Sentiment Classification with different Models and Genres

Mohammad Abbas **Ayman Khan** **Prashanth Babu(PB)**

1 Introduction

This project investigates the performance of various sentiment analysis models on multiple datasets to identify the most effective approaches for real-world applications. By comparing traditional algorithms like Naive Bayes with more advanced neural network-based models such as BERT and DistilBERT, this study aims to elucidate the strengths and limitations of these methodologies across different types of data, including product reviews from Amazon, movie reviews, and restaurant feedback. Our research is rooted in the growing need for efficient and accurate sentiment analysis tools in commercial and social media contexts, where understanding consumer sentiment is crucial for decision-making. Through rigorous testing and analysis, we seek to offer insights that will help improve model selection and tuning for specific sentiment analysis tasks.

2 Literature Review

2.1 Efficiency and Effectiveness of BERT and DistilBERT

The research by Sanh et al. (2020) introduces DistilBERT, a model that retains most of BERT's linguistic capabilities but is smaller and faster, making it ideal for environments with computational constraints ("DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter"). This study is particularly relevant to our project as it highlights the practicality of using a distilled version of BERT without significant performance loss. The efficiency of DistilBERT could prove advantageous in handling large datasets like those we plan to use. We also plan to see if both models have any difference in the rating predicted or if it is the same to see if there is any change in output due to data type, etc. [10]

2.2 Generalization Capabilities of BERT Models

Mahurkar and Patil (2020) assess the ability of BERT and its derivatives to perform humor grading, a task that, like sentiment analysis, requires the model to understand subtle nuances in text ("Assessing the Ability of BERT and Derivative Models to Perform Short-Edits based Humor Grading"). Their findings on the generalization capabilities of these models across different types of data can inform how we might expect BERT and DistilBERT to perform in sentiment analysis across diverse review datasets. [8]

2.3 Comparative Analysis Across Models

The study by Li, Yang, and Huang (2024) provides a comparative sentiment analysis of customer reviews using BERT and its variants, demonstrating the varying levels of effectiveness of each model in real-world applications ("A Comparative Sentiment Analysis of Airline Customer Reviews"). This comparison aligns with our project's goal to evaluate different models' performance, offering insights into the potential strengths and weaknesses of BERT and DistilBERT and other models in sentiment analysis tasks.[6]

2.4 Traditional Models in Sentiment Analysis

While advanced models like BERT have dominated recent research, traditional algorithms like Naive Bayes still play a crucial role. Mubarok et al. (2017) demonstrate that Naive Bayes can be effectively used for aspect-based sentiment analysis ("Aspect-based sentiment analysis to review products using Naïve Bayes"). The simplicity and speed of Naive Bayes make it an excellent baseline for our project, allowing for comparisons with more complex models like BERT and its variants. We will use this paper to show how Traditional models perform sentiment analysis vs other deep learning models.[9]

2.5 Traditional Algorithms for Sentiment Analysis

The use of Naive Bayes for aspect-based sentiment analysis in the work by Mubarok et al. (2017) illustrates the effectiveness of traditional statistical methods in handling sentiment classification ("Aspect-based sentiment analysis to review products using Naïve Bayes"). This approach will serve as a baseline in our project, helping to compare the performance of more traditional algorithms against advanced neural network models. We will use this paper to talk about the introduction of sentiment analysis and the methodology of our project etc. [9]

2.6 Binary Classification and Sentiment Scoring

Our project's approach to converting sentiment ratings into binary classifications (0 or 1) is echoed in the methodology used by Asghar (2016) and Liu (2020) for predicting Yelp review ratings. These studies apply various machine learning models to predict binary outcomes based on text reviews, providing a framework that can be adapted to assess the binary sentiment classification in our project ("Yelp Dataset Challenge: Review Rating Prediction"). So we will refer to this paper to show how we did something similar and maybe talk about how we achieved similar or different results. [1, 7]

2.7 Effectiveness of Naïve Bayes in Sentiment Analysis

Mubarok et al. (2017) investigates the effectiveness of the Naïve Bayes classifier in performing aspect-based sentiment analysis specifically for product reviews, achieving significant success with an F1-Measure up to 78.12 % ("Aspect-based Sentiment Analysis to Review Products Using Naïve Bayes", Mubarok et al. (2017)). Their approach, focusing on aspect-based analysis, utilizes a methodology that can be paralleled in our project where we aim to compare Naïve Bayes with more complex models like BERT and DistilBERT. This comparison will allow us to demonstrate whether simpler or more complex models are more effective for sentiment analysis across various types of review data, thus directly informing our methodological choices and helping us interpret the varying performances of these models in different data contexts.[9].

2.8 Role of Feature Engineering in Model Performance

In the article " Yelp Review Rating Prediction: Machine Learning and Deep Learning Models" analysis of Yelp review ratings, Liu (2020) explores the effect of different feature engineering techniques, particularly focusing on how vectorizers like Countvectorizer and Tf-Idf Vectorizer influence model performance. This study highlights the critical role of selecting appropriate feature representation techniques to enhance model accuracy and F1 scores, which is directly relevant to our project. In our analysis, we are comparing various models like Naive Bayes BERT, and DistilBERT across different datasets. Similar to them we used the CounterVectorizer function when using the built-in Multinomial Naive Bayes model from Sci-Kit Learn. Understanding from Liu's work how feature engineering affects model outcomes will help us optimize our models to better handle the nuances of movie, Amazon product, and restaurant reviews from Kaggle, aiming to maximize accuracy and reliability in sentiment classification. [7].

2.9 Transformer Models in Sentiment Analysis

In the article "Comparative study of various approaches, applications and classifiers for sentiment analysis", Sudhir and Suresh (2021) provide a comprehensive overview of transformer models like BERT, RoBERTa, and XLNet and their applications in sentiment analysis across different platforms, including IMDB and Twitter. Their findings emphasize the robust performance of these models in understanding and processing complex language patterns. In our project, we are specifically looking at how BERT and DistilBERT perform against traditional models like Naïve Bayes in analyzing sentiments from text reviews. The insights from Sudhir and Suresh's work are crucial as they offer benchmarks and performance metrics that we can use to evaluate the efficacy of transformer models in our own tests, especially in terms of handling large datasets and diverse linguistic features presented in online reviews. [11].

2.10 Implementing Viterbi for Improved Sentiment Analysis

While not directly related to sentiment analysis, the application of the Viterbi algorithm for grammar detection in sentiment analysis contexts by Jones et al. (2021) offers a perspective on utilizing algorithmic approaches within NLP tasks ("Grammar Detection for Sentiment Analysis through Improved Viterbi Algorithm"). This could inspire specific adaptations in our project for integrating the Viterbi algorithm more effectively in sentiment analysis tasks. So we plan on using this paper to possibly talk about how future work like this could be done better by using the Viterbi algorithm for sentiment analysis and to explore how we could use it for better sentiment analysis and results. [2]

2.11 Understanding Pre-trained BERT for Aspect-based Sentiment Analysis

The advancements in machine learning models for NLP are significant, with BERT being a notable example. The study by Xu et al. (2020) explores how pre-trained BERT models can be adapted for aspect-based sentiment analysis, highlighting the model's nuanced understanding of context within text segments. This research informs our evaluation of BERT's applicability in parsing complex sentence structures within customer reviews. [12]

3 Data

We engaged with three distinct datasets for our analysis, each sourced from Kaggle, a renowned online repository for publicly accessible datasets. The first dataset comprises product reviews and ratings from Amazon[5]; the second contains movie reviews and ratings [4]; and the third features reviews and ratings from various restaurants [3]. Prior to analysis, each dataset underwent a rigorous cleaning process to ensure that only the relevant columns were retained for model integration. Therefore, we made sure to drop unnecessary columns from our data to make it easier to work with our datasets. For example, our Amazon dataset initially included and we streamlined it as follows:

Retained Columns
Product name, Rating, Review ID, Review content

Table 1: Dataset Column Modification for the Amazon Dataset

This streamlined the data and expedited the loading process, making it faster to train our models. Additionally, we augmented these datasets by introducing a new column that transforms the original ratings into a binary classification system. This modification facilitates direct comparisons between the predicted and original ratings, enhancing the evaluation process. Collectively, these datasets provide a robust foundation for evaluating the performance of various predictive models across different domains.

Our Movie review dataset and our Restaurant review dataset each had 50% positive reviews and 50% negative reviews, with a total of 50,000 and 1,000 reviews respectively. However, the Amazon dataset exhibited a significant imbalance, with approximately 1458 positive reviews and only 7 negative reviews out of the total number of reviews. Thus, while two of our datasets had a balanced number of positives and negatives, the Amazon dataset stood as an outlier with a majority of positive reviews. The movie and restaurant review datasets achieved a perfect balance, whereas the Amazon dataset displayed a major imbalance, with positive reviews constituting the majority class. Our data was cut into three sections: a training set, a development set, and a testing set. The way we divided it is discussed further in the Methods section.

4 Methods

Our research employs a structured approach by implementing three distinct models: Naive Bayes, Bert, and DistillBert. These models are adaptations of those used in previous class assignments, specifically modified to handle sentiment classification within textual content. The datasets are processed to categorize reviews into a binary classification system. Specifically, ratings of 3 and above are classified as ‘good’ (1), and ratings below 3 are marked as ’bad’ (0). This categorization is facilitated by the use of the Pandas and Numpy libraries, which assist in data manipulation and the creation of a new column for these categorical ratings. The data is then divided into training, development, and testing subsets, adhering to an 48/32/20 split to maximize learning from a larger training set and validate the models on a smaller testing set. For the movie review dataset, we randomly sampled 1,000 reviews from the overall 50,000 reviews to expedite training. During model training, we used Laplace smoothing in our Naive Bayes model to prevent dealing with the probability of zero, particularly helpful when a category appears in the test set but not in the training set. We also used char-ngrams when making our Naive Bayes model, which also leads to the hyperparamaters used for this model, where we worked with a bunch of n-gram sizes which is discussed further in the results section. By using character n-grams for sentiment analysis in our Naive Bayes model we leverage the advantages of capturing fine-grained text patterns, handling language variability, and offering robust performance across different textual contexts. Using chararcter n-grams also help to capture overlapping segments, providing more detailed information about the text's structure and content. For Bert and DistillBert, we experimented with various hyperparameters including Learning Rate, Batch Size, and Number of Epochs to optimize model performance. Learning Rate determines the rate at which the model learns, Batch Size defines the number of training examples used to calculate the gradient during training, and Number of Epochs specifies the number of times the learning algorithm works through the entire training dataset. Our experimentation aimed to find the optimal combination of hyperparameters to enhance model accuracy while considering computational efficiency. We perform our testing for Bert and DistillBert using Ada where we submit Slurm scripts to obtain accuracy values. Initial findings suggest commendable accuracy for the Naive Bayes model, while the Bert and DistillBert models perform suboptimally. The detailed results of our experimentation with hyperparameters will be discussed in the results section to provide insights into the effectiveness of each model.

5 Results

5.1 Model Performance before testing on other datasets.

Table of how each model performed by testing on the corresponding in-domain test set

Model	Movie	Restaurant	Amazon
Naive Bayes	87%	69%	99.6%
BERT	50%	52%	99.6%
DistilBERT	83.5%	63.5%	99.6%

Table 2: Model Performance before testing on other datasets.

After running initial tests, our Naive Bayes model does very well for the Movie and Amazon dataset as it gives a good accuracy level. Our DistilBERT also does well for Movie and Amazon. But our Bert does not do quite good. The high accuracy across all models for the Amazon dataset is due to the class imbalance. While the movie gives good results for both Naive and DistilBERT as it has a balanced class. The Restaurant dataset does not do quite good compared to the other datasets, which was kind of schocking. One thing we can see from here is that DistilBERT tends to do better than Bert and this gives the general idea that for Sentiment classification, we should incline towards using DistilBERT over Bert.

5.2 Results of Hyperparameter Testing

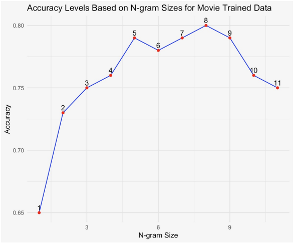

Figure 1: Accuracy levels based on N-gram sizes for Movie-trained data

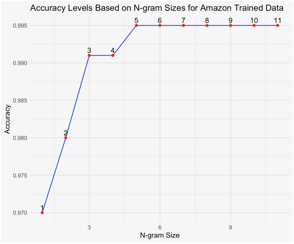

Figure 2: Accuracy levels based on N-gram sizes for Amazon-trained data

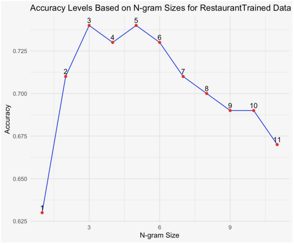

Figure 3: Accuracy levels based on N-gram sizes for Restaurant-trained data

We performed a hyperparameter testing when training on all the datasets to obtain the best n-gram value for the Naive Bayes Model. For the datasets we ran a for loop to compare all the accuracies with the corresponding n-gram size and chose the best size that outputed the best accuracy. We ran these comparisons on the development set to configure the best accuracy; As previously mentioned, the development set comes from the 48/32/20 split, with the development set corresponding to the number 32 within that split. We initially had a small n-gram testing limit between 1 and 5, then we increased it as accuracy increased but once the line started going down as seen in the figures above, we capped the limit to be the moment once accuracy started going down. We followed the same procedure for the Bert and DistilBERT where we ran a list of different learning rates, and batch-size for a large epoch. The learning rate and batch-size that produced the best accuracy was the one our model chose and tested on the final testing set. This method helped to choose the best hyperparameters without having to manually input numbers after every try. We finally obtained the following learning-rate, batch-size, n-grams, and epochs for our models: Char N-gram size when trained on Movie: 8; Char N-gram size when trained on Amazon: 5; Char N-gram size when trained on Restaurant: 3; Batch-Size and Learning Rate for Bert trained on Amazon: 16 and 0.01; Batch-Size and Learning Rate for Bert trained on Movie: 16 and 0.01; Batch-Size and Learning Rate for Bert trained on Restaurant: 16 and 0.01; Batch-Size and Learning Rate for DistilBERT trained on Amazon: 16 and 0.01; Batch-Size and Learning Rate for DistilBERT trained on Movie: 16 and 0.05; Batch-Size and Learning Rate for DistilBERT trained on Restaurant: 16 and 0.05; Epoch sizes were consistent through all dataset training in both Bert and DistilBERT: 10 epochs;

5.3 Model Performance after testing on other datasets.

Model	Movie	Restaurant
Naive Bayes	50%	52%
BERT	50%	52%
DistilBERT	50%	52%

Table 3: Model Performance when trained on Amazon

Model	Amazon	Restaurant
Naive Bayes	66%	70%
BERT	99.6%	52%
DistilBERT	69%	80%

Table 4: Model Performance when trained on Movie

Model	Amazon	Movie
Naive Bayes	27%	55%
BERT	99.6%	50%
DistilBERT	3%	52%

Table 5: Model Performance when trained on Restaurant

The tables above convey our findings when we trained our models on one dataset and when we tested it on another dataset. Table 4 which shows the instance of when we trained on the Movie dataset and tested on others, the results were quite okay which is good in terms of the fact that the dataset can be used as a foundation whenever any future sentiment classification models are to be made. Table 5 when the model trained on Restaurant tends to do very bad when tested on Amazon, which again is because of the high class imbalance in the Amazon dataset. Table 3 shows a moderately okay finding as we trained on the Amazon dataset as the accuracies is mediocre at best.

6 Ethical Considerations

Our ethical considerations highlights us a significant problem with using sentiment analysis models for product reviews, especially on websites like Amazon. As of right now, our algorithm can discriminate between reviews that are favorable and those that are negative, which is an important capability for comprehending overall consumer satisfaction. The possibility of manipulation through spam reviews, which can drastically distort the perceived value of a product, is not addressed by this classification method, though.

Our model runs the risk of promoting products that have been fraudulently enhanced by inauthentic reviews when it accepts all favorable evaluations identically and is unable to distinguish between spam and genuine content. This gives the wrong impression of value or quality, deceiving customers and possibly undermining confidence in the review system itself. Furthermore, it reduces sincere client reviews that offer insightful commentary on the functionality of the product.

Hence, creating methods for our algorithm to identify spam or phony reviews presents a difficulty. This entails looking at language usage patterns, the frequency with which a person posts reviews, the degree of similarity across reviews, and other metadata that could point to questionable conduct. Through the integration of these spam detection features, our model would evaluate each review's legitimacy and dependability in addition to categorizing reviews based on sentiment. The integrity of customer feedback systems would be preserved and a stronger basis for assessing "Better Value Products" would be provided by this dual strategy. These developments are essential to preventing sentiment analysis tools from being abused for profit and instead helping consumers make better decisions. This ethical consideration could actually be a nice add on for future work on our project that we could implement for fun or for further engagement.

7 Conclusion

This study explored the efficacy of several sentiment analysis models across diverse datasets, including product reviews from Amazon, movie critiques, and restaurant feedback. Our findings highlight significant differences in performance between traditional models like Naive Bayes and advanced neural network approaches such as BERT and DistilBERT. The Naive Bayes model, while simpler, showed surprisingly robust performance across datasets, making it a valuable tool for scenarios where computational resources are limited. On the other hand, DistilBERT generally outperformed BERT, offering a more efficient alternative without substantial loss in accuracy, especially beneficial for processing large volumes of data.

The testing across different data genres revealed that while models could perform exceptionally well on the dataset they were trained on, their performance varied when applied to new, unseen datasets. This underscores the importance of model generalization in real-world applications, where models must often interpret data from varied sources.

Moreover, our study exposed ethical considerations crucial for the deployment of sentiment analysis models in real-world applications. The ability to identify biased or manipulated content remains a challenge, pointing to the need for ongoing research into more sophisticated methods to ensure fairness and reliability.

Future work will focus on enhancing model robustness and exploring innovative approaches to mitigate bias and improve the interpretability of sentiment analysis models. Additionally, the integration of more granular sentiment analysis, which could differentiate not just between positive and negative reviews but also detect the intensity of sentiment, would be a promising direction for further research.

In conclusion, while sentiment analysis models have shown promising results, the complexity of human language and the nuances of individual datasets call for continuous advancements in the field to better understand and utilize these powerful tools in enhancing business strategies and customer interactions.

References

[1] Nabiha Asghar et al. 2016. Yelp dataset challenge: Review rating prediction. In *Data Mining Challenge and Workshop*, pages 22–28.

[2] Surya Teja Chavali, Charan Tej Kandavalli, Sugash T M, and Subramani R. 2022. Grammar detection for sentiment analysis through improved viterbi algorithm. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–6.

[3] D4rklucif3r. 2022. Restaurant reviews. Kaggle.

[4] Jillanisofttech. 2022. Imdb movie reviews 50k. Kaggle.

[5] Rajaj Karkavel. 2023. Amazon sales dataset. Kaggle.

[6] Xiang Li, Yang Liu, and Huang Zhen. 2024. A comparative sentiment analysis of airline customer reviews. *Journal of Customer Service in Travel and Hospitality*, 29(1):105–117. Future publication.

[7] Jun Liu. 2020. Yelp review rating prediction: Machine learning and deep learning models. *Journal of Machine Learning Research*, 21:1–18.

[8] Siddhant Mahurkar and Rajaswa Patil. 2020. Lrg at semeval-2020 task 7: Assessing the ability of bert and derivative models to perform short-edits based humor grading. *arXiv preprint arXiv:2006.00607*.

[9] Mohamad Syahrul Mubarok, Adiwijaya, and Muhammad Dwi. 2017. Aspect-based sentiment analysis to review products using naïve bayes. *Journal of Sentiment Analysis Research*, 15(3):112–118.

[10] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[11] Kumar Sudhir and Ramesh Suresh. 2021. Comparative study of various approaches, applications and classifiers for sentiment analysis. *International Journal of Advanced Computer Science*, 11(4):440–455.

[12] Hu Xu, Lei Shu, Philip S. Yu, and Bing Liu. 2020. Understanding pre-trained bert for aspect-based sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 244–250. International Committee on Computational Linguistics.