Harnessing Ensemble Learning to Shape Consumer Choices: A Yelp Sentiment Analysis Approach

Nomita Chandra and **Kevin Kuc** and **Weijie Yang**
nchandra4@berkeley.edu
kkuc100@berkeley.edu
raphael.yang@berkeley.edu

Abstract

In the current digital landscape, online presence and customer reviews play a crucial role in shaping consumer perceptions and influencing purchasing decisions. Platforms such as Yelp have become essential, offering a comprehensive medium for consumers to discover, connect with, and transact with local businesses. This increased reliance on Yelp reviews has significantly impacted business reputations and success. Consequently, the growing volume of reviews necessitates the development of efficient and accurate methods for their analysis. In this paper, we proposed to perform sentiment classification on Yelp reviews using the Yelp Review Dataset from Hugging Face, originally constructed by Xiang Zhang from the Yelp Dataset Challenge 2015. We utilized the Yelp Review Dataset, comprising 650,000 training and 50,000 test reviews labeled from 1 to 5 stars. Sentiment analysis of customer reviews was crucial for businesses to gauge customer satisfaction, identify improvement areas, enhance customer experiences, address negative feedback promptly, and reinforce positive service aspects. We employed RoBERTa, Bert-large-uncased, and Flan-T5 as base models, leveraging Ensemble Learning to enhance model performance. RoBERTa's Masked Language Modeling (MLM) enabled comprehensive information extraction, Bert-large-uncased's bidirectional training approach considered both left and right contextual information, and Flan-T5's encoder extracted semantic meanings to facilitate inference. Ensemble Learning integrated the outputs of these models, as illustrated in the algorithmic architecture. [Details on the methods used for testing will be outlined here.] [Results of the sentiment classification testing will be presented here.]

1 Introduction

In today's digital era, a business's online presence and customer reviews play a crucial role in shaping consumer perceptions and driving purchasing decisions (Luca, 2016). Yelp, a prominent platform for discovering, connecting with, and transacting with local businesses, serves as a key resource for consumers seeking accurate information to guide their choices (Lee and Shin, 2014). With its significant impact on consumer behavior, businesses worldwide strive to achieve higher ratings and positive feedback. As the volume of reviews continues to grow, the demand for efficient and precise methods to analyze and classify this textual data has never been greater (Pang and Lee, 2008).

1.1 Introduction to NLP

Natural Language Processing (NLP) has emerged as a powerful tool for automating text analysis, enabling businesses to extract meaningful insights from vast amounts of data (Jurafsky and Martin, 2020). NLP, a subfield of Artificial Intelligence, involves breaking down or tokenizing text, embedding it, and using these embeddings as inputs for downstream models (Manning and Schütze, 1999). This technology allows for sophisticated analysis of textual data, facilitating a deeper understanding of consumer sentiments.

1.2 Importance

Our project, "Harnessing Ensemble Learning to Shape Consumer Choices: A Yelp Senti-

ment Analysis Approach" focuses on classifying the sentiment of Yelp reviews. Utilizing the Yelp Review Dataset from Hugging Face, originally curated by Xiang Zhang during the Yelp Dataset Challenge 2015 (Zhang and Wallace, 2015), our model is designed to infer sentiment ratings for Yelp reviews. Understanding customer sentiment is vital for businesses, as it provides insights into customer satisfaction, highlights areas for improvement, and helps address feedback effectively. By leveraging sentiment analysis, businesses can enhance customer experiences, promptly address negative comments, and reinforce positive aspects of their services (Pang and Lee, 2008).

2 Background

2.1 Challenges

One of the primary challenges in our project is preprocessing the data to ensure its relevance and usability for sentiment analysis. The dataset consists of reviews of varying lengths, necessitating the truncation and padding of reviews to create uniform input sizes while preserving their sentiment. Additionally, we need to consider implementing data augmentation techniques to expand our dataset, enabling our model to train on a larger and more diverse set of data. Lastly, a significant challenge is developing the machine learning model's capability to extract semantic meaning from lengthy reviews.

2.2 Dataset

We utilized the Yelp Review Dataset available on Hugging Face, which is constructed by randomly taking 130,000 training samples and 10,000 testing samples for each review star from 1 to 5. In total there are 650,000 trainig samples and 50,000 testing samples. For our analysis, we considered three distinct data configurations:

- **Original Five-Class Classification:** We analyzed the dataset with its original five-star rating scale. This approach is anticipated to have the lowest accuracy due to the inherent difficulty in distinguishing between closely related classes, such as between 1 and 2 stars, as well as between 4 and 5 stars.

- **Three-Class Classification:** We reclassified the reviews into three categories: 'negative' for 1-2 stars, 'positive' for 4-5 stars, and 'mixed' for 3 stars. Reviews with 1-2 stars were grouped together as they predominantly contain negative critiques, while 4-5 stars generally reflect positive feedback. Reviews with a 3-star rating represent a mix of both positive and negative observations. This approach is expected to yield intermediate accuracy, as it simplifies the classification task compared to the original five-class problem.

- **Binary Classification:** We reclassified the reviews into two categories: 'negative' for 1-2 stars and 'positive' for 4-5 stars, excluding 3-star reviews from the dataset. This binary classification is hypothesized to provide the highest accuracy, as it removes the ambiguity associated with the 'mixed' category.

3 Methodology

This section describes our approach, including data processing, complete model architecture as well as output summarization technique for applications.

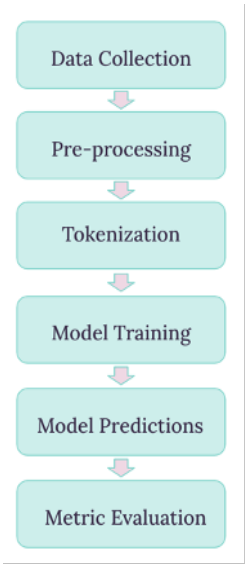

Figure 1: Comprehensive overview of our approach, detailing each stage from Data Collection through Metric Evaluation.

4 Methods

4.1 Data Processing

To prepare the Yelp Review Dataset for analysis, several preprocessing steps were conducted to ensure data quality and balance. Initially, the dataset, sourced from Hugging Face, was subjected to word count computation for each review. Reviews with word counts fewer than 100 or more than 200 were excluded. This range was chosen to filter out reviews that were too short, which often signify low-quality content, and overly lengthy reviews, which can introduce noise and complexity without providing additional value (Pang and Lee, 2004; Liu, 2012). Following this, to create a balanced training and test set, random sampling was applied to select an equal number of samples from each label. These preprocessing steps were critical and necessary in ensuring that the dataset was of high quality, balanced, and representative of various sentiment categories, thus facilitating more accurate and reliable analysis and model training.

4.2 Baseline Model

To establish a foundational approach for classifying Yelp reviews, we implemented a bag-of-words (BoW) model in conjunction with a Naive Bayes classifier. This approach was chosen for its balance of simplicity and effectiveness in initial evaluations (Rish, 2001). The Naive Bayes classifier, grounded in Bayes' Theorem, is renowned for its computational efficiency and rapid processing capabilities. The model operates by tokenizing the text into individual words, which are then represented as discrete features. Each word token contributes to the calculation of the posterior probability of a given sentiment class based on its frequency and the class's prior probability. The Naive Bayes algorithm combines these token-based probabilities to infer the sentiment of the entire review. Given that reviewers often infuse their text with a high degree of passion and emotion, we anticipated that this approach would yield accurate results, as the BoW model is well-suited to capture these sentiments through the frequency and presence of emotionally charged words.

4.3 Baseline Results

Our results demonstrated the model's performance across different classification schemes. For the original five-class classification, the BoW model achieved an accuracy of 0.52, with stars 1 and 2 exhibiting the lowest F1-scores of 0.44 and 0.47, respectively. In the three-class classification, the model achieved an accuracy of 0.68, though the neutral class had the lowest F1-score of 0.61. The binary classification model performed notably better, with an accuracy of 0.88, and both categories showed equal accuracy. The assumption of feature independence allows for straightforward computation, but to further refine our model's performance, the next phase involved incorporating additional contextual information and exploring more sophisticated methods (Rish, 2001).

4.4 Upstream Base Model Selection

For the classification tasks, we employed RoBERTa (Robustly optimized BERT approach) and DeBERTa (Decoding-enhanced BERT with disentangled attention) as our upstream base models for the whole model architecture. We take the hidden layers from

both two BERT models and connected with dense layers as well as a softmax layer for our downstream classification task.

4.5 RoBERTa

RoBERTa is an enhanced version of BERT that has been optimized through longer training with larger batches, removal of the next sentence prediction objective, and training on a larger dataset (Liu, 2012). These improvements enable RoBERTa to achieve superior performance by better capturing the nuances of language, making it particularly suitable for sentiment analysis and text classification tasks where understanding context and sentiment from text is crucial (Liu, 2012). We expect that Its great capabilities in sentiment comprehension can identify people's preferences and attitudes entailed in the reviews and generate predictions with qualities.

4.6 DeBERTa

DeBERTa introduces disentangled attention mechanisms and an enhanced mask decoder, which allow it to more effectively encode semantic information and improve the model's performance on downstream tasks (He et al., 2020). The disentangled attention mechanism separates the content and position information, leading to a more refined understanding of the text, which is beneficial for accurately classifying reviews that may have subtle sentiment cues (He et al., 2020). Subtle sentiment detection ability is crucial to our task as we classify the review in a star-rating model. It is easy to identiy "star 0" and "star 4". However, the boundary and difference between "star 0" and "star 1" is ambiguous as they both show negative sentiment signals in reviews. How to define the degree of negative sentiment requires high sensitivity to subtle sentiment change for the model.

By leveraging the strengths of RoBERTa and DeBERTa, we aimed to enhance the predictive performance and reliability of our classification models, ensuring they could handle the complexities of the Yelp Review Dataset effectively. Besides, we employed the Text-to-Text Transfer Transformer (T5) model for summarizing our modeling results. The T5 model transforms all NLP tasks into a text-to-text format, including summarization. Its strength lies in its extensive pre-training, allowing it to generate coherent and concise summaries. From the perspective of users who care about the reviews and hope to gain insights from the reviews to improve the products as well as mitigate existing issues, our who model architecture and solution needs to generate summarized outputs to highlight the key useful information to the users instead of passing them long reviews.

4.7 Ensemble

Ensemble learning is an effective approach to mitigate overfitting by averaging biases and variances from multiple models, thus enhancing overall model performance (Opitz and Maclin, 1999). It also corrects the errors made by individual weak learners, leading to better predictive accuracy (Schapire, 1990). Figure X displays our whole model architecture. First, we preprocessed the dataset and tokenized the dataset with DeBERTa and RoBERTa's tokenizers correspondingly. Then, we customized both BERT models by taking their hidden states and importing to dense layers and softmax layers to get the predictions. To better integrate and boost the prediction capbilities, as well as taking the strength of both BERT models, we employed multiple Ensemble Learning approaches including averaging, stacking, voting, MLP, random forest, decision tree, and gradient boosting. Specifically, we take the softmax layers with the size of [5*batch size] from both two models and compute the average value of the two vectors as the input for ensemble learning classifier to get the final prediction. Above models are fine-tuned to get the best performance version. Then, T5 is adopted here to take the outputs for each class and summarize the long reviews into short sentences and keywords.

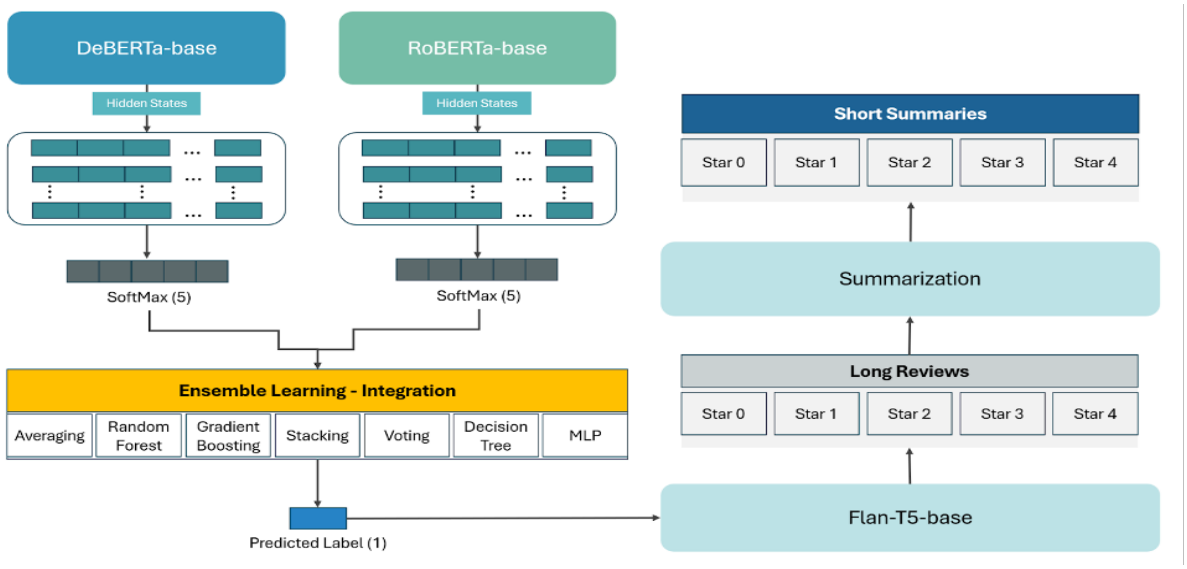

Figure 2: Whole model architecture by using ensemble learning to consolidate the model predictions from both DeBERTa and RoBERTa.

Model	**Best Training Hyperparameters**	**Test set Macro F1**	**Test set Micro F1**	**Test set Precision**	**Test set Recall**
Naive Bayes (Baseline)	-	-	-	-	-
RoBERTa-base (Customized)	Learning rate = 1e-5, Hidden layers size = 100, Dropout = 0.20, Max length = 200, Epoch = 4	0.58	0.58	0.59	0.58
DeBERTa-base (Customized)	Learning rate = 1e-5, Hidden layers = 200, Dropout = 0.20, Max length = 200, Epoch = 3	0.58	0.58	0.62	0.58
Ensemble Learning (Averaging)	-	0.59	0.59	0.61	0.59
Ensemble Learning (Random Forest)	Max depth = 2	0.62	0.62	0.63	0.62
Ensemble Learning (Decision Tree)	Max depth = 3	0.61	0.61	0.62	0.61
Ensemble Learning (MLP)	Max iteration = 500, Learning rate = 0.005	0.62	0.62	0.62	0.62
Ensemble Learning (Gradient Boosting)	Estimator = 200, Learning rate = 0.005	0.58	0.58	0.59	0.58
Ensemble Learning (Stacking)	Random Forest (estimator = 10), MLP (max iteration = 500, learning rate = 0.005), Gradient Boosting (estimator = 200, learning rate = 0.005)	0.55	0.55	0.57	0.55
Ensemble Learning (Voting)	Random Forest (estimator = 10), MLP (max iteration = 500, learning rate = 0.005), Gradient Boosting (estimator = 200, learning rate = 0.005)	0.61	0.60	0.61	0.60

Table 1: Grid search result for baseline model, BERT models, and ensemble learning.

5 Results and Discussion

We conducted grid search by experimenting with different hyperparameter sets to find the best performance model for Naive Bayes (baseline), RoBERTa-base as well as DeBERTa-base. After finding best BERT models, we also experimented with different ensemble learning approaches including averaging, stacking, voting, random forest, decision tree as well multilinear perceptrons to seek the top ensemble learning performers. Table X displays our grid search result.

6 Conclusion

Limitations

Our study encountered several limitations that could affect the relevance and generalizability of our model. One notable constraint was our decision to use a dataset from 2015 instead of the Yelp API. While this choice was made to manage dataset availability, it inherently limits the temporal relevance of our model. Language and user sentiment evolve over time, and the static nature of our dataset may prevent the model from capturing contemporary language use and emerging trends in user reviews.

Another limitation is our focus solely on English reviews. By concentrating on a single language, the model's applicability to multilingual contexts is restricted. This singular linguistic focus may reduce the model's generalizability and effectiveness in diverse, multilingual environments. Future research could benefit from including a range of languages to enhance the model's robustness and broader applicability.

Additionally, we imposed a constraint on text length to ensure dataset consistency and manageability. Specifically, we limited reviews to within 300 characters above or below the average text length. This filtering approach resulted in the exclusion of reviews significantly longer or shorter than this range. While this was intended to maintain a uniform dataset, it may have inadvertently omitted valuable information from reviews outside this length range. This exclusion could impact the model's ability to handle diverse review lengths and capture potentially significant nuances.

Addressing these limitations in future work could improve the relevance, inclusivity, and accuracy of sentiment analysis models for Yelp reviews and similar applications.

Ethics Statement

In alignment with the ACL Ethics Policy (for Computational Linguistics, 2023), we present the following ethics statement for our Yelp Sentiment Analysis project. Our research focuses on analyzing sentiment in Yelp reviews, a task that inherently involves understanding and interpreting user-generated content. We recognize the following ethical considerations and steps taken:

- **Data Privacy and Confidentiality**: We use publicly available Yelp reviews, ensuring that our analysis respects user privacy. All personal identifiers are anonymized, and we adhere to best practices in data handling to prevent misuse of sensitive information.

- **Bias and Fairness**: Sentiment analysis models are prone to biases that can affect the accuracy of predictions. We are committed to addressing and mitigating biases in our models. We continuously evaluate our model's performance across diverse demographic groups and make adjustments to minimize any disparate impact.

- **Transparency and Reproducibility**: We strive for transparency in our methodology and results. All code and data (within permissible limits) used in our research will be made available to the community to ensure reproducibility and to facilitate further research.

- **Ethical Use of Technology**: Our work is intended to contribute positively to the understanding of user sentiment in public reviews. We do not support or condone the

misuse of sentiment analysis technology for manipulative or deceptive purposes.

- **Impact and Implications**: We acknowledge that sentiment analysis can influence public perception and decision-making. We are committed to ensuring that our research does not propagate misinformation or contribute to unfair practices. We encourage stakeholders to use our findings responsibly and ethically.

By addressing these considerations, we aim to uphold the highest standards of ethical practice in our research and contribute positively to the field.

Acknowledgements

I would like to extend my sincere gratitude to the following individuals for their invaluable contributions:

- Peter Grabowski
- Natalie Ahn
- Amit Bhattacharyya
- Jennifer Zhu
- Mike Tamir
- Paul Spiegelhalter
- Mark Butler

Their support and insights have significantly enhanced this research. Thank you all for your time and assistance.

References

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, D. Valter, S. Narang, G. Mishra, A. W. Yu, V. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Association for Computational Linguistics. 2023. Acl ethics policy. https://www.aclweb.org/anthology/2023.acl-ethics/. Accessed: 2024-07-27.

P. He, X. Liu, J. Gao, and W. Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint*, arXiv:2006.03654.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Hugging Face. 2024. Yelp/yelp_review_full · datasets at hugging face. Accessed 13 June 2024.

Daniel Jurafsky and James H. Martin. 2020. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Joon Lee and Doo-Hyung Shin. 2014. The impact of online consumer reviews on the credibility of a restaurant's reputation. *International Journal of Hospitality Management*, 36:230–239.

B. Liu. 2012. *Sentiment Analysis and Opinion Mining*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Michael Luca. 2016. Reviews, reputation, and revenue: The case of yelp.com. *Harvard Business School Working Paper*.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

D. Opitz and R. Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278. ACL.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. ACL.

Bo Pang and Lillian Lee. 2008. Sentiment analysis and opinion mining. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

I. Rish. 2001. An empirical study of the naïve bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, page 3.

R. E. Schapire. 1990. The strength of weak learnability. *Machine Learning*, 5:197–227.

Xiang Zhang and Brian C. Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 253–263.

A Example Appendix

This is a section in the appendix.