KU LEUVEN

RESEARCH CENTRE FOR INFORMATION SYSTEMS ENGINEERING (LIRIS)

Transformer-based Sentiment Analysis: Investigating Misclassification Patterns in Amazon Book Reviews

Group report

Hassan Kamran	Ayesha Riaz	Mohamad Al-Homsi
R0974307	R0974212	R0714957

Thesis submitted to obtain the degree of
Master of Information Management

Promoter: Prof. Dr. Galina Deeva

Daily Supervisor: Christopher Bockel genannt Rickermann

Academic year: 2023-2024

Abstract. In the e-commerce landscape, sentiment analysis has gained immense popularity in identifying and categorizing opinions expressed in review texts to make informed decisions. Despite having numerous comparisons between models based on performance, literature often lacks an exploration of the misclassification reasons for particular models. This study fine-tunes four pre- trained transformer models namely BERT, RoBERTa, DistilBERT, and GPT-2 for a fine-grained (5-rating) sentiment analysis on Amazon books review data. It not only analyzes performance using robust evaluation metrics but also attempts to understand the rationale behind model misclassifications based on textual challenges or certain limitations of the models themselves. On the 5-class scale, DistilBERT excels in predicting ratings '1' and '2', RoBERTa in predicting ratings '4' and '5', and GPT-2 in predicting the rating '2'. Subsequently, we aggregate ratings into positive, neutral, and negative sentiments, and conduct an in-depth inspection of the misclassified reviews to uncover common error patterns. On the aggregated scale with 3 labels, RoBERTa outperforms other models in detecting negative sentiments, GPT-2 in detecting neutral sentiments, and DistilBERT in classifying positive sentiments. The analysis of misclassifications further reveals distinctive biases among the models. RoBERTa tends to label reviews as negative, GPT-2 leans towards neutrality, while BERT and DistilBERT exhibit a propensity towards positive sentiment classification. Additionally, 46% of the common misclassifications across the four models are attributed to user errors in rating a review, 31% to the subjective nature of the reviews, and 23% to contextual complexities in the text which the models struggle to effectively capture. This research offers insights into the nuanced performance of transformer models in sentiment analysis tasks, highlighting both their strengths and inherent biases, which could inform future advancements in NLP methodologies and contribute to enhancing decision-making processes in e-commerce platforms.

Keywords: Sentiment analysis, Amazon book reviews, Transfer learning techniques, Pre-trained models, Model errors

Acknowledgements. We extend our deepest gratitude to our thesis promoter Dr. Galina Deeva along with our daily supervisor Christopher Bockel genannt Rickermann, who provided us with their unwavering support, guidance and invaluable insights throughout this research endeavor. Their encouragement and constant support have been instrumental throughout the process. They consistently provided us valuable feedback and suggestions throughout our journey to ensure that we produce a compelling thesis. Therefore, this thesis is a product of not only our efforts and commitment, but their constant support and encouragement as well.

Table of Contents

1 Introduction 5

2 Literature Review 7

2.1 Traditional approaches to Sentiment Analysis 8

2.2 The gradual shift towards deep learning approaches 8

2.3 Pre-trained Transformer models: A breakthrough in existing NLP tasks 9

3 Methodology 12

3.1 Theoretical background 12

3.2 Experimental setup 14

4 Results and Discussion 19

4.1 Model fine-tuning 19

4.2 Model performance based on evaluation metrics 20

4.3 Misclassification analysis 23

5 Critical Reflection and Future Scope 28

6 Conclusion 29

References 31

Appendix 36

List of Figures 40

List of Tables 41

1 Introduction

Today's e-commerce landscape is significantly shaped by user-generated content in the form of online reviews that serve as a source of word-of-mouth marketing to influence consumer perceptions and purchase decisions (Chevalier and Mayzlin, 2006). Several studies have found that products/services with a higher volume of positive reviews tend to attract more trust, and in return, sales as social proof (the influence of others' opinions) has a crucial impact on our own behavior (Chen et al., 2008; Chevalier and Mayzlin, 2006; Clemons et al., 2006). These studies have explored various industries, spanning from beers to movies, and extending to digital marketplaces like Amazon. Buyers can use existing reviews for information search and alternatives evaluation to make better purchase decisions (Kohli et al., 2004). Moreover, the presence of reviews can improve their perception of the credibility of an online store and create a sense of community among them (Kumar and Benbasat, 2006). This is the reason why websites like Amazon, Yelp, and TripAdvisor give utmost importance to reviews, for example, Amazon strategically places reviews above the product specifications and details on its website.

Not only do online reviews aid consumers in their purchase decisions, but they also provide sellers with invaluable insights into purchasing behaviors (Chevalier and Mayzlin, 2006). Any organization needs to know the needs and expectations of consumers to effectively tailor strategies to aid product improvements. Reviews have provided new opportunities for improved and more accurate assessment of market needs, target customer segments, and customer behaviors, to move from the status quo of the existing propositions of an organization to the desired state of customer satisfaction levels (Wang et al., 2018; Mejova, 2009). This underscores the importance of leveraging reviews to gain perspectives into consumer tastes and refine product offerings to improve overall satisfaction.

Amazon.com, one of the largest e-commerce companies globally, boasts a vast catalog of over 12 million products (Reisinger, 2017). With a staggering 310 million active customers as of February 2017 (Statista, 2018), Amazon serves as a rich source of user-generated reviews which serve as an integral part of the platform; encompassing diverse categories such as books, clothing, electronics, fashion, and more. A common approach Amazon employs for quantitative feedback is the star ratings on a Likert scale from 1-5, that accompany reviews and are pivotal in sentiment identification (Mejova, 2009). In addition to providing customers a quick overview of the product quality without having to go through lengthy review texts, ratings often also play a crucial role in product recommendations and search algorithms in digital marketplaces like Amazon, enhancing product visibility and potential for sales (Godes and Mayzlin, 2004). The issue, then, arises from the fact that many reviews contain textual evaluations only, that offer less utility to consumers, and are hard to integrate with the automated systems; especially when quantitative comparisons between products are needed. This emphasizes the importance of tools capable of predicting ratings from textual reviews.

Even if ratings already exist, another challenge lies in whether they accurately reflect customer sentiment or not. Several studies have analyzed the distributions of ratings on major platforms and concluded that they tend to be heavily skewed towards the positive side (Hu et al., 2009). For example, a comparison between the ratings on Airbnb and TripAdvisor revealed that Airbnb ratings are evidently more inflated than those on TripAdvisor for the same property (Zervas et al., 2021). This can be due to two reasons; either the reviews are incentivized, or simply, the ratings are inflated even if the reviews are reflective of the true experience and indicative of a potential issue. This is because guests may feel pressured to leave higher ratings (Whalen, 2023). While Amazon somewhat tackled the incentivized reviews by the introduction of Amazon Vine in 2007; an invite-only program that invites trusted Amazon reviewers to review new products, issue of inflated ratings might persist. Therefore, models capable of predicting accurate user ratings from textual reviews are of paramount importance.

The most popular method for the automatic classification of reviews is sentiment analysis (SA); a subfield of natural language processing (NLP), which is commonly defined as the process of studying subjective elements like words or sentences to identify and extract the emotions or thoughts of a text (Medhat, 2014; Mejova, 2009). Traditional approaches to SA including lexicons and machine learning classifiers are well-documented in extant literature and can be utilized by platforms like Amazon to classify textual reviews into corresponding ratings. In the past, researchers have discussed the challenges associated with traditional approaches including the requirements for excessive data pre-processing and manual feature selection. Recent studies and practical applications have increasingly favored deep learning models over traditional machine learning, owing to the availability of huge datasets and advancements in computer hardware and word embedding algorithms.

In particular, the transformer architecture (Vaswani et al., 2017) has gained prominence as a groundbreaking neural network architecture increasingly being employed by researchers to deliver remarkable performance in NLP tasks. It relies on self-attention mechanisms, eliminating the need for hand-crafted features and recurrent or convolutional layers, that are common challenges associated with traditional machine/deep learning algorithms. Researchers have been using transfer learning approaches to fine-tune pre-trained transformer models on review datasets, comparing the performance to traditional deep learning models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs). They have revealed state-of-the-art performance with much less effort through the use of pre-trained transformer models.

Our research seeks to address several gaps in existing literature on sentiment analysis in the context of Amazon book reviews. Previous studies have predominantly focused on demonstrating the transformers' superior performance to traditional approaches, and lack a theoretically grounded explanation for differences in model performance, typically relying on architectural comparisons. In contrast, our approach shifts from evaluating these models based on standard metrics to an in-depth analysis of different factors influencing the model predictions. By doing so, we aim to identify and explore the

various errors made by the models based on certain textual properties of the reviews. Our objective is to identify prominent challenges within the texts that present difficulties for the models, and discern how they are interpreted and addressed by each model. Additionally, we seek to determine whether certain errors are common across all models or unique to each, setting them apart from one another.

In this thesis, we fine-tune four pre-trained transformer models namely BERT, RoBERTa, DistilBERT, and GPT-2 on a dataset of Amazon book reviews in an effort to predict the rating value. The desired output to an input of a text review is a "star" rating on a continuum from 1 to 5. To gain a deeper understanding of the reviews that each model misclassified, we conduct an in-depth analysis of the predictions; with the goal of uncovering potential limitations of each model, issues related to the dataset, or the process of annotation. We introduce a novel approach by leveraging the GPT-2 model for classifying book reviews. While BERT and its variants have been widely acclaimed as state-of-the-art models, we believe that exploring the capabilities of GPT-2 in this context could provide valuable insights. Additionally, previous studies have primarily utilized binary or three-class classification approaches. In contrast, we intend to push the boundaries by employing a five-class classification approach. Although this may lead to a potential drop in performance, we are prepared to explore this uncharted territory to gain a deeper understanding of SA in the context of Amazon book reviews.

This remainder of the paper is structured as follows. Section 2 presents a review of the existing literature. Section 3 describes the methodology including the theoretical background of pre-trained transformers and related concepts, as well as the details of the experiments conducted along with the evaluation measures used. Section 4 discusses the research results and their analysis, while section 5 critically reflects on the approach and findings of the research, taking into account the limitations and potential avenues for future research. Lastly, section 6 concludes the methodology and results, focusing on the main research takeaways.

2 Literature Review

Extensive research in the field of SA focuses on various techniques for extracting features and classifying review polarities within datasets. Studies have utilized SA in various domains, for example, in e-commerce to classify reviews, in marketing to classify social media posts, and in politics to classify tweets. Within e-commerce, researchers have attempted to classify a vast category of reviews; movie reviews (Rehman et al., 2019), hotel reviews (Ishaq et al., 2021), drug reviews (Vijayaraghavan and Basu, 2020), and product reviews on platforms like Amazon and Yelp (Rathor et al., 2018; Sadhasivam and Kalivaradhan, 2019; Tan et al., 2018; Ali et al., 2024; Kusal et al., 2023; Durairaj et al., 2021; Alves et al., 2022).

2.1 Traditional approaches to Sentiment Analysis

A major research field has emerged around the subject of how to extract the most accurate method to categorize the customers' reviews into negative or positive opinions. Globally, the traditional approaches to SA can be classified into two main categories; lexicon-based and machine learning classifiers. Lexicon-based techniques use pre-compiled dictionaries or vocabularies to classify texts. SentiWordNet is one of the most popular lexical resources used for SA (Nguyen et al., 2018). However, developing or maintaining a sentiment lexicon poses significant challenges in the modern era, particularly with the rise of huge volumes of unstructured, user-generated data (Lagrari et al., 2021).

Several machine learning algorithms, when trained on labeled datasets, can classify the polarity of textual reviews using features extracted from the text like word frequencies or n-grams. The algorithms namely Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), Maximum Entropy (ME), and Naive Bayes (NB) have received considerable attention in the literature. In this regard, most papers focused on comparing the performance of various machine learning algorithms and proposing the model with the highest accuracy, focusing on review datasets spanning vast categories such as movies and clothing (Pang et al., 2002; Mullen and Collier, 2004). These researchers concluded SVMs to be the best performing model. Commonly used feature extraction techniques on review datasets include Bag-of-Words (BoW), N-grams, and Term Frequency–Inverse Document Frequency (TF–IDF) (Chauhan et al., 2020; Vijayaraghavan and Basu, 2020).

Similarly, in the context of Amazon reviews, Rathor et al. (2018) concluded that SVMs outperform NB and ME. Sadhasivam and Kalivaradhan (2019) went one step further to propose an Ensemble approach to classify the reviews and found it highly effective as compared to NB and SVM. Quite a few papers have also attempted to draw comparisons between machine learning classifiers and lexicon-based approaches, revealing the superior performances of machine classifiers in classifying Amazon reviews (Bhavitha et al., 2018; Nguyen et al., 2018). Most of these works classify the reviews at the document level, where an entire review (document) is treated as a single entity and broken down into sentences. Each sentence is then examined for its structure and each word within the sentence is analyzed for its contextual dependency to determine the sentiment orientation (Choi et al., 2020). It is also noteworthy that these studies employ binary classification, categorizing reviews as either 'positive' or 'negative'. Furthermore, their top-performing models achieved F1-scores, precision, and accuracy of approximately 85 - 90% at maximum.

2.2 The gradual shift towards deep learning approaches

There have been significant advancements in NLP methodology over the past decade. Researchers have been incorporating pre-trained word embeddings such as Word2vec and GloVe in their models to map discrete words of sequences to real valued

representations (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al.,2017). They have been motivated by the ability of word embeddings to solve challenges associated with traditional feature extraction techniques by extracting semantic and syntactic details from word representations. Kokab et al. (2022) highlights the limitations of word embeddings including their inability to handle out of vocabulary words and the fact that there is only embedding for a word by which it is represented in any context.

Subsequently, more sophisticated algorithms have laid the foundation for building deep learning (DL) architectures capable of extracting the syntactic, semantic, and sentiment features without domain dependency constraints. Most research in deep learning is centered around highlighting its superiority to traditional machine learning approaches that primarily focus on manual feature extraction and struggle with high-dimensional data, short texts, and new data patterns characterized by diverse linguistic forms. The shift towards deep learning stems from the rising amounts of training data and the inherent strengths of DL methods in understanding the sequential nature of language and automatically extracting hierarchical features and contextual representations from text data (Kokab et al., 2022; Mohbey, 2021; Ali et al., 2024, Zhang et al., 2018).

CNNs, RNNs, long short-term memory (LSTM), Bi-LSTM, and multilayer perceptron (MLP) are examples of popular models used in deep learning. Studies have predominantly concentrated on comparing different deep learning models with each other, as well as with traditional machine learning models, primarily in terms of model accuracy. Gadri et al., (2022), Mohbey (2021), and Tan (2018) revealed the superior performance of CNN and LSTM models to predict ratings from reviews by achieving performance evaluation measures above 90%, surpassing models like SVM, NB, and LR. When comparing DL methods amongst each other, researchers concluded that LSTMs integrated with GloVe and Word2vec representations achieve better performance as compared to CNNs and RNNs, because of their ability to deal with long-length reviews (Kim and Yeong, 2019; Ishaq et al., 2021). To deal with the very long dependencies of sentences, Bi-LSTM was utilized to achieve an accuracy that surpassed the traditional ML as well as other DL approaches (Ali et al., 2024). A few studies also advocated the adoption of hybrid models in SA, as deep learning models demonstrate improved performance when combined. A combination of CNN and LSTM has frequently been employed to leverage their respective strengths to detect polarity in Amazon reviews, achieving an accuracy exceeding 90% (Min, 2019; Rehman et al., 2019).

2.3 Pre-trained Transformer models: A breakthrough in existing NLP tasks

Vaswani et al. (2017) introduced the transformer architecture on top of the traditional deep learning approaches for sequence-to-sequence tasks, proposing self-attention mechanisms to overcome the excessive reliance on convolution and recurrence. The encoder/decoder layers and self-attention in this architecture allows it to attend to all positions in the input sequence, simultaneously enabling greater parallelization and the effective capture of long-range dependencies and contextual information. The authors

compared it to state-of-the-art RNN and CNN models and concluded that it outperformed those. Subsequently, most of the recent research in the field of SA has been shaped by the use of transformers, whereby numerous researchers have attempted to discuss their success in modeling long-range dependencies, learning from unlabeled data sets through self-supervised learning, and capturing syntactical, semantical, and contextual representations without the need for manual features (Kokab et al., 2022; Kotei and Thirunavukarasu, 2023).

More recently, transfer learning approaches have entered the discussion. Transfer learning in the context of NLP entails pre-training a network on large amounts of text and fine-tuning the weights afterwards on a downstream task with labeled data (Aßenmacher and Heumann, 2020). Pretrained models eliminate the need to build models from scratch and can optimize performance in downstream tasks without changing the representations, even on smaller datasets. They, therefore, mitigate the issue of overfitting, which is prevalent in deep learning applications with limited training datasets (Erhan et al., 2010; Zhou and Srikumar 2022).

Typical examples of NLP pretrained models include BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), RoBERTa (Robustly optimized BERT approach) (Liu et al., 2019), GPT-2 (Generative Pre-trained Transformer)'s variant (Radford et al., 2019), DistilBERT (Sanh et al., 2019), and XLNet (Yang et al., 2019). These models have been repeatedly employed in literature with the goal of substantially improving sentiment detection accuracy, particularly for Amazon reviews. The experimentation in most papers spans a diverse array of machine, ensemble, and deep learning and reveals state-of-the-art results with the use of transformer-based models. Ali et al. (2024) achieved an accuracy of 89% using BERT and 88% using XLNet, which surpassed the figures they achieved using Bi-LSTM and LR. Their BERT model, when applied to the Amazon dataset utilized by Tan et al. (2014), attained an impressive accuracy score of 93.7% and an F1 score of 93%, surpassing the 71.5% accuracy figure achieved by the LSTM model used in Tan et al. (2014)'s study. Similarly, when the same model was applied to Qorich and El Ouazzani (2023)'s study on Amazon reviews that achieved an accuracy of 90% with CNNs and word embeddings, it achieved an exceptional accuracy of 91%, further supported by an F1 score of 91.4% in the context of binary-class classification.

Kusal et al. (2023) also performed a comparison analysis between deep learning (LSTM and Bi-LSTM) and pre-trained models (Bert, RoBERTa, DistilBERT, Distil RoBERTa) where RoBERTa outperformed the other transformer and deep learning models resulting in an accuracy of 80.28% and F1 score of 0.800, followed by BERT (accuracy of 78.91% and F1 score of 0.789). Durairaj et al. (2021) evaluated linear models like SVM and fastText, deep learning models like BiLSTM and hybrid fastText-BiLSTM, as well as a BERT-base-cased model on Twitter, IMDB Movie Reviews, Yelp, and Amazon customer reviews datasets. They concluded that a fine-tuned pre-trained BERT model scores the highest in terms of an F1 score of 0.90, whereas the other models scored less than 0.85; thus, establishing BERT as the state-of-the-art

model with respect to performance, accuracy, training and testing on customer review datasets. Each of these studies conducted either binary classification of reviews, achieving accuracy levels exceeding 90%, or divided the sentiment scale (1-5) into three labels (negative, neutral, or positive). In the latter case, the accuracy measures dropped below 90%, and in some instances, below 80%.

Researchers have also been adopting hybrid approaches; for example, combining RoBERTa with LSTM (Tan et al., 2022) and BERT with CNN, RNN, and Bi-LSTM (Bello et al., 2023), and achieved state-of-the-art performances in accuracy, recall, and precision. While most studies have focused on a comparative analysis between pre-trained models and traditional machine/deep learning approaches in sentiment classification, Alves et al. (2022) performed an analysis to compare five pre-trained transformer models (FinBERT, BERT-base, FinBERTtone, DistilRoBERTa, and Twitter-RoBERTa) along with a sixth ensemble model in the classification of automotive product reviews on Amazon, again employing a three-class classification approach. They found that adding an ensemble of transformers to make the final prediction improves the performance.

In terms of model errors, researchers have recently started to investigate the predictions made by transformers. Ali et al. (2024) made one such attempt to gain insights into how linguistic cues affect sentiment predictions. They found out that positive reviews were characterized by high weights for positive words and downplayed negative words, while negative reviews emphasized negative sentiment words. This underscores the significance of understanding model interpretations for enhancing predictive capabilities in text analysis. To gain deeper insights into inaccurately classified comments, a thorough analysis of predictions was also conducted. The focus was on examining comments that the model incorrectly predicted, aiming to uncover potential issues with dataset annotation. Alves et al. (2022) also tried to analyze the best model for each prediction (positive or negative) and concluded that the ensemble and RoBERTa models were best for all metrics for negative and positive sentiment. For neutral sentiment, no model performed well. While these studies make initial attempts at exploring model performance, it remains an area that is largely unexplored and requires further investigation to uncover the intricacies of different models.

In summary, extensive research efforts have emphasized the superiority of pre-trained transformers over traditional methods in sentiment analysis of Amazon reviews. While traditional approaches have been valuable in achieving accuracies exceeding 80%, researchers have consistently highlighted the time-consuming nature of data preprocessing and feature extraction tasks associated with them. Conversely, deep learning methods, such as CNNs, RNNs, and Bi-LSTMs, have outperformed machine learning models, but also face challenges with the effective capture of long-range dependencies and contextual information. The recent shift in focus towards pre-trained models like BERT is attributed to the reduced training efforts and exceptional performance associated with them. However, existing studies primarily focus on comparisons with traditional approaches, with few attempts to compare transformer models based on the

rationale behind their predictions. Moreover, most studies employ binary or three-class classification at most, neglecting scenarios requiring more than three labels. This highlights the need for further research to explore the capabilities of transformer models comprehensively and address classification challenges beyond binary and ternary scenarios.

3 Methodology

3.1 Theoretical background

Pretrained models

BERT. Devlin et al. (2018) introduced BERT as a new language representation model, as opposed to previous techniques that were uni-directional. BERT's architecture is rooted in the Transformer model which is explained in **Fig. 1.** The self-attention mechanism allows the model to weigh the importance of different words in a sentence when encoding its meaning. The Transformer architecture is renowned for its effectiveness in capturing long-range dependencies and contextual information in sequences, and BERT uses the encoder part of this architecture. In addition to using a bi-directional transformer that takes the context on the left as well as the right while training, BERT also applies the concepts of Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) for pre-training, which make it better than the models before it. Moreover, it can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of natural language processing tasks (Devlin et al., 2018). The details of its architecture are shown in **Table 1.,** as adopted from Devlin et al. (2018)'s paper.

*RoBERTa***.** Introduced by Liu et al. (2019), RoBERTa was an attempt to alleviate the need to extensively fine tune the model, which required resources and time. It rather focuses on optimizing the model based on careful evaluation of the effects of hyperparameter tuning and training set size. In addition to being trained on a larger corpus of data, RoBERTa differs from BERT in the fact that it changes the masking strategy from static to dynamic, as well as removes the NSP objective. RoBERTa was considered as the best performing model compared to BERT, DistilBERT, and XLNet in classifying sentiments (Mathew and Bindu, 2021). Another study conducted to compare these models for the task of emotion detection resulted in RoBERTa achieving the best results (Adoma et al., 2020). The details of the model are shown in **Table 1**.

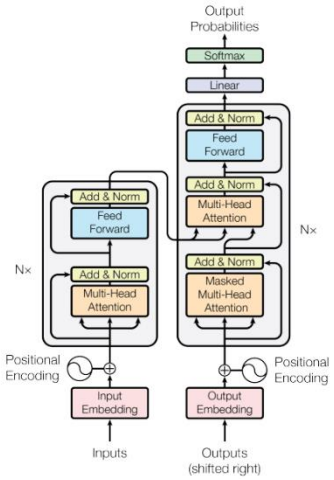


Fig. 1. The Transformer Architecture (Vaswani et al., 2017)

*DistilBERT***.** In the wake of limited computational resources, DistilBert was introduced by Sanh et al. (2019), that reduced the size of the original model by 40%, while retaining 97% of its language understanding capabilities, making it 60% faster through knowledge distillation. In addition, Sanh et al. (2019) claimed that DistilBERT is only 0.6% behind the previously launched BERT in terms of test accuracy, while being considerably smaller. The details of the model are shown in **Table 1.**

OpenAI GPT-2. A little while after BERT was launched, OpenAI came up with their version of a pre-trained model called GPT-2. Although it inherits most features from the same transformer architecture, the key features distinguishing it from other models are its sheer size (1.5B Parameters), its ability to generate text, and the fact that it is a decoder-only transformer (Radford et al., 2019). A downside of it is that unlike BERT, which is bi-directional, GPT-2 is unidirectional, meaning that it can only consider the context to its left side. The details of the model architecture are summed up in **Table 1.**

Table 1. Overview of the architecture of the pretrained models

Model	Number of Layers	Hidden Layer Size	No of self-attention heads	Total Parameters
BERT-Base	12	768	12	110M
RoBERTa-Base	12	768	12	125M
DistilBERT	6	768	12	66M
GPT-2	12	768	12	117M

Tokenizer. Tokenizer extract features from the text that can be fed to a pretrained model for fine-tuning.

BERT and DistilBERT Tokenizer. These tokenizers use the concept of WordPiece tokenization, whereby, the algorithm progressively learns a given number of merge rules

based on the most significant likelihood. It adds those tokens that increase the likelihood the most. The efficacy hinges on its adeptness at contextual comprehension, necessitating a meticulous structuring of input data. This entails the incorporation of specialized tokens, such as [CLS] (classification), denoting the beginning of the input sequence, and [SEP] (separator) marking the boundaries between sentences. This tokenization strategy is instrumental in facilitating the model's understanding of contextual nuances and sentence relationships. In scenarios where these models evaluate multiple sentence sets, the [SEP] token plays a pivotal role in delineating the boundaries between sentences. By adhering to this tokenization approach, we optimize the ability to process input text comprehensively, thereby, enhancing the capability to generate precise and contextually informed representations for subsequent tasks.

RoBERTa and GPT-2 Tokenizer. These tokenizers also work in a similar way except for the fact that they use byte-level Byte-Pair Encoding. This is a modern technique that can handle rare and out of vocabulary words, helping better handling of domain-specific language. As opposed to WordPiece, where the likelihood is used to merge characters, Byte-Pair Encoding iteratively uses the frequency of frequently occurring pairs to add to its current inventory. This technique not only handles rare vocabulary but also ensures that words with common root words are represented using common sub words to help the model better understand variations.

3.2 Experimental setup

Overall process. The overall process adopted in conducting experiments is illustrated in **Fig. 2**. The process initiated with data exploration and preprocessing that involved removal of missing words, feature selection, and rating recalibration from '1-5' to '0-4' to allow us to feed it into the models. Subsequently, tokenization converted the review text into input IDs to be used by the pretrained models for fine tuning. After the fine-tuning, we conducted an evaluation of the results that included a quantitative as well as qualitative analysis. Quantitative metrics used for evaluation are shown in **Fig.2**. Qualitative analysis included analyzing misclassified reviews individually by grouping the ratings '0' and '1' as negative, '2' as neutral, '3' and '4' as positive. By doing so, we analyzed major misclassifications amongst the models to identify potential issues that could relate to either the review text or possible biases in the models.

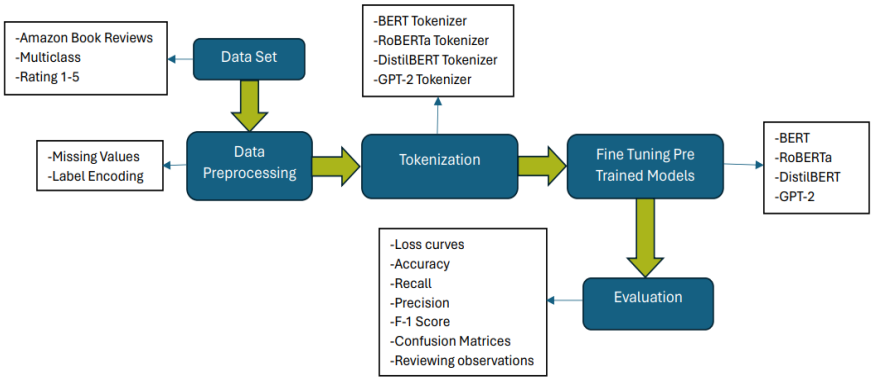


Fig. 2. Methodology Summarized

Data and exploratory data analysis. The original dataset consisted of a .json file that consisted of 17,79,608 observations. While selecting the sample size, we aimed at striking a balance between a sufficient data sample for a robust analysis while keeping it small enough to allow us for thorough examination of the individual test observations to facilitate our misclassification analysis. We experimented with sizes of 1000, 5000, and 10,000 observations, and found that 5000 observations were enough to achieve this objective. So, we parsed a sample of 5000 stratified observations. We further split this dataset in the ratio 80/20 to create our training and validation datasets. An extract of the dataset is shown in **Table 2**. The features of interest in our experiment were 'reviewText' and the rating that is labeled as 'overall'.

The distribution of the target variable across the data set is shown in **Fig. 3.** The figure shows a high class imbalance with 56% of observations lying within the rating 5. Similarly, rating '1' , '2', and '3' make up just 19% of the total dataset, indicating the high skew of the dataset.

Table 2. Extract of the dataset

reviewText	overall
"Was too complex, complicated for me to work. Keep it simple. Costs more than you can realize in a return."	1
"Interesting, fun, good light read. I really enjoyed it and will be getting more from this author in the future. I recommend for a rainy day."	3
"This series is not as good as some of her others but this book is by far the better one in the series."	4

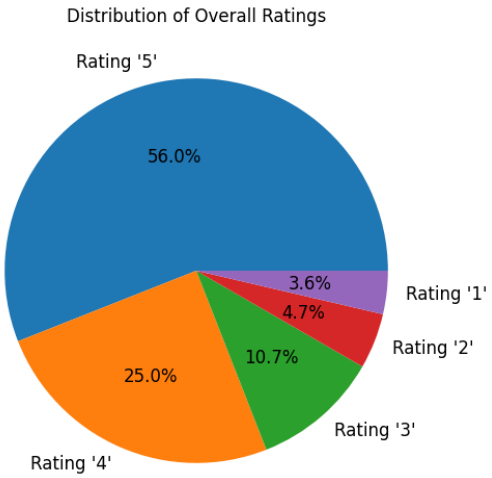


Fig. 3. Distribution of ratings across the full data set

Data preprocessing

Handling missing values. In the resulting dataset of 5000, there were 3 observations that were missing review text. We had to remove these to make the dataset compatible with PyTorch.

Feature selection and label encoding. To prepare data that was compatible with Hugging Face library, we kept the features named 'Review Text' and 'Overall', and created a 'Datasetdict' object containing the training and validation split. Moreover, the ratings had to be rescaled from 1-5 to 0-4 as illustrated in **Table 3**.

Table 3. Rescaled Ratings

Rating	Meaning	Rescaled Rating
1	Very Bad	0
2	Bad	1
3	Average	2
4	Good	3
5	Very Good	4

Tokenization. In order for the text to be converted into inputs that could be fed to the model, we used the respective pretrained tokenizer for each of the models including BERT Tokenizer, RoBERTa Tokenizer, DistilBERT Tokenizer, and GPT-2 Tokenizer.

Hyperparameters used. Hyperparameters play a pivotal role in determining a model's performance, convergence speed, and generalization ability.

Learning rate. Studies have consistently chosen a learning rate of 2e-5 for fine-tuning these models and demonstrated its superiority to other learning rates, for example, 4e-4, that can lead to training instability and a model's failure to converge effectively (Sun et al., 2019; Tang et al., 2020). So, we chose a learning rate of 2e-5 for our experiments.

Batch size. A batch size of 8 was used for fine-tuning BERT, RoBERTa, and DistilBERT models. For GPT-2, a batch size of 3 was chosen based on the maximum memory available. According to a research, small batch training has been recommended as it facilitates more up-to-date gradient calculations, resulting in training that is more stable and reliable (Masters and Luschi, 2018). This approach offers the potential for improved generalization performance when compared to using larger batches.

Number of epochs. The selection of the number of epochs involves a balance between ensuring adequate model learning and preventing overfitting. In this study, 10 epochs were chosen to fine-tune the models. This decision aligns with the understanding that transformer models often converge relatively quickly during fine-tuning, and few epochs are usually sufficient to achieve high performance as conveyed by Devlin et al., (2018) recommending 3-4 epochs. We chose 10 epochs as we aimed at capturing convergence behavior and overfitting as well.

Optimizer. AdamW optimizer, an extension of the Adam optimizer, incorporates weight decay directly into its optimization algorithm to improve model stability and prevent overfitting (Loshchilov & Hutter, 2017). Widely regarded as one of the most popular optimizers in literature, AdamW has been extensively used for natural language processing tasks (Durairaj et al., 2021; Ali et al., 2024). In a study by Sanjana et al. (2021), which explored various optimization methods for language models, AdamW emerged as the top performer for a multiclass classification task. Their findings underscored the efficacy of AdamW in achieving high accuracies, which made it a suitable choice for optimizing our pre-trained models.

The hyperparameters employed in our study are summarized in **Table 4.**

Table 4. Hyperparameters used for fine tuning.

Model	Epochs	Learning rate	Training Batch Size	Optimizer
BERT	10	2e-5	8	AdamW
RoBERTa	10	2e-5	8	AdamW
DistilBERT	10	2e-5	8	AdamW
GPT`-2	10	2e-5	3	AdamW

Fine tuning. The next step in our experiment was fine tuning the four models on our dataset. Experiments were conducted using Google Colab with GPU support. The transformer models were implemented using Keras and PyTorch. Training and evaluation of the models was done through the Google Colab cloud environment. We also used generative AI tool "ChatGPT 3.5" to assist us in fixing codes wherever necessary.

Evaluation metrics. The evaluation process involved assessing each model's performance on a separate validation set after each training epoch for continuous monitoring of its progress and convergence over the course of training. Metrics such as loss, accuracy, precision, recall, and F1 score, calculated from the multi label confusion matrices of actual versus predicted rating, were tracked to evaluate the models. Since we dealt with multiclass classification, we used the weighted measures for precision, recall, and F1 that consider the total number of instances in a certain rating. The evaluation metrics are described below.

Accuracy refers to the number of correctly predicted instances among the total number of instances and is formulated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{(Grandini et al., 2020)}$$

Precision refers to the proportion of true positives to the sum of true and false positives. It is formulated as:

$$Precision = \frac{TP}{TP+FP} \qquad \text{(Grandini et al., 2020)}$$

Recall is the fraction of positive labels correctly identified by the model as is formulated as:

$$Recall = \frac{TP}{TP+FN} \qquad \text{(Grandini et al., 2020)}$$

F1-Score is the harmonic mean of precision and recall and is calculated as:

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \qquad \text{(Opitz \& Burst, 2021)}$$

Next, we calculated the individual F-1 scores per rating for each model. We relied more on F-1 as accuracy is not well-suited in multiclass classification problems, especially when the datasets are imbalanced. F1-Score is a more robust measure in this case, as it effectively balances both recall and precision measures (Grandini et al., 2020; Noori, 2021). The confusion matrices for each model were also used to dig deeper into the model performances.

Creating sentiment polarities. Keeping in view the fact that from a business standpoint, misclassifications within one notch difference (e.g. between ratings '3' and '4') may not be as problematic as ones within higher notch differences (e.g. between ratings '0' and '4'), the next step was to consolidate the labels '0' and '1' into a negative polarity, '2' into a neutral polarity, and '3' and '4' into a positive polarity. Through this approach, we aimed to simplify the rating system and prioritize misclassifications with significant impacts, such as those transitioning from negative to neutral to positive sentiments. This time, we calculated the individual F-1 scores per polarity for each model from the

new confusion matrices to highlight the number of reviews classified into each sentiment polarity.

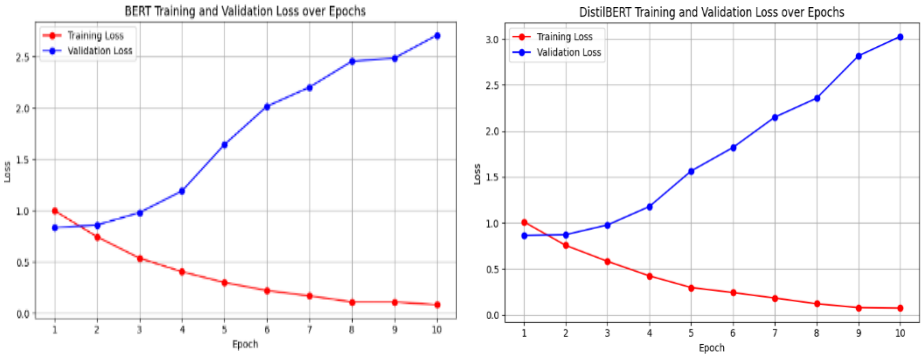
Misclassification analysis. To deepen our understanding into the misclassifications made by the models, our next step involved conducting an in-depth analysis of the model predictions. We closely examined and assessed the textual properties of the inaccurately predicted reviews, with the primary objective of understanding the rationale behind these predictions and uncovering potential issues related to the dataset or the models. We used the sentiment polarities instead of individual ratings for this analysis to put greater emphasis on significant misclassifications. We aimed to identify and analyze misclassifications that were common across all models as well as those specific to each model. For the latter, we compared the predictions of each model to see how they classified the same review.

4 Results and Discussion

4.1 Model fine-tuning

The fine tuning of the four transformer-based models involved monitoring training loss, validation loss, and accuracy metrics across the epochs. **Fig. 4.** displays the loss graphs for each model. A consistent decrease in training loss is observed for all models, indicating effective learning from the training data. However, the validation loss consistently increases as training progresses. This pattern suggests that these models can achieve the desired performance in a small number of epochs and can start to overfit rapidly.

Regarding accuracy, each model experienced fluctuations across epochs, with peak accuracies reached at different numbers of epochs. BERT achieves its highest accuracy of 66.0% at epoch 2, RoBERTa peaks at epoch 6 with an accuracy of 67.2%, DistilBERT reaches 64.0% at epoch 2, and GPT-2 attains a 65.6% accuracy at epoch 2. This confirms that when fine tuning pre-trained models, a lower number of epochs is sufficient to achieve maximum accuracy.



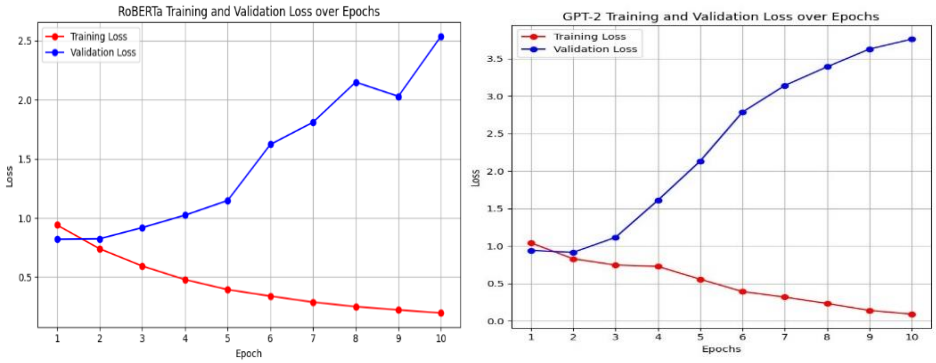


Fig. 4. Loss Curves over epochs

4.2 Model performance based on evaluation metrics

The evaluation of the models based on the chosen metrics is summarized in **Table 5.**

Table 5. Evaluation metrics for fine-tuned models

Model	Accuracy	F1 Score	Recall	Precision
BERT	0.667	0.646	0.667	0.644
RoBERTa	**0.670**	**0.665**	**0.670**	**0.666**
DistilBERT	**0.639**	**0.628**	**0.639**	**0.624**
GPT-2	0.656	0.638	0.656	0.658

Even though there does not seem to be much deviation across the results between the different models, the results show that RoBERTa outperforms other models across all evaluation metrics. Conversely, DistilBERT yields lower values for each metric; however, the result is not far off from RoBERTa.

Table 6. presents the individual F-1 scores per rating for each model.

Table 6. F-1 measure per rating.

Model	Rating '0'	Rating '1'	Rating '2'	Rating '3'	Rating '4'
BERT	0.424	0.409	0.448	0.441	0.824
RoBERTa	0.552	0.351	0.412	**0.514**	**0.827**
DistilBERT	**0.578**	**0.447**	0.332	0.461	0.790
GPT-2	0.481	0.154	**0.507**	0.452	0.807

For the ratings '0' and '1', DistilBERT has the highest F-1 scores, for '2', GPT-2 scored the highest, and, for '3' and '4', RoBERTa is the best performing model in terms of F-1 score.

Fig. 5. shows the multi-label confusion matrices for each model.

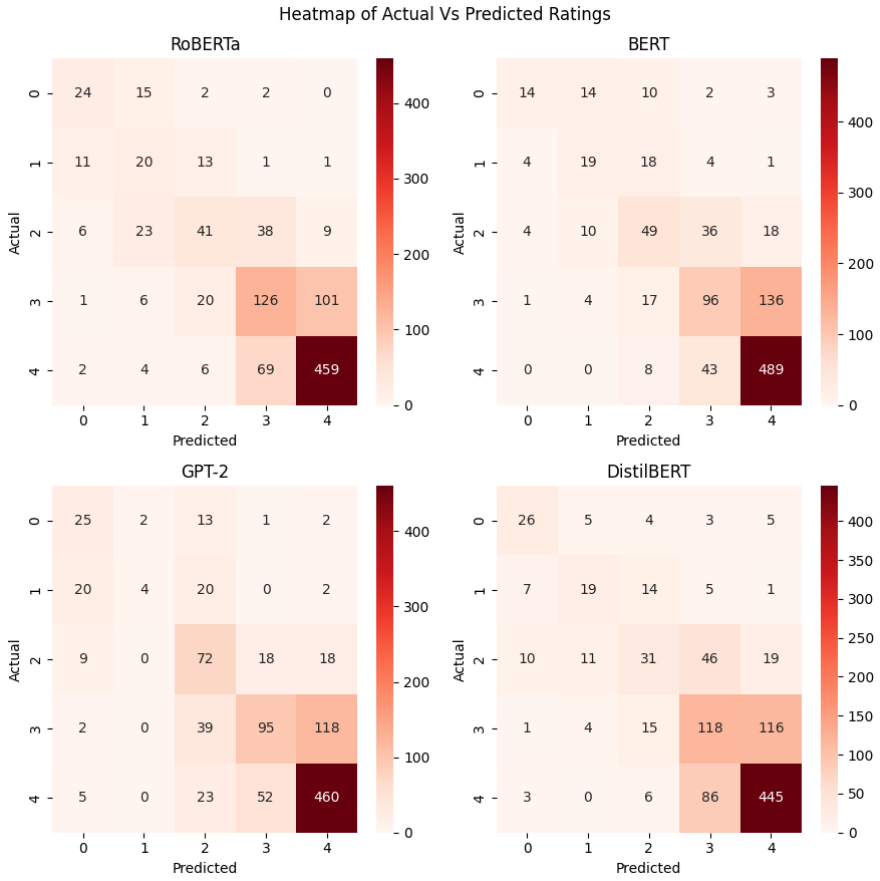


Fig. 5. Confusion matrices of actual vs predicted rating.

Upon analyzing the confusion matrices, a consistent pattern of behavior emerges. Across all models, there is a recurring trend of misclassifications occurring primarily between adjacent sentiment scores. The models exhibit challenges in accurately distinguishing between fine-grained sentiment categories, such as "good" (rating ‘3’) and "very good" (rating ‘4’).

Analyzing these matrices together with the F1-measures for all models reveals interesting insights. Even though the number of correctly classified ‘0’ and ‘1’ rating reviews is almost similar for DistilBERT and RoBERTa, nevertheless, DistilBERT demonstrates a superior F-1 score. Upon closer examination of misclassifications, it appears that in cases where RoBERTa misclassifies a review with a rating ‘0’ or ‘1’, it makes a smaller error as compared to DistilBERT. In these cases, RoBERTa tends to preserve the negative sentiment, with its misclassifications lying between the ratings ‘0’ and ‘1’ (i.e., predicting negative reviews as negative). Conversely,

DistilBERT's misclassifications for the ratings ‘0’ and ‘1’ tend to occur more frequently in higher scores (‘3’ or ‘4’), indicating its higher propensity to erroneously label negative reviews as positive. To sum up, when DistilBERT makes an error, it is larger than that of RoBERTa. Therefore, it is prudent to interpret the F1 scores cautiously and understand that DistilBERT displaying higher F1-scores for negative ratings does not necessarily indicate its superiority in classifying negative reviews.

Sentiment polarities. The individual F-1 scores per polarity for each model, as well as the confusion matrices are shown in **Table 7**. and **Fig. 6.** respectively.

Table 7. F-1 Measure for aggregated ratings.

Model	Negative	Neutral	Positive
BERT	0.642	0.448	0.942
RoBERTa	**0.697**	0.412	**0.944**
DistilBERT	0.651	0.332	0.934
GPT-2	0.654	**0.507**	0.930

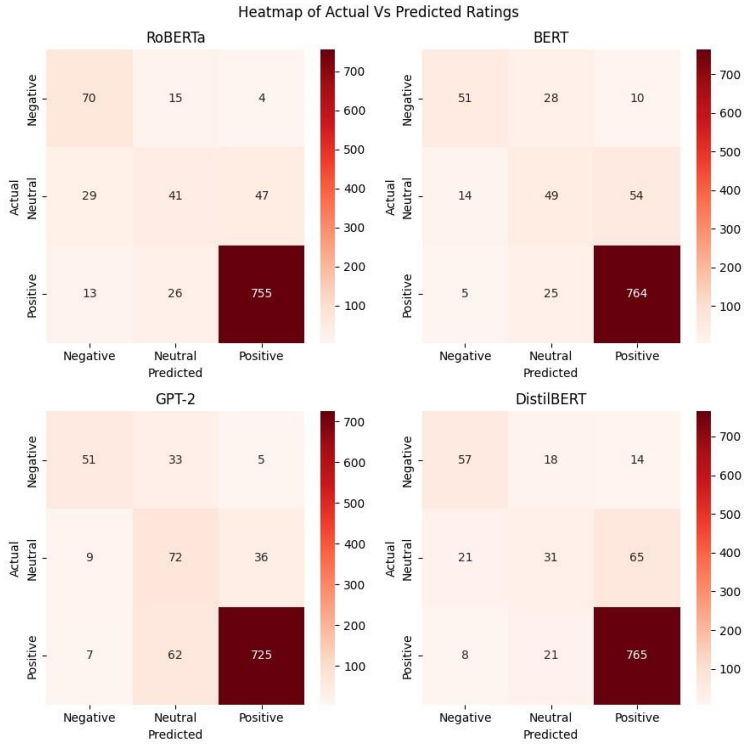


Fig. 6. Confusion matrices for actual vs predicted polarity.

According to **Table 7**., RoBERTa outperforms other models on the negative and positive polarities in terms of F-1, while GPT-2 outperforms all models in classifying neutral reviews. It is interesting to note that the F-1 scores for the positive polarity are

consistently high (above 0.9) for all models, while they are the lowest for the neutral polarity (between 0.3 and 0.5). This indicates the weaker performance of models in classifying the neutral sentiment. The high F-1 scores for the positive polarity can also be attributed to the skewness of the training dataset.

As seen from the confusion matrices in **Fig.6,** when we define negative reviews as an aggregate of both '0' and '1' ratings, RoBERTa outperforms DistilBERT in terms of the number of correctly classified negative reviews, since most of its misclassifications had been within the '0' and '1' ratings. Moreover, it misclassifies the highest number of positive and neutral reviews as negative. This denotes the existence of a possible bias of RoBERTa towards the negative sentiment.

When we define positive reviews as an aggregate of both '3' and '4' ratings, it can be seen from **Fig.6** that BERT and DistilBERT outperform RoBERTa in terms of the number of correctly classified positive reviews. Both models also misclassify a higher number of neutrals and negatives to positives. This indicates a potential inclination of BERT and DistilBERT towards positive sentiments.

4.3 Misclassification analysis

Common misclassifications. A total of 48 instances were identified where all models misclassified the actual polarity of that text into another polarity. Upon manually analyzing all 48 of these reviews, we concluded that these errors fall into three overarching categories, each with its distinct characteristics and prevalence within the sample, as summarized in **Table 8**.

Table 8. Common misclassification categories

Misclassification Category	Frequency	Percentage (%)
User Errors	22	46
Subjectivity of Sentiment	15	31
Contextual Complexities	11	23

User (reviewer) error. The misclassified reviews attributed to user errors are ones for which the rating clearly does not match the sentiment expressed in the review. In these cases, the model effectively predicts the sentiment from the text. These errors arise when reviewers assign a significantly lower rating than the tone of the review suggests, or conversely, an inflated rating that is disproportionately high. The mismatch between the review tone and the rating can also be attributed to mislabeling of the dataset. **Table 9.** illustrates the most representative examples of such reviews. Additional examples are included in **appendix 1**.

Table 9. Common misclassifications due to user errors

Review text	Actual polarity	Predicted polarity	Possible Explanation
"I think I wasn't in the mood to read it; I'll try again later. Darlene, she is a great lady."	Negative	Neutral	Neutral tone of the text which is not indicative of any negative sentiment is predicted by the models.
"Great job with your book. I love it!!!I am a new fan!!!!I would recommend it to my teacher Mrs. Kushner!"	Neutral	Positive	Purely positive sentiment with words and phrases such as 'great job', "love", "fan", and "recommend" aligns with the models' prediction.
"Big disappointment.... I knew it was coming, I just didn't want to believe it. I have a lot to say but I don't know where to start without spoiling it. Grrrr...."	Neutral	Negative	Negative sentiment with the phrase "big disappointment" is detected rightfully by the models.
"I had such high hopes for Jeremy's story, and I felt it was such a disappointment.…Kinda boring. Wasn't really thrilled with Danny either."	Positive	Negative	Negative sentiment with use of the words "disappointment", and "boring" is detected rightfully by the models.

Subjective interpretation. The reviews attributed to this category contain text that is highly ambiguous and subjective in nature, often lacking depth in explaining the assigned rating. A discrepancy between the review text and its corresponding rating is observed once again, which complicates the model's interpretation. In these cases, both the user's and the model's rating appear appropriate, and their interpretation can be subjective, varying from person to person. **Table 10.** illustrates a few examples of such reviews. Additional examples are found in **appendix 2**.

Table 10. Common misclassifications due to subjectivity of reviews

Review text	Actual polarity	Predicted polarity	Possible Explanation
"Kind of a slow read with an ending I didn't quite expect. I will be reading the next book just to see what happens."	Positive	Neutral	The text lacks clarity, making it difficult to determine whether it conveys a positive, neutral, or negative sentiment.
"It was just ok. I loved possessive and alpha male Josh. However, there were	Negative	Neutral	Some may perceive it as negative due to the mention of 'just ok' and 'boring filler'; others might view it as neutral, particularly those

just too many pages of boring filler."	who appreciate elements like the character Josh being described as possessive and alpha male.

Contextual complexities. These reviews mostly express initial admiration for the author's previous works or the book's overall reputation, setting a tone that conflicts with the subsequent criticism of the book itself. Additionally, lengthy reviews focusing heavily on the storyline, often include positive or negative words unrelated to the reviewer's actual opinion on the book, setting a tone contraindicated by the real sentiment. The models struggle to grasp the real context of these reviews, leading to misclassifications. Similarly, a few reviews focusing solely on one negative aspect of the book without balancing it with positive points, are misclassified as negative. **Table 11** illustrates examples highlighting each of these issues, with additional examples in **appendix 3**.

Table 11. Common misclassifications due to contextual complexities

Review text	Actual polarity	Predicted polarity	Possible Explanation
"If you love the details of forensics work, you'll like this book. I thought it was a little heavy on the nitty-gritty of the investigation and a little light on story and plot, but if you're a CSI fan, you will probably enjoy it."	Neutral	Positive	The review focuses on potential reader enjoyment rather than the reviewer's personal feelings. Despite criticism of heavy forensic detail and light plot, the model predicts it as positive.
"I find it amazing that a control freak who is also a stalker and a pervert could be anyone's dream man, but apparently, it's possible. The author made Christian Grey likable by making him truly decent and loving. Oh, and obviously, he is a brilliant Harvard drop-out. There's some obsession in our society with dysfunctional personalities, which I do not seem to share."	Negative	Positive	Due to words like "amazing", "dream man", "likable", "loving", "brilliant", an "obsession", the model classifies the text as positive, failing to grasp the underlying negative context.
"The only problem is the time issues compared with the end of the other novels. No real explanation of how they ended up when and where the book is set."	Positive	Negative	The phrase, "the only problem", at the start suggests a single issue amidst an otherwise positive review but sets an initial negative tone.
"I'm actually glad that the book didn't go on any longer. This was my first and last sojourn into the vampire camp. I'm glad I didn't spend more or have to read longer."	Negative	Neutral	The models fail to grasp the ironic negative context in the text.

Misclassifications specific to each model. Upon analyzing the reviews misclassified by each model individually, it was revealed that RoBERTa, BERT, and DistilBERT exhibit challenges in interpreting mixed sentiments, and differ in the weights they assign to positive or negative sentiments within the same text. As opposed to these models, GPT-2 is more biased towards neutral sentiment. The misclassifications specific to each model are discussed below.

RoBERTa. As seen from **Fig.6**, RoBERTa misclassifies the highest number of neutral and positive reviews (43) as negative. We ignored the common misclassifications and observed that in most of the remaining cases, RoBERTa was the only model that was misclassifying reviews as negative. Thus, we can conclude that when faced with a mixed sentiment review, RoBERTa tends to give more weight to negative words in the text than positive words. **Table 12** shows a sample of these reviews, with additional examples in **appendix 4**.

Table 12. Misclassified reviews for RoBERTa

Review text	Actual polarity	Predicted polarity	Possible Explanation
"I didn't know how I would feel about this book when I first started reading it, but I'm glad I stuck with it. EJ is a damaged woman, damaged by her mother who was too busy with her career and herself to be a mother. I did not give this book 5 stars because 1) the slow start, and more importantly 2) the inserts of Lilly's (EJ's mother) writing which, although I got used to it and found it mildly interesting, was totally superfluous to the story."	Positive	Negative	A mix of positive and negative words such as 'glad', 'damaged', 'slow', 'mildly interesting', and 'superfluous' exists where RoBERTa emphasized more on the negative words.
"I liked this book, but it was not that funny like I don't get the purpose of it. It isn't Jeff Kinney's best book he ever wrote"	Neutral	Negative	The model has assigned a greater weightage to the negative connotation for this review than to the positive.

BERT and DistilBERT. As seen earlier from **Fig.6**, BERT and DistilBERT misclassify the highest number of neutral and negative reviews as positive. Ignoring the common misclassifications, we observed that BERT and DistilBERT were the only models misclassifying them in most of these cases. After careful evaluation of these reviews, we conclude that when faced with a mixed sentiment review, BERT and DistilBERT tend to give more weight to positive words in a review text. Despite the presence of criticisms or negative aspects within the reviews, the model often assigned higher scores words indicating positive sentiment. **Table 13** shows a sample of these reviews which

have been inaccurately classified by BERT and DistilBERT uniquely, with additional examples in **appendix 5**.

Table 13. Examples of misclassified reviews for BERT and DistilBERT

Review text	Actual polarity	Predicted polarity	Possible explanation
"I'm a big fan of the Horus Heresy novels and have been very impressed with the story writing and stories which have been released. And I think it is very fun that the Black Library is having many authors each write a book (or two). Mitchel Scanlon wrote 15 hours which, in my mind, was a very good book. Descent of Angels is not. For those of you who are fans of the 40k fluff, it is, of course, a must read. And it is not a "bad book". It just doesn't really live up to the other books written and seems rather incomplete."	Neutral	Positive	The presence of positive expressions, particularly in relation to the broader series and the author's previous work, may lead the model to assign a higher score despite the mixed sentiments expressed in the review.
"I purchased this along with many other Disney books and found it a huge disappointment. It is filled with beautiful color pictures which is nice but the information is all extremely simple and not helpful at all."	Negative	Positive	The overall disappointment is not captured because of high weightage to words like "beautiful" "nice", "extremely simple".

GPT-2. As seen from **Fig.6**, GPT-2 misclassifies the highest number of negative and positive reviews (95) as neutral. We ignored the common misclassifications and observed that in most of the remaining cases, GPT-2 was the only model making this misclassification. When investigated, it became clear that when faced with a review with mixed sentiment, discussing both positives and negatives, GPT-2 has a strong tendency to classify that review as neutral. **Table 14** shows a few such examples where GPT-2 was the only model misclassifying the review and more examples can be found in **appendix 6**.

Table 14. Examples of misclassified reviews by GPT-2

Review text	Actual polarity	Predicted polarity	Possible explanation
"I really wanted to give this book 5 stars but I couldn't because the way it ended was too abrupt. I loved the book though it was really a great story there where some grammar errors but not so many that it distracts from the book."	Positive	Neutral	While this is an overall positive review with some negative feedback, GPT-2 classified it as neutral.

"I read about 75% of the book but then decided to put it away. Contains some interesting insights but for a layman difficult to understand the differences between different types of hallucinations"	Negative	Neutral	While this is an overall negative review with some positive sentiment, GPT classified it as neutral.

5 Critical Reflection and Future Scope

Our study represents a step forward from merely evaluating pre-trained models based on standard metrics to an in-depth analysis of the model predictions. In terms of the evaluation metrics, our maximum achieved accuracy stands at 67% with the RoBERTa model, amidst challenges such as class imbalance and a lack of hyperparameter optimization. Due to a lack of dedicated GPU support, we could not set up an extensive grid search to optimize hyperparameters such as learning rates or batch sizes while fine-tuning our models. Instead, we referred to existing research and used the hyperparameters that have been recommended frequently. Even though we tried methods such as incorporating class weights and extracting a uniformly distributed sample of data, these approaches proved to be ineffective in addressing the issue of class imbalance in our dataset. The dataset was heavily skewed towards the rating '4', which could have potentially compromised our models' performances and resulted in a significant skew of the predicted ratings towards the positive end. We used robust evaluation metrics such as F1, which are well-suited for imbalanced datasets.

While the results of the evaluation metrics show minimal deviation across different models, our detailed misclassification analysis uncovers intriguing hidden biases that set them apart. Moreover, we see that while metrics like F-1 score provide valuable insights, they do not fully capture the intricacies of model performance, especially in a multiclass sentiment classification context. Our approach mainly focuses on understanding the magnitude of errors made by the models in predicting each rating and uncovering issues with different models in predicting certain ratings. Consolidating the ratings into polarities helped us disregard the misclassifications within one notch difference and focus on the ones with more impact.

As seen from our results, each of our models struggles with three common categories; user errors, subjectivity of reviews, and contextual complexities. It was interesting to see that approximately 50% of the commonly misclassified reviews could be attributed to user errors, whereby, the corresponding rating did not match the review.

The prevalence of such problems within the dataset reveals how important it is for companies like Amazon to understand that user ratings may be either inflated or deflated. It emphasizes the models' proficiency in predicting the correct rating. This insight holds significance for businesses seeking to ascertain the genuine sentiment and perceptions of their products. By discerning the true rating from sentiment analysis,

they can gain clarity on customer views and make informed decisions accordingly. Moreover, errors in labeling reviews can significantly impact model performance, underscoring the importance of precise and consistent annotation practices, and the periodic review and refinement of annotations.

Such an analysis is not without its set of challenges and limitations. A challenge we faced in our misclassifications analysis was that some misclassified reviews exhibited a blend of common errors, including mixed sentiments, subjectivity of sentiment, and contextual understanding. This overlapping made it difficult to analyze and categorize them into particular reasons. Additionally, we encountered cases with unique challenges and unspecified issues not fitting neatly into the categories we defined.

In the future, researchers can investigate optimizing hyperparameters and recommend the most effective pre-trained transformer model for a multi-class classification of Amazon reviews. Future researchers can also investigate modern methods to handle class imbalance like data augmentation and leverage large language models (LLMs) to generate synthetic data. This could be a path to find out ways that could improve results when fine-tuning models on a skewed dataset. Another avenue for future research lies in investigating the underlying motives presented by customers behind rating a certain review, which is extremely important for businesses like Amazon. Moreover, the analysis can be expanded to include other popular models like XLNET and GPT-3 to uncover how they perform on different sentiments.

6 Conclusion

In the modern era, where e-commerce is thriving and generating huge amounts of unstructured textual data in the form of user reviews, it is essential for businesses to be able to extract and analyze this data on the sentiment that these reviews convey. Particularly, Amazon generates millions of reviews for various products including books that it sells online. In this paper, we set out to leverage modern transfer learning techniques by fine-tuning four pretrained transformer models namely BERT, RoBERTa, DistilBERT, and GPT-2 on an Amazon book review dataset to classify reviews on a scale from 1-5.

To achieve this, we preprocessed the data by removing missing values and rescaling ratings on a scale from 0-4. This was followed by tokenization by the respective tokenizer of each model and lastly, fine-tuning of the models based on carefully chosen hyperparameters. We set out to conduct a meticulous analysis of the results including not only a quantitative analysis but also a qualitative analysis where we delved into the results to identify possible issues relating to the text or the models.

Based on the chosen metrics, RoBERTa was the best performing model while DistilBERT lagged behind. We then looked at the F-1 scores per rating for each model. It was interesting to see that even though DistilBERT was performing the best for the

ratings ‘0’ and ‘1’ based on this metric, whenever it made an error in these ratings, it had a large impact. The F-1 score also showed that GPT-2 was the best performing model for the rating ‘2’.

To conduct an analysis of the misclassifications and draw interesting insights, we grouped the ratings ‘0’ and ‘1’ as negative, ‘2’ as neutral and ‘3’ as positive to analyze misclassifications from one polarity to the other. Next, we reviewed the common misclassifications and concluded that approximately 50% of these could be attributed to user errors; whereby, the review would not correspond to the given rating. Another reason included subjective interpretation that accounted for approximately 30% of the common misclassifications. Here, the review was ambiguous and lacked clarity due to which the models had a hard time interpreting it. Yet, another issue identified involved reviews with contextual complexities where reviewer delved deep into the theme of the book discussing story lines, which made it difficult for the models to identify the true sentiment expressed. Such reviews accounted for 22% of the total cases.

On reviewing unique misclassifications for the models, we concluded that when faced with reviews containing mixed sentiments (both positive and negative), RoBERTa exhibited a bias towards giving more weight to negative words and phrases, thus, misclassifying more reviews as negative. Conversely, BERT and DistilBERT were biased towards giving a higher weightage to positive words and phrases, misclassifying a large number of reviews as positive. On reviewing GPT-2, it was interesting to see that when faced with a mixed sentiment review, GPT-2 had a tendency to classify it as neutral, regardless of the overall sentiment that the review carried.

While this research gives enough insight into the performance of these models and certain biases they exhibit, future research can focus on optimizing the model performance by setting up a grid search to optimize hyperparameters such as learning rate, batch size, and number of epochs, and investigate if it improves performance and removes these biases. Moreover, since the dataset is highly skewed, there is a lot of scope for future research to find out ways that could address this issue to improve model performances.

References

Adoma, A. F., Henry, N. M., & Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. *International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp.117-121). IEEE. https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379.

Ali, H., Hashmi, E., Yayilgan Yildirim, S., & Shaikh, S. (2024). Analyzing amazon products sentiment: a comparative study of machine and deep learning, and transformer-based techniques. *Electronics*, *13*(7), 1305.

Alves, M. D., Lobo, A. G., & Reis, A. M. (2022). Assessing the use of pre-trained transformers to classify customer reviews. *5th International Conference on Quality Engineering and Management: A Better World with Quality* (pp. 329-337).

Aßenmacher, M., & Heumann, C. (2020). On the comparability of pre-trained language models. arXiv preprint arXiv:2001.00781. https://doi.org/10.48550/ARXIV.2001.00781

Bello, A., Ng, S. C., & Leung, M. F. (2023). A BERT framework to sentiment analysis of tweets. *Sensors*, *23*(1), 506. https://doi.org/10.3390/s23010506

Bhavitha, B. K., Rodrigues, A. P., & Chiplunkar, N. N. (2017). Comparative study of machine learning techniques in sentimental analysis. *2017 International conference on inventive communication and computational technologies (ICICCT)* (pp. 216-221). IEEE. https://doi.org/10.1109/ICICCT.2017.7975191

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146. https://doi.org/10.1162/tacl_a_00051

Chauhan, U. A., Afzal, M. T., Shahid, A., Abdar, M., Basiri, M. E., & Zhou, X. (2020). A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews. *World Wide Web*, *23*, 1811-1829.

Chen, P., Dhanasobhon, S., & Smith, M. D. (2008). All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon. com. *Social Science Research Network*. https://doi.org/10.2139/ssrn.918083

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, *43*(3), 345-354. https://doi.org/10.1509/jmkr.43.3.345

Choi, G., Oh, S., & Kim, H. (2020). Improving document-level sentiment classification using importance of sentences. *Entropy*, *22*(12), 1336. https://doi.org/10.3390/e22121336

Clemons, E. K., Gao, G. G., & Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of management information systems*, *23*(2), 149-171. https://doi.org/10.1109/HICSS.2006.534

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/ARXIV.1810.04805

Durairaj, A. K., & Chinnalagu, A. (2021). Transformer based contextual model for sentiment analysis of customer reviews: A fine-tuned bert. *International Journal of Advanced Computer Science and Applications*, *12*(11). http://dx.doi.org/10.14569/IJACSA.2021.0121153

Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, *9*, 201–208.

Gadri, S., Chabira, S., Ould Mehieddine, S., & Herizi, K. (2022). Sentiment analysis: developing an efficient model based on machine learning and deep learning approaches. In *Intelligent Computing & Optimization* (Vol. 371, pp. 237–247). Springer International Publishing. https://doi.org/10.1007/978-3-030-93247-3_24

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing science*, *23*(4), 545-560. https://doi.org/10.1287/mksc.1040.0071

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*. https://doi.org/10.48550/ARXIV.2008.05756

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd nordic conference on computational linguistics* (pp. 187-196).

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, *52*(10), 144-147. https://doi.org/10.1145/1562764.1562800

Ishaq, A., Umer, M., Mushtaq, M. F., Medaglia, C., Siddiqui, H. U. R., Mehmood, A., & Choi, G. S. (2021). Extensive hotel reviews classification using long short term memory. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 9375-9385.

Kim, H., & Jeong, Y. S. (2019). Sentiment classification using convolutional neural networks. *Applied Sciences*, *9*(11), 2347.

Kohli, R., Devaraj, S., & Mahmood, M. A. (2004). Understanding determinants of online consumer satisfaction: A decision process perspective. *Journal of Management Information Systems*, *21*(1), 115-136. https://doi.org/10.1080/07421222.2004.11045796

Kokab, S. T., Asghar, S., & Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, *14*, 100157.

Kotei, E., & Thirunavukarasu, R. (2023). A systematic review of transformer-based pre-trained language models through self-supervised learning. *Information*, *14*(3), 187.

Kumar, N., & Benbasat, I. (2006). Research note: the influence of recommendations and consumer reviews on evaluations of websites. *Information Systems Research*, *17*(4), 425-439. https://doi.org/10.1287/isre.1060.0107

Kusal, S., Patil, S., Gupta, A., Saple, H., Jaiswal, D., Deshpande, V., & Kotecha, K. (2024). Sentiment analysis of product reviews using deep learning and transformer models: A comparative study. In *Artificial Intelligence: Theory and Applications* (Vol. 843, pp. 183–204). Springer. https://doi.org/10.1007/978-981-99-8476-3_15

Lagrari, F. E., & Elkettani, Y. (2021). Traditional and deep learning approaches for sentiment analysis: A survey. *Advances in Science, Technology and Engineering Systems Journal*, *6*(4), 1-7. https://doi.org/10.25046/aj060501

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. https://doi.org/10.48550/arxiv.1711.05101

Masters, D., & Luschi, C. (2018). Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*. https://doi.org/10.48550/arxiv.1804.07612

Mathew, L., & Bindu, V. R. (2022). Efficient classification techniques in sentiment analysis using transformers. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 1* (pp. 849-862). Springer Singapore. https://doi.org/10.1007/978-981-16-2594-7_69

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department*, 5.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/ARXIV.1301.3781

Min, Z. (2019, March). Drugs reviews sentiment analysis using weakly supervised model. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)* (pp. 332-336). IEEE.

Mohbey, K. K. (2021, March). Sentiment analysis for product rating using a deep learning approach. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 121-126). IEEE.

Mullen, T., & Collier, N. (2004, July). Sentiment analysis using support vector machines with diverse information sources. *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 412-418).

Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, *1*(4), 7.

Number of active Amazon customer accounts worldwide from 1st quarter 2013 to 4th quarter 2021. (2016, January 28). Statista. https://www.statista.com/statistics/476196/number-of-active-amazon-customer-accounts-quarter/Top of Form.

Opitz, J., & Burst, S. (2019). Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*. https://doi.org/10.48550/arxiv.1911.03347

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*. https://doi.org/10.48550/arxiv.cs/0205070

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543). https://doi.org/10.3115/v1/D14-1162

Qorich, M., & El Ouazzani, R. (2023). Text sentiment classification of Amazon reviews using word embeddings and convolutional neural networks. *The Journal of Supercomputing*, *79*(10), 11029-11054.

Rajapakse, T. (2021, November 10). *General usage*. Simple Transformers. https://simpletransformers.ai/docs/usage/

Rathor, A. S., Agarwal, A., & Dimri, P. (2018). Comparative study of machine learning approaches for Amazon reviews. *Procedia computer science*, *132*, 1552-1561.

Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, *78*, 26597-26613.

Reisinger, Don. (2021, June 9). *Here's how much Amazon Prime customers spend per year*. Fortune. https://fortune.com/2017/10/18/amazon-prime-customer-spending/.

Sadhasivam, J., & Kalivaradhan, R. B. (2019). Sentiment analysis of Amazon products using ensemble machine learning algorithm. *International Journal of Mathematical, Engineering and Management Sciences*, *4*(2), 508.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. https://doi.org/10.48550/arxiv.1910.01108

Sanjana, R., Tandon, C., Bongale, P. J., Arpita, T. M., Palivela, H., & Nirmala, C. R. (2021). Comparative analysis of various language models on sentiment analysis for retail. In A. Tiwari, K. Ahuja, A. Yadav, J. C. Bansal, K. Deep, & A. K. Nagar (Eds.), *Soft Computing for Problem Solving* (Vol. 1392, pp. 725–739). Springer Singapore. https://doi.org/10.1007/978-981-16-2709-5_55

Silva Barbon, R., & Akabane, A. T. (2022). Towards transfer learning techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for automatic text classification from different languages: A case study. *Sensors*, *22*(21), 8184. https://doi.org/10.3390/s22218184.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification?. *Chinese Computational Linguistics*, *11856*, 194–206. https://doi.org/10.1007/978-3-030-32381-3_16

Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, *10*, 21517-21525. https://doi.org/10.1109/ACCESS.2022.3152828

Tang, T., Tang, X., & Yuan, T. (2020). Fine-tuning BERT for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, *8*, 193248-193256. doi.org/10.1109/ACCESS.2020.3030468

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*. https://doi.org/10.48550/ARXIV.1706.03762

Vijayaraghavan, S., & Basu, D. (2020). Sentiment analysis in drug reviews using supervised machine learning algorithms. *arXiv preprint arXiv:2003.11643*.

Wang, Y., Lu, X., & Tan, Y. (2018). Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. *Electronic Commerce Research and Applications*, *29*, 1-11. https://doi.org/10.1016/j.elerap.2018.03.003

Whalen, J. R. (Host). (2023, March 10). Airbnb Ratings: Do Five Stars Tell the Whole Story? [Audio podcast episode]. *Your Money Matters*. WSJ. https://www.wsj.com/podcasts/your-money-matters/airbnb-ratings-do-five-stars-tell-the-whole-story/85e5f599-8f62-4611-bbd1-e8ac371e087dTop of Form

Whalen, J.R. (Host). (2023). Airbnb Ratings: Do Five Stars Tell the Whole Story? [Audio podcast]. WSJ. https://www.wsj.com/podcasts/your-money-matters/airbnb-ratings-do-five-stars-tell-the-whole-story/85e5f599-8f62-4611-bbd1-e8ac371e087d

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*. https://doi.org/10.48550/arxiv.1609.08144

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

Zervas, G., Proserpio, D., & Byers, J. W. (2021). A first look at online reputation on Airbnb, where every stay is above average. *Marketing Letters*, *32*(1), 1-16. https://doi.org/10.1007/s11002-020-09546-4

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/widm.1253

Zhou, Y., & Srikumar, V. (2021). A closer look at how fine-tuning changes BERT. *arXiv preprint arXiv:2106.14282*. https://doi.org/10.48550/ARXIV.2106.14282

Appendix

Appendix 1. More examples of misclassifications due to user error.

Review text	Actual polarity	Predicted polarity
"This was an interesting book. I would say that there is something for everyone in this book. No doubt that reading this book will make a person more effective with time management and day to day productivity."	Neutral	Positive
"Interesting, fun, good light read. I really enjoyed it and will be getting more from this author in the future. I recommend for a rainy day."	Neutral	Positive
"This book took me over a month to finish. Why did I keep reading? Bored I guess. The story goes back and forth between two different people Hahp and Sadima. It gets you really mad with most of the characters expecially Somiss. Hardly anything possitive happens and you keep reading hoping that this depressing book will get better. I'm not always looking for happy endings but geesh.... This was bad."	Neutral	Negative
"Not my cup of tea given my personal taste. A really great storyline but just too slow and unromantic for my part. Definitely a good on for most other readers though so enjoy."	Negative	Neutral
"This was helpful and got my friend started playing the guitar - it gave my friend the basics to start."	Neutral	Positive

Appendix 2. More examples of misclassifications due to subjective reviews

Review text	Actual polarity	Predicted polarity
"This book is mainly concerned with reiterating old tried-and-true techniques to combat procrastination; new information is not presented. This work does serve as a reminder to follow certain techniques to increase your productivity."	Negative	Positive
"I am sorry to say this, because it seems like a work of love by those who created the book...but it really doesn't translate well into English! There is actually a better translation out there by Michael Saso (which is still somewhat confusing since that's the way it was originally written). That translation is entitled The Gold Pavilion.It's nice that this work includes the Chinese characters (the Gold Pavilion does not)...however, only one English translated word was used for each Chinese character, when actually there could be a multitude of subtle meanings. So I think in this version, the meaning is almost completely lost. Besides the translation	Negative	Neutral

itself, it seems that the authors have a hard time communicating clearly in English at all...you can see this for yourself by reading the free preview. It takes a lot of work on behalf of the reader.However, for the diehard Taoism scholar, maybe this book contains some hidden gems.”		

Appendix 3. More examples of misclassifications due to contextual complexities

Review text	Actual polarity	Predicted polarity
“While I didn't look forward to this as much as Brott's New Father, let alone Lamott's Operating Instructions, I did find parts of it really funny. And the best part is that it's got little quotes from fathers throughout, including the most anguished, troubled kinds of dilemmas, making me feel more human for my difficulty rising to the challenge sometimes. One Dad laments, "It all seems like a blur. I sometimes have to fight very hard to remember what he was like even one month or one week ago. Every time I turn around, he's doing something new. I don't want to miss anything. It all makes me wonder whether I've been around enough." But then beautiful moments too, "She screams with delight and babbles incessantly. Who knows what she's thinking." "My son's crawling. He comes bouncing up to me, smiling his head off. It cracks me up, and I laugh hysterically. He crawls right up on my lap and nibbles my nose." I'd grade this a B.”	Negative	Positive
“Not my normal read. A little slower than I expected, however.....the author kept me in the book. Anyone with an autistic child should read this one. The blend of suspense, familial ties, and art made for an interesting read”	Neutral	Positive
“I've read every book in this series although I haven't published reviews for all of them. This latest book continued the narrative in Kevin Rau's solid and easy to read style. For lovers of super-hero fiction its definitely a fine example.Unfortunately, that means that it has some serious flaws that in my mind drag it down to 3 stars.The first problem is that character development seems to have been almost completely abandoned at this point. Viewpoint characters change constantly so its rare that you get to fully experience a characters thoughts and feelings. Hints are given out, but nothing complete.This also contributes to the next issue I have which is that the enemies seem to be pulled out of comic-book central casting. They are uber-powerful except for 1-2 hard to discover weaknesses and they can put together elaborate plots and plans in hours. H.E.R.O. - Malice is a wonderful example of that as the bad guys develop part of the plan in less than 2 days despite the extremely high scientific barriers to success they themselves describe. Later they managed to modify a drug that has to be manufactured for a specific DNA signature in minutes and inject someone they captured far after their initial foray into capturing test subjects.That said, I like most of the	Neutral	Positive

characters, even if I feel that Kevin would do far better to choose one character and follow them throughout a book rather than jump from character to character every chapter. This pattern reminds me way too much of a comic book and while it can be fun, it definitely prevents these books from reaching the quality of Wearing The Cape, Confessions of a D-list Supervillain and Velveteen v the Junior Super Patriots."		

Appendix 4. Additional example for RoBERTa's misclassification

Review text	Actual polarity	Predicted polarity
"I will admit i didn't want to read this book after reading the reviews of it. Plus, it started off kind of slow. Not to mention there was a lot of internal dialogue going on. But let me just say by the time I got to the last page i gave myself a think in the head for not reading it when i first got it. I can't wait for the next one. I made up theories about the series. I'm just waiting for them to be confirmed now."	Positive	Negative

Appendix 5. More examples for BERT and DistilBERT's misclassifcations

Review text	Actual polarity	Predicted polarity
"I purchased this along with many other Disney books and found it a huge disappointment. It is filled with beautiful color pictures which is nice but the information is all extremely simple and not helpful at all.Disney is very expensive but certainly a worthwhile trip. It is very necessary to do a lot of research before planning to get the most for your buck and from your trip. I highly recommend THE UNOFFICIAL GUIDE TO WALT DISNEY WORLD and the PASSPRTER GUIDE for extensive research, guidance on everything needed to plan an amazing trip.For pictures of what you will see when there get this book. For me it ruins the surprises and some of the memories I want to make on my own.Disney is truly magical and shop wisely for the necessary guide books."	Negative	Positive
"The really only good thing about this Parker/Spenser is the reappearance of The Gray Man, who shot and almost killed Spenser in "Small Vices."The Bad is ever-present. Parker is surely coasting. There really is no plot--a fact which Spenser and Hawk try repeatedly to make into a joke but which is less funny the more they try.The standard characters appear, but mostly as cardboard cutouts. (The exception is Vinnie, who is likable but too brief.) There are really no themes, except the tired one of Vengeance. Yawn.And the Ugly is quite intrusive. The ugliest thing about the Spenser books has always been Susan. She is odious. Narcissistic, vain, shallow, not very bright, self-important and self-absorbed, unwitty, unlikable. (The amazing thing is that every character supposedly likes her. Why?) And in this one, there are TWO Susans--one named Cecile. Cecile is not as disgusting as Susan, but that's only because she appears less.	Negative	Neutral

Fortunately, at the end, she disappears entirely. Would that the real Susan would go and do likewise.(Note: this is not the worst Spenser. That dishonor goes to "Double Deuce." But it's close.)"		

Appendix 6. Additional example for GPT-2's misclassification

Review text	Actual polarity	Predicted polarity
"I listened to this book that is on several CDs. I listened in my car and enjoyed it very much. (We envy so many people before we know what is actually going on in their lives... Wynonna has been through so much. I hope it will be smooth sailing for her from here on out. I sure would never want to trade places with her). I enjoyed this book on CD and I learned a LOT from it as well. However it does get a tad slow in places so be prepared."	Positive	Neutral

List of Figures

Fig. 1. The Transformer Architecture (Vaswani et al., 2017) 13
Fig. 2. Methodology Summarized 15
Fig. 3. Distribution of ratings across the full data set 16
Fig. 4. Loss Curves over epochs 20
Fig. 5. Confusion matrices of actual vs predicted rating. 21
Fig. 6. Confusion matrices for actual vs predicted polarity. 22

List of Tables

Table 1. Overview of the architecture of the pretrained models 13
Table 2. Extract of the dataset 15
Table 3. Rescaled Ratings 16
Table 4. Hyperparameters used for fine tuning 17
Table 5. Evaluation metrics for fine-tuned models 20
Table 6. F-1 measure per rating 20
Table 7. F-1 Measure for aggregated ratings 22
Table 8. Common misclassification categories 23
Table 9. Common misclassifications due to user errors 24
Table 10. Common misclassifications due to subjectivity of reviews 24
Table 11. Common misclassifications due to contextual complexities 25
Table 12. Misclassified reviews for RoBERTa 26
Table 13. Examples of misclassified reviews for BERT and DistilBERT 27
Table 14. Examples of misclassified reviews by GPT-2 27