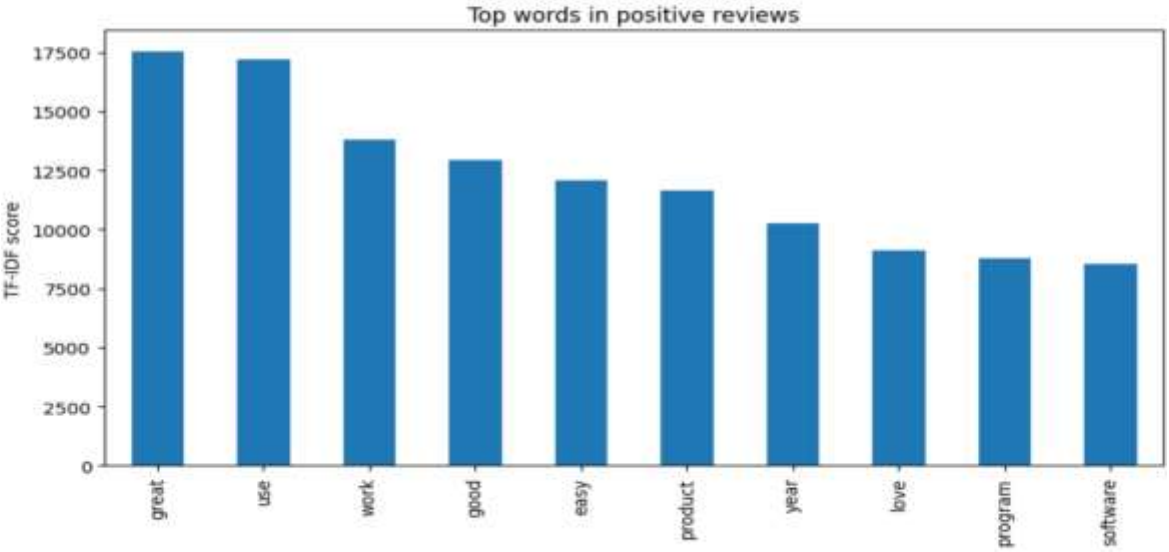


Figure 3.4: Word cloud representation of the most frequent words in negative reviews.



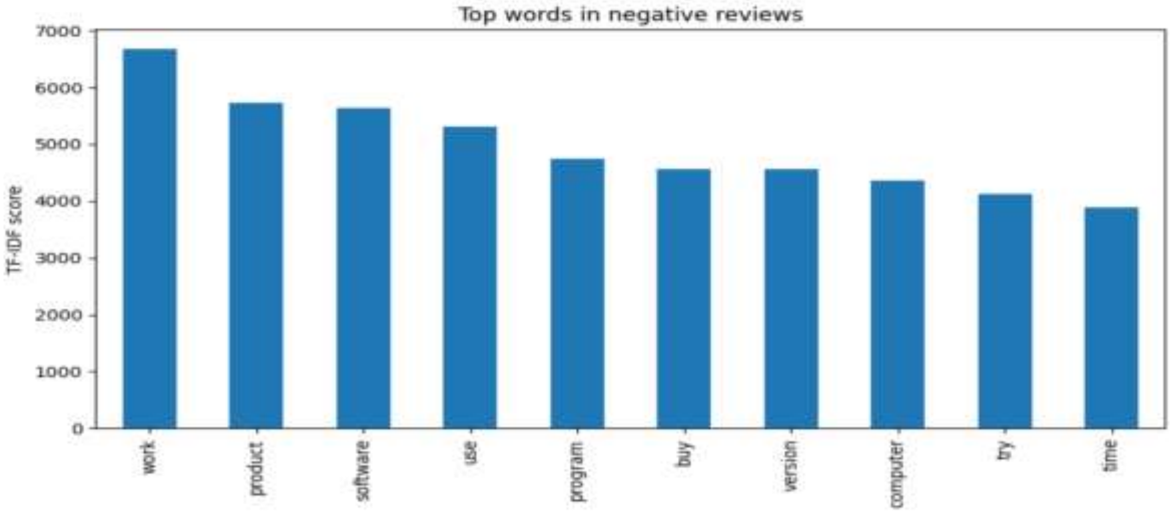


Figure 3.5: Top 10 most frequent words in negative and positive reviews

3.5 Performance Metrics of Traditional models

The Random Forest sentiment model showcases a strong performance with an accuracy of 0.82, recall of 0.86, and precision of 0.85, which is further affirmed by its 0.85 F1 score and 0.90 ROC score. The RNN sentiment model, while slightly less accurate at 0.81, maintains a good balance of precision (0.86) and recall (0.83), resulting in an F1 score of 0.85 and an ROC score of 0.89, indicating a slight preference towards precision and a robust discriminatory power. In the star rating comparison, the Logistic Regression model stands out with the highest accuracy of 0.86, a uniform score of 0.85 across recall, precision, and F1 score, and a notable ROC score of 0.93, demonstrating excellent classification and balance. On the other hand, the RNN star rating model, though slightly lagging with an accuracy of 0.83 and scores around 0.82-0.83 for recall, precision, and F1, still presents a solid performance with an ROC score of 0.91, hinting at slightly lesser differentiation ability between rating categories.

Model Class	Model Name	Validation Accuracy	Precision	Recall	F1 Score	ROC
Sentiment Model	Random Forest	0.82	0.85	0.86	0.85	0.90
Sentiment Model	Recurrent Neural Network	0.81	0.86	0.83	0.85	0.89
Star Rating Model	Logistic Regression	0.86	0.85	0.85	0.85	0.93
Star Rating Model	Recurrent Neural Network	0.83	0.83	0.82	0.83	0.91

Table 3.1: Performance metrics of traditional models.

3.5 Performance Metrics of Hybrid models

- Random Forest (Sentiment model) & Logistic Regression (Ratings model)

The model performs well when the sentiment weights are kept in the range of 0.0 and 0.6. In this range, the ROC score only slightly improves without impacting other performance metrics. However, the model's performance declines noticeably after the sentiment weight exceeds 0.6, indicating that a high sentiment weight may reduce the model's overall efficacy. With more substantial ROC scores and consistent performance in other areas within this range, it is a good idea to put the sentiment weight at 0.6 or less to get the best outcomes.

Sentiment weight	Ratings weight	Validation Accuracy	Precision	Recall	F1 Score	ROC
0.0	1.0	0.86	0.85	0.84	0.85	0.926
0.1	0.9	0.86	0.85	0.85	0.85	0.929
0.2	0.8	0.86	0.86	0.85	0.84	0.930
0.3	0.7	0.86	0.85	0.84	0.84	0.931
0.4	0.6	0.86	0.85	0.84	0.85	0.932
0.5	0.5	0.86	0.85	0.85	0.85	0.930
0.6	0.4	0.86	0.85	0.85	0.85	0.927
0.7	0.3	0.85	0.85	0.84	0.84	0.922
0.8	0.2	0.85	0.83	0.83	0.83	0.914
0.9	0.1	0.84	0.83	0.82	0.82	0.904
1.0	0.0	0.82	0.81	0.81	0.81	0.891

Table 3.2: Random forest (sentiment model) & Logistic Regression (Rating Model)

- **RNN (Sentiment model) & RNN (Ratings model)**

It is evident from looking at the data set that the model performs consistently and well when the sentiment weight is between 0.0 and 0.6. The precision metric shows a progressive increase, reaching its highest values at sentiment weights of 0.9 and 1.0, while the validation accuracy within this range peaks at a sentiment weight of 0.3. The recall value is somewhat consistent up to a sentiment weight of 0.2 before beginning to moderately fall, showing that as the sentiment weight increases, it becomes less and less effective at detecting positive cases. Up until a sentiment weight of 0.6, which indicates balanced classification performance, the F1 score, which offers a thorough assessment of the model's precision and recall, remains consistent. The ROC score, which shows how well the model can differentiate between classes, gradually rises until a sentiment weight of 0.4, after which it slightly decreases while maintaining a score above 0.9, indicating good model performance. Since greater ROC scores and other stable metrics are seen within this range, maintaining a sentiment weight at or below 0.6 may thereby optimise performance.

Sentiment weight	Ratings weight	Validation Accuracy	Precision	Recall	F1 Score	ROC

0.0	1.0	0.83	0.85	0.88	0.87	0.906
0.1	0.9	0.83	0.85	0.88	0.87	0.908
0.2	0.8	0.83	0.86	0.88	0.87	0.910
0.3	0.7	0.85	0.86	0.88	0.87	0.911
0.4	0.6	0.84	0.86	0.87	0.87	0.912
0.5	0.5	0.84	0.87	0.87	0.87	0.911
0.6	0.4	0.84	0.87	0.86	0.87	0.910
0.7	0.3	0.83	0.87	0.86	0.86	0.909
0.8	0.2	0.83	0.87	0.85	0.86	0.907
0.9	0.1	0.83	0.88	0.84	0.86	0.904
1.0	0.0	0.82	0.88	0.83	0.85	0.901

Table 3.3: RNN (Sentiment model) & RNN (Ratings model)

- **Random Forest (Sentiment model) & RNN (Ratings model)**

According to the data, the hybrid model works best when sentiment weights are between 0.0 and 0.6. At these sentiment weight levels, the model maintains constant validation accuracy, precision, and recall metrics, with a modest peak in performance occurring at 0.3 sentiment weight levels. Throughout this range, the F1 score remains constant, demonstrating a balanced model. A noteworthy improvement in the model's classification performance can be seen in the ROC, which peaks at a sentiment weight of 0.5. Therefore, it is advised to set the sentiment weight at 0.6 or lower to get the best results, as this range shows the greatest ROC values and the most stable other metrics.

Sentiment weight	Ratings weight	Validation Accuracy	Precision	Recall	F1 Score	ROC
0.0	1.0	0.83	0.85	0.88	0.87	0.906
0.1	0.9	0.83	0.85	0.88	0.87	0.911
0.2	0.8	0.84	0.86	0.87	0.87	0.910
0.3	0.7	0.84	0.86	0.88	0.87	0.918
0.4	0.6	0.84	0.87	0.88	0.87	0.920
0.5	0.5	0.84	0.87	0.88	0.87	0.921
0.6	0.4	0.84	0.87	0.88	0.87	0.920
0.7	0.3	0.84	0.87	0.87	0.87	0.916

0.8	0.2	0.84	0.86	0.87	0.87	0.910
0.9	0.1	0.83	0.86	0.86	0.86	0.902
1.0	0.0	0.82	0.86	0.85	0.85	0.89

Table 3.4: Random forest (Sentiment model) & RNN (Ratings model)

- **RNN (Sentiment model) & Logistic Regression (Ratings model)**

The performance metrics show that the hybrid model, which combines Logistic Regression (used for ratings) and RNN (used for sentiment analysis), performs reasonably well when the sentiment weight is between 0.0 and 0.4. This performance is characterised by consistent validation accuracy and F1 score, as well as a slight improvement in the ROC score, which peaks at a sentiment weight of 0.3. The ROC score noticeably decreases with a sentiment weight of 0.5 despite other metrics remaining consistent, suggesting a decline in the model's discriminating power. Validation accuracy, recall, and ROC scores gradually decrease as the sentiment weight gets closer to 1.0, suggesting declining performance. For optimal results, maintaining a sentiment weight below 0.5 seems to be beneficial, as it aligns with higher ROC values and a stable performance across other metrics.

Sentiment weight	Ratings weight	Validation Accuracy	Precision	Recall	F1 Score	ROC
0.0	1.0	0.85	0.87	0.90	0.88	0.926
0.1	0.9	0.86	0.87	0.90	0.88	0.929
0.2	0.8	0.86	0.88	0.90	0.89	0.931
0.3	0.7	0.86	0.88	0.89	0.89	0.932
0.4	0.6	0.86	0.88	0.89	0.89	0.931
0.5	0.5	0.86	0.88	0.88	0.88	0.830
0.6	0.4	0.85	0.88	0.87	0.88	0.927
0.7	0.3	0.84	0.88	0.86	0.87	0.923
0.8	0.2	0.84	0.88	0.85	0.87	0.917
0.9	0.1	0.83	0.88	0.84	0.86	0.910
1.0	0.0	0.82	0.88	0.83	0.85	0.901

Table 3.5: RNN (Sentiment model) & Logistic Regression (Ratings model)

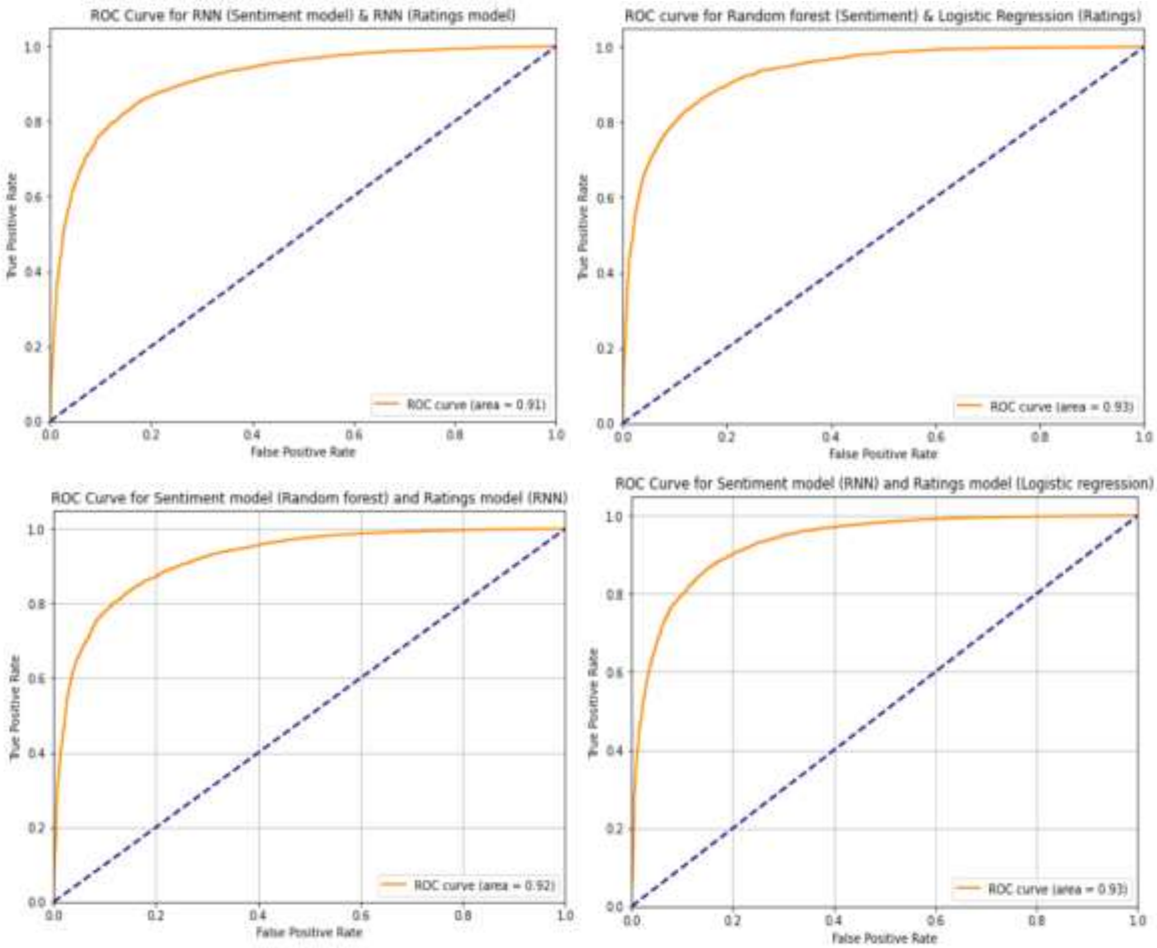


Figure 3.6: ROC curve for optimal weights across the hybrid models.

4 Discussion

4.1 Evaluation

The previous chapter's comparative analysis of the performance metrics for both traditional and hybrid models provide essential insights into the advantages and disadvantages of each strategy. A thorough analysis of these models can aid in understanding the nuances of how specific algorithms behave when sentiment and rating components are given varying weights and how these configurations affect the overall performance.

4.1.1 Traditional Models

The Random Forest sentiment model stands out due to its stellar performance characterized by a high degree of accuracy and exceptional classification abilities, a testament to its high ROC score. It showcases an adept ability to discern between positive and negative sentiment classes, thereby cementing its reliability for sentiment analysis tasks. In contrast, the Recurrent Neural Network, utilized both for sentiment analysis and star rating predictions, might slightly trail in the accuracy department compared to other models. However, it preserves a harmonious balance between precision and recall, thereby exhibiting significant discriminatory prowess. This underscores the model's slight inclination towards prioritizing precision, thereby making it an optimal choice in scenarios where minimizing false positives is key. On the other hand, the Logistic Regression model, specifically designed for star rating predictions, surfaces as a dominant contender with unparalleled accuracy. It demonstrates a uniform distribution of scores across various critical metrics such as precision, recall, and the F1 score. Moreover, its high ROC score echoes its excellent classification capabilities and well-rounded performance, positioning it potentially as the safest bet for rating prediction tasks, particularly when striving for a balanced classification outcome.

4.1.2 Hybrid Models

Hybrid models have been acknowledged for their superior flexibility, facilitating enhanced tuning of sentiment and rating weights to further hone performance. Through meticulous assessment of varying combinations of sentiment and rating weights, a discernible pattern has emerged; the sentiment weight spectrum between 0.0 and 0.6 frequently procures optimal performance across diverse setups.

The synthesis of Random Forest and Logistic Regression models substantiates this observation, exhibiting solid performance specifically within the 0.0 to 0.6 sentiment weight range. This duo sustains steady performance indicators with a subtle ascendancy in ROC scores, suggesting that fine-tuning the balance between sentiment and rating weights could potentially amplify the model's effectiveness. Similarly, the RNN model, applied both for sentiment and rating analyses, manifests a stable and commendable performance within the same sentiment weight range. A notable trait is the gradual enhancement in precision correlating with an increment in sentiment weight, hinting at its appropriateness for assignments where precision is a priority.

Furthermore, the combination of Random Forest and RNN mirrors this preference for the sentiment weight domain of 0.0 to 0.6, maintaining uniformity in most of the performance metrics and attaining a zenith in the ROC score specifically at a sentiment weight of 0.5, indicating an optimum classification competency within this bracket. In contrast, the alliance of RNN and Logistic Regression finds its forte in a slightly restrained sentiment weight range of 0.0 to 0.4. While the metrics remain stable, a marked decline in the ROC score at a sentiment weight of 0.5 forecasts a possible diminution in classification effectiveness beyond this juncture.

4.1.3 Personal Evaluation

The process of developing and fine-tuning models for sentiment analysis and star rating prediction has been a journey filled with learning and growth. The limitations encountered during the project have not only been challenges but also opportunities to adapt and innovate in finding solutions. Despite the hurdles, witnessing the performance metrics of the developed models has been gratifying. The hybrid models showcased an interesting potential in leveraging the strengths of both sentiment analysis and star rating predictions, indicating a promising avenue for further research and development.

In my opinion, the research has demonstrated that even with limited resources, it is possible to develop models that can offer substantial performance in sentiment analysis and star rating prediction tasks. The experience has also underscored the importance of resource management and planning in the development of machine learning models, as the computational demands can be quite significant. Moving forward, I envision further refining these models, perhaps exploring options for optimizing computational resources and addressing the dataset imbalance more effectively. It also opens a pathway for exploring more sophisticated hybrid models that can potentially offer even more reliable predictions. This research has been a significant step in my journey in the field of machine learning and data science, providing a robust platform to build upon in future projects. It has been a rich learning experience, where the challenges encountered were equally opportunities for growth and innovation.

4.2 Limitations

One significant limitation encountered during this study was the constraint of computational resources. Despite leveraging the capabilities of Google Colab's GPU subscription, the fine-tuning of intricate pre-trained models like BERT proved to be quite laborious. The heightened computational demands circumscribed the scope of model optimization, potentially curtailing the performance and efficiency of the tested models. Concurrently, the dataset utilized in this study demonstrated a noticeable degree of imbalance, potentially undermining the reliability of the predictive models. This imbalance could engender models with a bias towards the majority class, consequently affecting the precision and recall metrics for the minority class. Implementing strategies to equilibrate the dataset might foster enhancements in the overall performance of the predictive models, thereby mitigating the adverse effects arising from both computational constraints and dataset imbalance.

5 Conclusion

This research presented a comprehensive analysis of various machine learning and deep learning models applied in the domain of sentiment analysis and star ratings prediction. The performance metrics showcased that the individual traditional models hold strong positions in their respective specialties, with Random Forest and Logistic Regression particularly demonstrating potent capabilities in sentiment analysis and rating predictions, respectively. The exploration into hybrid models further substantiated the potential of combining different analytical approaches to optimize performance. It was observed that adjusting the weights assigned to sentiment and ratings could significantly influence the performance outcomes, often offering avenues to enhance the accuracy and classification abilities of the models.

Considering the insights derived from the analysis, it can be inferred that while traditional models offer strong and reliable performance, hybrid models open doors to further optimizations, allowing for tailored approaches to specific tasks. It suggests that the path to achieving the best performance might lie in a careful and considered combination of various model strengths, leveraging the complementary advantages that different models bring to the table.

This research underscores the importance of not only selecting the right models but also the right combination of features and weight configurations to achieve optimal performance in sentiment analysis and ratings prediction tasks. It sets a promising stage for further exploration and innovations in this domain, paving the way for more sophisticated and efficient predictive analytics solutions.

References

Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). *Random forest and support vector machine based hybrid approach to sentiment analysis*. *Procedia Computer Science*, 127, 511–520.

Alamoudi, E. S., & Alghamdi, N. S. (2021). *Sentiment classification and aspect-based sentiment analysis on Yelp reviews using deep learning and word embeddings*. *Journal of Decision Systems*, 30(2–3), 259–281.

Alqaryouti, O., Siyam, N., Abdel Monem, A., & Shaalan, K. (2024). *Aspect-based sentiment analysis using smart government review data*. *Applied Computing and Informatics*, 20(1–2), 142–161.

An, S., Kim, D., & Kim, S. (2018). *Recurrent neural network training with dark knowledge transfer*. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4254–4263.

Asthana, A. (2022). *Hybrid Recommendation System: A Review*. *International Journal of Computer Applications*, 184(32), 1–7.

Drolet, M.-J., Rose-Derouin, E., Leblanc, J.-C., Ruest, M., & Williams-Jones, B. (2023). *Ethical issues in research: Perceptions of researchers, research ethics board members, and research ethics experts*. *Journal of Academic Ethics*, 21(2), 269–292.

Hao, Y., Zhang, Y., & Li, X. (2023). *Sentiment recognition and analysis method of official document text based on BERT-SVM model*. *Journal of Physics: Conference Series*, *2318*(1), 012034.

Heguerte, L. B., Bugeau, A., & Lannelongue, L. (2023). *How to estimate carbon footprint when training deep learning models? A guide and review*. *Environmental Research Communications*, 5(8), 085001

Javed, A., Rafi, M., & Baig, A. R. (2021). *A Survey on Content-Based Filtering Recommender Systems*. *Journal of Information Science and Engineering*, 37(4), 1011–1031.

Kulkarni, S., & Rodd, S. F. (2020). Context Aware Recommendation Systems: A review of the state of the art techniques. *Computer Science Review*, 37, 100255

Kulkarni, S., & Rodd, S. F. (2020). *Context Aware Recommendation Systems: A review of the state of the art techniques*. *Computer Science Review*, 37, 100255.

Lengkeek, S., van der Lee, C., & Hogenboom, F. (2023). *Leveraging hierarchical language models for aspect-based sentiment analysis on financial data*. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1234-1245.

Li, Z., Zhang, Y., & Chen, L. (2022). Personalized recommendation model of electronic commerce in new media environment. *Frontiers in Psychology*, 13, 952622.

Li, Z., Zhang, Y., & Chen, L. (2022). *Personalized recommendation model of electronic commerce in new media environment*. *Frontiers in Psychology*, 13, 952622.

Logesh, R., Subramaniyaswamy, V., & Vijayakumar, V. (2019). *A Hybrid Personalized Recommender System Using Clustering and Association Rule Mining*. *Journal of Ambient Intelligence and Humanized Computing*, 10(5), 1955–1965.

Mutinda, J., Mwangi, W., & Okeyo, G. (2023). *Sentiment analysis of text reviews using lexicon-enhanced BERT embedding (LeBERT) model with convolutional neural network*. *Applied Sciences*, 13(3), 1445

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). *Deep learning applications and challenges in big data analytics*. *Journal of Big Data*, 2(1), 1

Osman, T., Noah, S. A. M., & Darwich, M. (2019). *A Hybrid Recommender System Based on Context Awareness and Sequential Behavior*. *IEEE Access*, 7, 186264–186278.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., & others. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 15.

Roy, S., & Dutta, P. (2022). *A Hybrid Recommender System Using Collaborative Filtering and Content-Based Filtering for Improved Recommendation*. *Procedia Computer Science*, 198, 85–91.

Shakhovska, N., Medykovskyy, M., & Syerov, Y. (2020). *Logistic regression for star rating prediction*. *Advances in Intelligent Systems and Computing*, 1080, 697–707.

Suresh, A., & MJ, C. M. B. (2020). A Comprehensive Study of Hybrid Recommendation Systems for E-Commerce Applications. *International Journal of Advanced Science and Technology*, 29(3), 4089–4101.

Suresh, A., & MJ, C. M. B. (2020). *A Comprehensive Study of Hybrid Recommendation Systems for E-Commerce Applications*. *International Journal of Advanced Science and Technology*, 29(3), 4089–4101.

Walek, B., & Fajmon, P. (2023). *A Hybrid Recommender System for E-commerce Based on Customer Reviews and Product Features*. *Applied Sciences*, 13(1), 123.

Wang, J., Zhang, Y., Yuan, S., & Zeng, D. D. (2020). A Systematic Study on the Recommender Systems in the E-Commerce. *IEEE Access*, 8, 45459–45470.

Wang, J., Zhang, Y., Yuan, S., & Zeng, D. D. (2020). *A Systematic Study on the Recommender Systems in the E-Commerce*. *IEEE Access*, 8, 45459–45470.