Sentiment Analysis to Classify Restaurant Reviews

Alejandro Gutierrez Acosta, Sana Barakat and Sofia Moreno Lasa
School of Science and Technology, IE University, Madrid, Spain

agutierrez.ieu2020@student.ie.edu, sbarakat.ieu2021@student.ie.edu,
smoreno.ieu2021@student.ie.edu

Prof. Suzan T. S. Awinat
AI: Natural Language Processing & Semantic Analysis

ABSTRACT

In this study, we examined how user-generated metadata can enhance the accuracy and robustness of sentiment analysis models for classifying restaurant reviews. As online reviews significantly influence consumer choices and business reputation, it is crucial to develop precise sentiment analysis tools. We compared two sentiment analysis models: the traditional Random Forest and the advanced BERT Transformer. The study involved a dataset of restaurant reviews, which were processed and analyzed using these models. Our findings indicate that the BERT Transformer model, which incorporates user-generated metadata such as ratings and reviewer profile details, outperforms the Random Forest model in terms of accuracy and robustness. This suggests that deeper, context-aware analysis models are more effective in interpreting complex user sentiments. The results not only confirm our hypothesis but also reveal the potential for metadata to significantly refine sentiment analysis. Surprisingly, some anomalies in model predictions underscored the challenges of sarcasm and implicit sentiments in textual reviews. Further research could explore these aspects more deeply, potentially improving model sensitivity and broadening its applicability. The insights gained from this study are valuable for businesses aiming to enhance their strategies based on customer feedback and for the academic community that continues to advance the field of natural language processing.

KEYWORDS

Sentiment analysis, Restaurant reviews, BERT Transformer, Random Forest, Natural language processing

1. INTRODUCTION

The rise of online review sites has changed the way that the restaurant business operates by giving customers a voice and influencing the perceptions of the restaurants they eat at. Understanding and correctly interpreting the sentiment expressed in these evaluations has become essential for restaurants in this digital age that want to draw in discerning customers and keep a strong brand image. [1]

By investigating the possibility of using user-generated metadata to improve the precision and resilience of sentiment analysis models in classifying restaurant reviews as positive or negative, this study seeks to advance the rapidly developing field of sentiment analysis. In particular, we aim to tackle the subsequent research query: How does using user-generated metadata enhance the accuracy and robustness of sentiment analysis models in categorizing restaurant reviews as positive or negative?

We compare two different sentiment analysis models—the conventional Random Forest model and the cutting-edge BERT Transformer model—in order to fully answer this question. The BERT Transformer model embodies the most recent developments in deep learning architectures for nlp tasks, whereas the Random Forest model represents a traditional approach to machine learning. [2]

This discovery has profound implications for company strategy and customer behavior that go well beyond the boundaries of computational linguistics. Restaurants can improve their comprehension of customer sentiment and adjust their strategies to meet customer expectations and boost overall satisfaction by analyzing the effects of incorporating user-generated metadata, such as user ratings, review sentiments, and other contextual factors.[1][3]

The following sections of this report cover the details of our methodology, explain the conclusions drawn from our experimentation, analyze the implications of our findings for researchers and businesses, and offer thoughtful suggestions for more research in this area. By means of this complex project, we hope to shed light on the best methods for sentiment analysis of restaurant reviews, which will lead to better decision-making and better customer experiences in the ever-changing restaurant industry.

2. RELATED LITERATURE

Sentiment analysis, a pivotal aspect of natural language processing, has garnered substantial attention in recent literature, particularly within the context of social media platforms and customer reviews of restaurants. In a study conducted in Karachi, Pakistan, researchers delved into the sentiments embedded within restaurant reviews collected from a prominent Facebook community. Employing advanced sentiment analysis techniques, the study aimed not only to categorize comments as positive or negative but also to employ text categorization methods to delve into specific aspects such as food taste, ambiance, service, and value for money. The findings from this research revealed that customers often express sentiments related to service quality, indicating its paramount importance in shaping overall restaurant experiences [3].

Meanwhile, another notable study emphasized the challenges and significance of sentiment analysis on Twitter, a widely-used social media platform renowned for its brevity and real-time nature. Despite its pervasive usage, accurate sentiment representation on Twitter encounters hurdles such as class imbalance and feature richness. To address these challenges, researchers proposed a novel framework for sentiment analysis tailored specifically for Twitter data. Leveraging advanced machine learning classifiers including Support Vector Machines (SVM), Logistic Regression, Random Forest, and Naive Bayes, the study aimed to extract nuanced sentiments from tweets, highlighting implications for businesses, governments, and public figures [4].

Furthermore, a comprehensive research endeavor focused on sentiment analysis of restaurant reviews sourced from the Yelp website, a renowned platform for consumer feedback. Employing both binary and ternary classifications, the study aimed to categorize sentiments as positive, negative, and neutral, providing a nuanced understanding of customer perceptions. Leveraging a diverse array of machine learning, deep learning, and transfer learning models, the research achieved commendable accuracy rates in sentiment classification. Additionally, a novel unsupervised approach for aspect-level sentiment classification was introduced, leveraging semantic similarity and pre-trained language models like GloVe. Noteworthy findings included a maximum accuracy of 98.30% using the ALBERT model, showcasing the efficacy of the framework in extracting sentiment-rich insights from online reviews [5,6].

Collectively, these studies underscore the profound impact of sentiment analysis in extracting valuable insights from online reviews and its implications for businesses, consumers, and decision-making processes. They shed light on the evolving landscape of sentiment analysis research within the realm of social media platforms and online consumer feedback, emphasizing the growing importance of leveraging sentiment analysis for informed decision-making and enhanced customer experiences.

3. EXPERIMENTAL METHODOLOGY

3.1 Dataset and Preparation

The dataset used in this study was obtained from Kaggle and consists of restaurant reviews with 10,000 rows and 8 columns. The columns include: restaurant's name, reviewer's name, review, rating given by the reviewer, metadata associated with the reviewer, time of the review, and pictures attached to review.

In preparing the dataset for analysis, several steps were undertaken to clean and preprocess the data. Firstly, missing and irrelevant data were addressed through careful examination and filtering processes to ensure data integrity. Additionally, data type conversion was performed to standardize formats across different columns, facilitating uniform processing. Understanding the distribution of ratings allowed for insights into overall sentiment trends and helped identify any anomalies or outliers. Moreover, examining the performance of restaurants based on reviews provided valuable context for interpreting sentiment analysis results. To focus the analysis on relevant content, reviews were filtered based on criteria such as length and relevance to the study objectives. Subsequently, the text data underwent tokenization and cleaning processes to break down sentences into individual words or tokens and remove unnecessary characters and symbols. Lastly, stemming techniques were applied to standardize word forms and reduce variation, enhancing the efficiency of subsequent analysis tasks. These preprocessing steps were crucial in preparing the dataset for sentiment analysis, ensuring accuracy and reliability in the interpretation of results.

To filter negative reviews, those with low ratings were selected, defining negativity based on these ratings. Following this, the reviews were tokenized and cleaned, involving the removal of stopwords and punctuation to streamline further analysis. Additionally, lemmatization was employed to standardize words to their base or dictionary form, enhancing consistency in analysis. For further simplification, stemming was optionally used to reduce words to their root forms. To gain insights into the distribution of words, frequency distribution was calculated, revealing the occurrence of terms across the dataset. Visualizing the most common words through plots facilitated a clearer understanding of word frequency distribution, aiding in identifying prevalent themes or sentiments within the reviews. These steps collectively prepared the data for sentiment analysis, enabling comprehensive exploration and interpretation of customer feedback.

The analysis of common words within the dataset reveals recurring terms such as "food," "place," "order," "service," and "chicken," indicating these aspects are frequently discussed in negative reviews. Notably, there is a noticeable drop in frequency after the initial words, suggesting that a select few terms are heavily used while others are less common. Additionally, the presence of specific words like "service," "time," "quality," and "experience" highlights potential areas of concern for customers, possibly relating to the speed and quality of service, as well as the overall dining experience. Furthermore, the appearance of words like "never" towards the end of the frequency plot hints at strong negative sentiments, potentially indicating dissatisfaction or a vow to avoid returning or recommending the establishment. These insights shed light on key themes and sentiments prevalent in the dataset, offering valuable guidance for further analysis and improvement efforts within the restaurant industry.

Data cleaning and preprocessing were performed on all reviews to enhance feature engineering and model training. Stemming and lemmatization were experimented with to determine the best text normalization technique. Stemming followed by lemmatization and vice versa were applied to tokenize and clean the reviews. Features and labels were prepared using TFIDF, and the dataset was split into training and testing sets. The training set comprised 80% of the data, while the testing set contained 20%. The resulting shapes of the training and testing data for the lemmatization-first approach were (7964, 9955) and (1991, 9955), respectively, indicating successful data splitting after cleaning missing values.

3.2 Experimental Tools

In this study, we deployed two different implementations of the Random Forest algorithm and one BERT Transformer model to classify sentiment in restaurant reviews. The choice of algorithms was driven by their distinct processing capabilities and proven efficiency in handling natural language data.

Random Forest

The Random Forest algorithm, an ensemble learning technique known for its robustness, creates numerous decision trees during training and determines the mode of the classes predicted by the individual trees. This algorithm was chosen due to its ability to effectively address overfitting and its capability to handle non-linear data. As previously mentioned, we experimented with variations during the data preprocessing phase:

1. Stemming then Lemmatizing: This approach, despite appearing counterintuitive as stemming typically involves more aggressive reduction of words, could've proved beneficial for certain complex word forms potentially improving the effectiveness of subsequent lemmatization.
2. Lemmatizing then Stemming: By applying lemmatization first, words were reduced to their dictionary forms before being further reduced by stemming, potentially normalizing different lemmas that share the same stem.

The variation in the order of preprocessing implementation proved significant, as shown in the results later on, the accuracy, precision, recall, and F1-scores for both models diverged from one another.

As for Random Forest's asymptotic behaviour, the time complexity is primarily determined by several factors, including the number of trees in the forest (T), the number of samples (N), and the number of features (M). The complexity of building a single tree is denoted as O(N * M * log(N)), assuming that each tree is constructed using all features. Consequently, when considering T trees, the overall training complexity becomes O(T * N * M * log(N)). As a result, the Random Forest algorithm can be computationally expensive when dealing with large datasets and a high number of trees.

For our case, we configured each RF with 300 estimators (trees) and a maximum tree depth of 15. This setup was chosen to balance between model complexity and the risk of overfitting. However, thanks to the use of a GPU and its parallel processing capabilities, the computation time was mitigated, taking less than 10 seconds to completely train most of the time.

BERT

The BERT model was chosen for its state-of-the-art performance in various natural language processing tasks. Its architecture enables it to comprehend the context of a word by taking into account all of its neighboring words (both to the left and right). Our implementation leveraged the Hugging Face Transformers library, specifically adapted to our sentiment analysis task. Data preprocessing included the unique tokenization and encoding suitable for BERT processing, which meant no experimentation similar to Random Forest.

While the base code for both the Random Forest models and the BERT Transformer was sourced from existing implementations—predominantly Hugging Face's BERT Documentation[7] and the original BERT paper[8]—notable modifications were made to tailor the algorithms to the specific challenges and data characteristics of restaurant review sentiment analysis.

As for BERT's asymptotic behavior, it is primarily characterized by its dependency on the sequence length and the size of the model. The primary factors which contributed to this are: Sequence Length (S), Number of Layers (L), and Hidden Size (H). Combining these factors, the overall time complexity for BERT in processing an input sequence can be approximated as O(L * S^2 * H).

We implemented the "bert-base-uncased" version, which means our factors were: Maximum Sequence Length of 256 (S), 12 Layers (L), and a Hidden Size of 768 (H). This configuration led to a complex model, even with a GPU it took several minutes to train 4 Epochs.

Environment configuration

We primarily used Python and Google Colab to develop, test and expand the code. Python was an obvious decision for its well-known for its robustness in data science and machine learning tasks. The presence of a wide range of libraries and frameworks, including Pandas for data manipulation, Scikit-Learn for machine learning, and PyTorch in conjunction with Hugging Face's Transformers for deep learning applications, greatly enhanced the speed of development and testing processes.

We chose Google Colab for its generous free powerful computing resources (GPUs and TPUs) in a cloud-based environment. This was crucial when it came to training BERT. Colab notebooks' interactive features allowed for quick feedback and revisions, making it ideal for fine-tuning models and debugging. Furthermore, by using Google Colab, we ensured that our experimental setups are easily reproducible by anyone with internet access, as all dependencies and the execution environment are standardized and maintained by the platform.

Benchmarking

For both the Random Forest models and the BERT Transformer, we implemented the following metrics to evaluate the effectiveness of our models in sentiment classification:

- Accuracy
- Classification Report
    - Precision, Recall, F1-Score
- Confusion matrix

We also implemented testing for both models by data splitting the dataset into training and testing sets, ensuring that the models were evaluated on unseen data to mimic real-world applicability and avoid overfitting. Typically, we used an 80-20 split.

4. RESULTS DISCUSSION

Random Forest models

As seen in the plots below, both models demonstrated similar overall accuracy, with a marginal improvement seen in the model where lemmatization was performed first. The most notable differences lie in the slightly improved recall for the negative ratings in, again, the lemmatization-first approach, suggesting that this combination was slightly better at capturing the nuances of negative reviews.

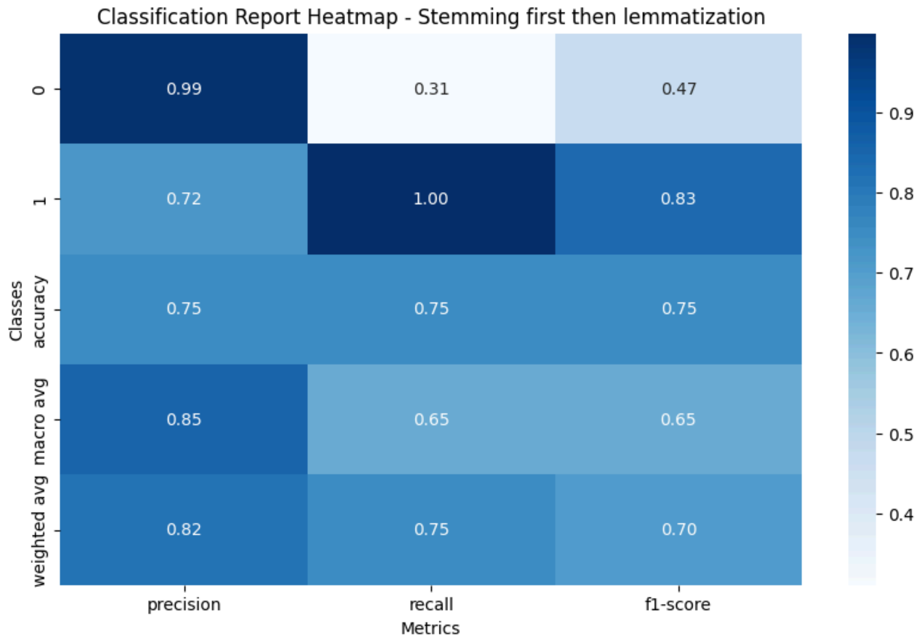

Figure 1. Classification Report of Stemming first, then lemmatization

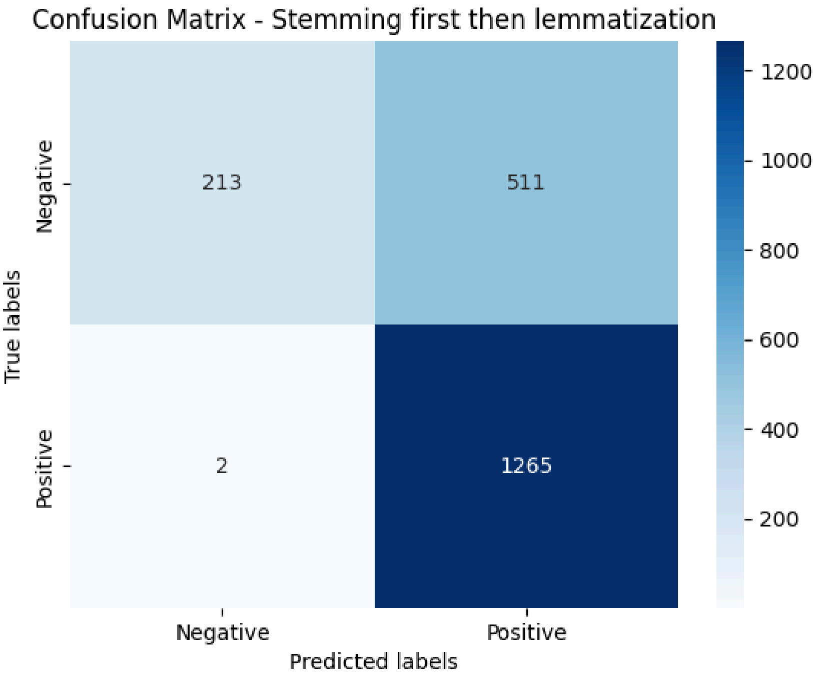

Figure 2. Confusion Matrix of Stemming first, then lemmatization

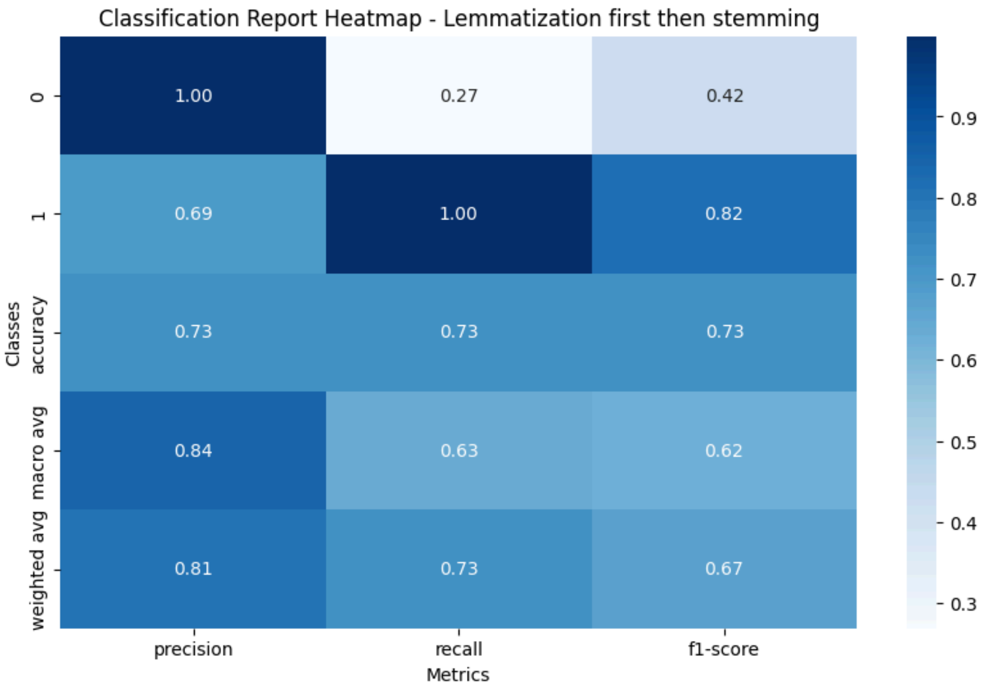

Figure 3. Classification Report of Lemmatization first, then stemming

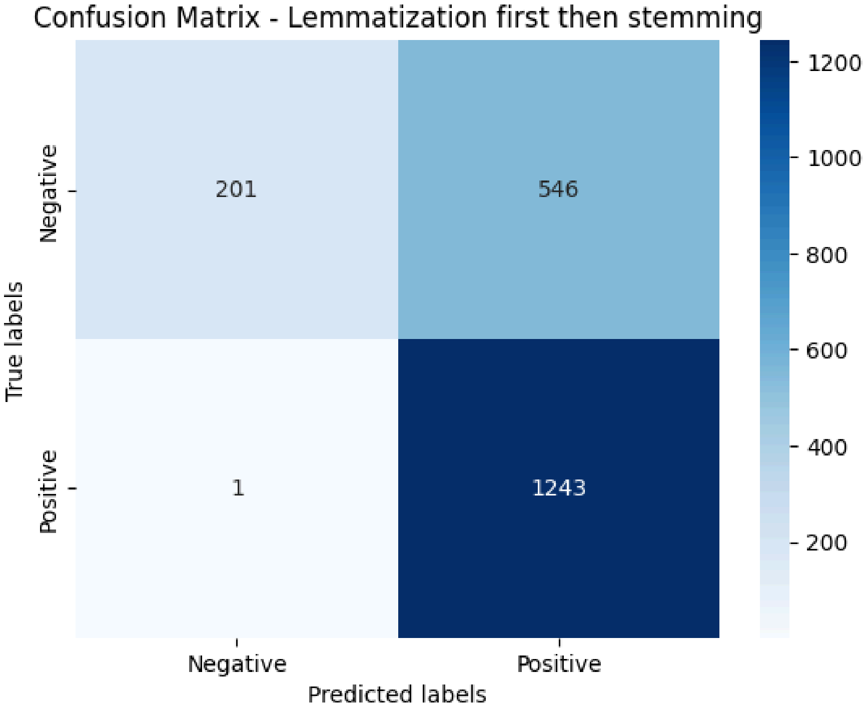

Figure 4. Confusion Matrix of Lemmatization first, then stemming

For negative reviews, the precision remained extremely high for both models, however, the low recall indicates that many were missed, confirming our initial observations of the imbalanced number of reviews. As for the positive reviews, the almost identical performance between the two models suggests that the effect of the preprocessing sequence variation wasn't too impactful in this case.

Overall, the marginal differences in model performance metrics suggest that while both preprocessing pipelines are effective, lemmatization followed by stemming offers a slight advantage. This could be attributed to the less aggressive reduction of word forms when lemmatization is applied first, potentially preserving more semantic information that is beneficial for classification.

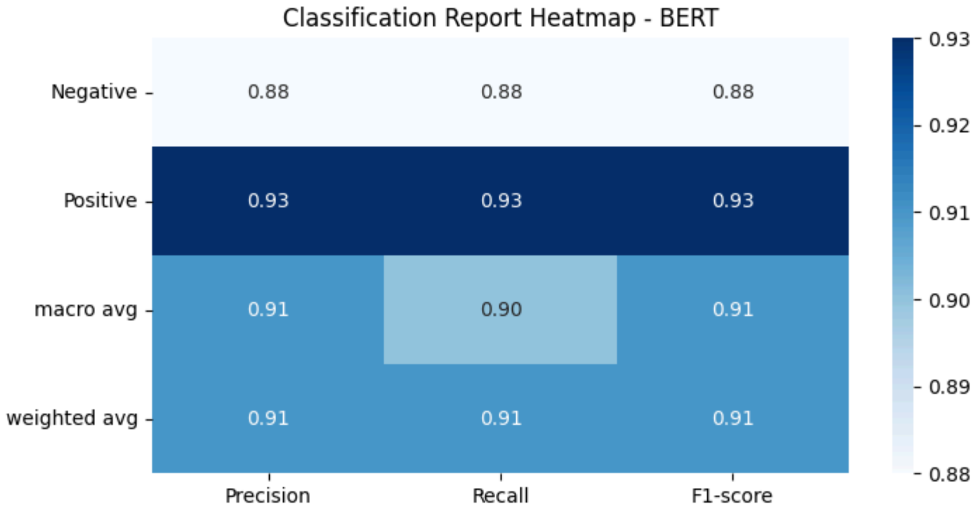

Figure 6. Classification Report of BERT

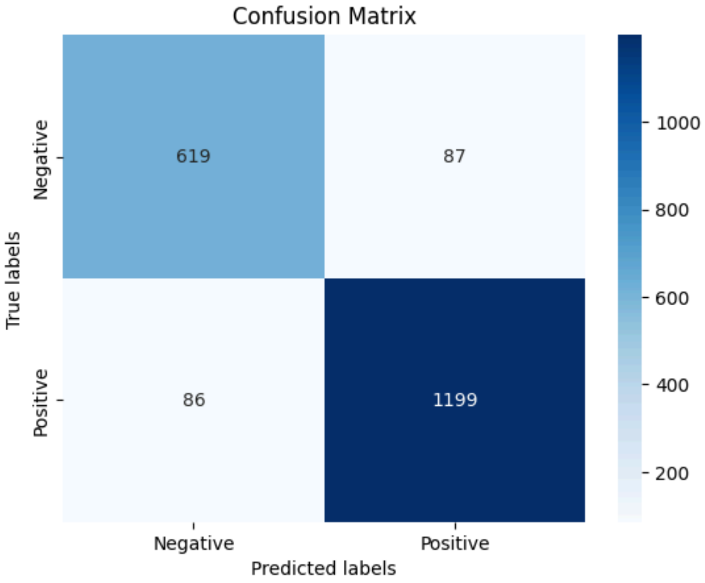

Figure 6. Confusion Matrix of BERT

As seen above, these are the metric evaluations for the BERT transformer. It is quite clear from the results that the BERT model outshines the RF models due to three key factors: contextual understanding, preprocessing, and generalization.

Firstly, BERT’s architecture allows a deeper comprehension of context and linguistic subtleties compared to RF models, which is particularly essential in sentiment analysis due to the potential for significant shifts in meaning depending on context. Secondly, BERT uses a distinct tokenization technique that effectively maintains semantic coherence, surpassing the effectiveness of stemming and lemmatization techniques used in RF models. Thirdly, the BERT transformer architecture exhibits superior generalization capabilities when applied to text data, thereby enhancing its proficiency in effectively handling the diverse characteristics of restaurant reviews.

All in all, the BERT transformer model exhibited exceptional accuracy and well-balanced classification for both classes, surpassing the RF models. BERT's ability to effectively handle linguistic context and subtle nuances, combined with its sophisticated neural network architecture, positions it as a more appropriate option for intricate NLP tasks like sentiment analysis on restaurant reviews. Which implies that deep learning models such as BERT are more desirable when dealing with text data that demands a nuanced comprehension. However, it’s important to emphasize that the findings were derived from a single dataset and specific context; performance outcomes may differ based on the input data and objectives of the task.

5. FUTURE WORK

Building upon the achievements obtained through the use of BERT Transformer and Random Forest models, several avenues for further exploration and improvement present themselves. Firstly, fine-tuning the BERT model with domain-specific data or restaurant-specific embeddings could enhance its performance in capturing nuanced sentiments and domain-specific context. Moreover, exploring ensemble methods that combine the strengths of BERT Transformer and Random Forest models could potentially yield even better results, leveraging the complementary aspects of both approaches. Additionally, investigating techniques for model interpretation and explainability, especially for complex deep learning models like BERT, would provide valuable insights into the decision-making process and boost trust in the results generated. Furthermore, conducting extensive cross-validation and sensitivity analysis across various hyperparameters and training configurations would offer robustness assessments and insights into model stability. Lastly, deploying the trained models in real-world restaurant review platforms or business environments and evaluating their performance in practical settings would provide valuable feedback for refinement and optimization, ultimately enhancing their usability and impact in real-world scenarios.

5. CONCLUSION

This study demonstrated that incorporating user-generated metadata significantly enhances the accuracy and robustness of sentiment analysis models for classifying restaurant reviews. Our comparison of the traditional Random Forest and advanced BERT Transformer models showed that context-aware deep learning technologies, like BERT, outperform simpler models in complex sentiment analysis tasks. These findings suggest a promising direction for future research in applying sophisticated NLP techniques to improve business insights and customer satisfaction. As the field progresses, further advancements in machine learning could lead to even more precise and actionable outcomes for the restaurant industry.

Overall, the study not only supports but also emphasizes the transition towards more complex, context-aware systems in the processing of natural language for business analytics. This shift promises to enhance the reliability of automated sentiment assessment, paving the way for more nuanced and actionable insights into consumer behavior.

ACKNOWLEDGMENTS

We thank our teacher Prof. Suzan T. S. Awinat for her feedback which proved very insightful and helpful towards the development of this work.

REFERENCES

[1] *The effects of online review platforms on restaurant revenue, Consumer Learning, and Welfare,management science - X-MOL*. X. (2022, February 1). https://www.x-mol.com/paper/1534549102819086336?recommendPaper=1356729489131159552

[2] Horev, R. (2018, November 17). *Bert explained: State of the art language model for NLP*. Medium. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[3] Sentiment Analysis and Classification of Restaurant Reviews using Machine Learning. IEEE Xplore. (n.d.). https://ieeexplore.ieee.org/Xplore/home.jsp

[4] Sentiment Analysis on Food Review using Machine Learning Approach. IEEE Xplore. (n.d.-b). https://ieeexplore.ieee.org/Xplore/home.jsp

[5] Alamoudi, E. S., & Alghamdi, N. S. (2021). Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. Journal of Decision Systems, 30(2–3), 259–281. https://doi.org/10.1080/12460125.2020.1864106

[6] Yu, B., Zhou, J., Zhang, Y., & Cao, Y. (2017, September 20). Identifying restaurant features via sentiment analysis on yelp reviews. arXiv.org. https://arxiv.org/abs/1709.08698

[7] “Transformers: State-of-the-art Natural Language Processing for Pytorch and TensorFlow 2.0.” Hugging Face. https://huggingface.co/transformers/

[8] Wolf, Thomas, et al. “Huggingface's Transformers: State-of-the-art Natural Language Processing.” ArXiv, 2020. https://arxiv.org/abs/1910.03771