A Deep Dive into User Reviews and Machine Learning on Yelp Dataset

Tanvir Hasan
Depertment of ITEE
University of Oulu
Oulu, Finland
tanvir.hasan@student.oulu.fi

Abstract—**Using a variety of libraries, such as TextBlob and VADAR, I performed a thorough analysis of sentiment in user reviews for this research project. To predict sentiment, I simultaneously used different techniques like Averaged Embeddings with logistic regression and Bag of Words (BoW) with logistic regression. Using Pearson correlation as a metric, the study investigates the relationship between user ratings and the feelings predicted by TextBlob, VADAR, BoW, Averaged Embeddings, and logistic regression. In addition, I compared the sentiment distributions of the expected sentiments from TextBlob and VADAR, showing the matching extents in positive, negative, and neutral circumstances with a bar plot. WordCloud are used to visually depict individual emotion cases, especially positive sentiment. The study explores important characteristics seen in user reviews, with an emphasis on price, cuisine, location, and service. The Parts of Speech (POS) tagger was used to extract adverb and adjective phrases that were associated with each attribute. I looked into Pearson correlations for positive and negative situations and determined the percentage of reviews that were written correctly. The development of a machine-learning model to predict user ratings from reviews signified the study's culmination, and I presented the model's results. Notably, I indicate an unresolved issue statement and disclose several constraints, including not using the SentiStrength library and the general inquirer corpus, which adds to the transparency and integrity of my work.**

Keywords—restaurant review, NLP, machine learning

I. Introduction

Analyzing user-generated content—especially online reviews—has become a crucial component of data analytics in today's digital world. This study explores sentiment analysis by sifting through a carefully selected portion of Yelp's massive collection of user data, reviews, and business information. Yelp, well known for enabling user-generated reviews and suggestions, offers a comprehensive dataset with a variety of customer voices from different businesses and organizations.

The main goal is to find latent insights that are hidden within this large dataset. Using a variety of techniques and instruments, my goal is to reveal the feelings that are conveyed in review papers. Sentiment analysis is a fundamental task of this investigation that is carried out with the aid of libraries like TextBlob and VADER. Sentiments are divided into three categories by this classification process: "positive," "negative," and "neutral." Beyond basic sentiment classification, though, the study aims to evaluate the relationship between user ratings and sentiment predictions made using various methodologies. Along with logistic regression, these include Averaged Embeddings and Bag of Words (BoW). This work is essential to understanding the connection between sentiments that are judged by humans and sentiments that are inferred by machines. It also delves into the fascinating field of sentiment distribution comparison across several libraries, revealing the subtleties that go into sentiment classifications. WordClouds are one innovative way to visually represent the terms that are most frequently used in reviews sorted by sentiment.

Turning away from sentiment analysis, this work delves into attribute extraction, emphasizing important aspects like price, food, location, and service. By carefully examining reviews, the study finds adjective and adverb phrases linked to these characteristics, providing businesses with priceless insights into what customers think. The study doesn't end there; it also looks at the syntax used in reviews, setting out to discover the relationship between clear and concise language and user satisfaction. A comprehensive analysis of linguistic accuracy is conducted, assessing the proportion of factually presented material in reviews and how it relates to sentiment. Ultimately, the research delves into the field of machine learning by creating and examining models for text sentiment classification to automate sentiment assessment. Models such as Linear Support Vector Classification (Linear SVC), Random Forest, Multinomial Naive Bayes, Logistic Regression, and Decision Tree are trained using rigorous data cleaning and feature engineering. Their performance is improved via hyperparameter adjustments, and the outcomes are provided for a thorough comprehension of their effectiveness in sentiment classification.

Sentiment analysis, linguistic quality assessment, and machine learning are all included in this exhaustive investigation, which has been carefully designed to uncover a wealth of information about Yelp reviews. With a thorough grasp of consumer sentiment, this journey creates a path for organizations to improve their offers and boost customer satisfaction. Sentiments are an essential part of humanity. Sentiment substantially influences human decision-making and increases our communication skills. Over the first few years, academics have made tremendous progress toward recognizing sentiment automatically. Sentiment from the text has become one of the hottest pursuits among scholars. It is commonly interchangeably used yet different and plays key roles across various fields, including business, healthcare, education and many more. The relevance and diverse uses of sentiment and emotion analysis offer insights into the hurdles faced by researchers in the development of effective approaches. A study [1] provided numerous strategies on how to extract significant words utilizing semantic approaches such as POS tagger by detecting nouns, verbs, adverbs, and adjectives of phrases. They also presented a Chi-square approach to eliminate the weak semantic features and achieved an improvement in the emotion recognition rate using the ISEAR dataset. Ezhilarasi & Minu [2] proposed a method for automatically detecting emotions in the text by developing an emotion ontology in English using WordNet. In another study [3], Shafana et al. evaluated the sentiments of people in South Asian countries regarding the online education system during the COVID-19 outbreak using Twitter data. Their findings suggested a mainly positive attitude toward the online education system in the region. Meanwhile, Putri and Kusumaningrum [4] proposed a way to extract overarching themes from tourist reviews, dividing them into positive and negative attitudes in their research on the Indonesian tourism business. Notably, sentiment analysis based on interview data has been a relatively untapped subject. Parmar et al. [5] addressed this issue by conducting one-on-one interviews with Indian millennials and managers from varied sectors. Their research indicated that sentiment variances are driven by criteria such as gender, industry, industry experience, and millennial status. The rise in natural language processing can be linked to improvements in machine learning, artificial intelligence, and deep learning. While teaching machines to understand human language remains a tough task, it is increasingly doable. The adoption of NLP has delivered considerable benefits to many educational institutions. Nevertheless, the integration of AI technology within numerous enterprises in worldwide has not undergone considerable modifications. In a recent examination [6], efforts were made to anticipate emotion in tweets produced by users of Pathao, a popular bike-sharing firm, using a range of machine learning algorithms. The study found Support Vector Machine (SVM) as the top-performing model, outperforming NaiveBayes and Logistic Regression. Additionally, an analysis by Arafat et al. [7] analyzed data from English newspaper headlines to discover patterns. They detected a higher prevalence of negative sentiment phrases compared to positive ones, with roughly 400 instances in 2018 and 2019. Their research provides insights into the broader socio-political and cricket-related landscape of Bangladesh during that period.

Notable advancements have been made in the field of sentiment analysis research in the context of restaurant reviews, with a particular emphasis on the Yelp dataset. Liu [8] carried out a thorough analysis, contrasting the ability of several deep learning and machine learning models to predict user sentiment. Significantly, machine learning models benefited from improvements in text representation and normalization, while deep learning models performed better when maximum length constraints and pre-trained word embeddings were used. The macro F1 score reveals that surprisingly, less complex models—such as Support Vector Machine and Logistic Regression—performed better than their more complex counterparts, like Gradient Boosting, LSTM, and BERT. In another study, Alamoudi & Alghamdi [9] used the ALBERT model to classify sentiment in binary and ternary categories with an impressive 98.30% accuracy. Their novel unsupervised method of aspect-level sentiment classification, which combined pre-trained language models such as GloVe with semantic similarity, showed an impressive 83.04% accuracy. Additionally, Hemalatha and Ramathmika [10] used machine learning methods from the NLTK toolkit to analyze the sentiment of Yelp restaurant reviews. The efficacy of natural language processing in understanding and classifying attitudes in textual evaluations was highlighted by their methodical examination and comparison of algorithmic efficiency and confidence.

II. METHODOLOGY

A. Dataset Description & Ethical Consideration

The foundation of this research is the examination of the Yelp business dataset [11], which is a vast collection of user-generated reviews, ratings, and other information linked to a wide range of companies. With information on more than 150,000 companies, 6.9 million reviews, and 200,000 images, the original dataset offers a comprehensive picture of customer experiences. A specific subset of 20,000 reviews was carefully selected for this study, with particular emphasis on four key columns: review content, user ratings, URL, and date. Every entry in the collection provides qualitative comments and insights into user satisfaction levels, representing a distinct user perspective. The fact that the dataset is available for academic use adds to the transparency of the study and promotes cooperation between researchers and students. The research hopes to provide nuanced insights into customer feelings, preferences, and the variables driving online reviews by utilizing this publicly available dataset.

With each star rating denoting a different degree of user satisfaction, the dataset under examination figure 1 shows a distribution of user ratings for a certain collection of items. The data indicates that a significant proportion of users hold extremely positive feelings, as evidenced by the 55% allocation of 5-star ratings. Furthermore, 22% of people have given their reviews a 4-star rating, which adds to the general upward trend in user comments. A significant fraction of the dataset, however, falls into the lower rating range; 10% gives 3 stars, and 6% gives 1 and 2 stars. The exploration of the dataset was made free from copyright restrictions (CC0: Public Domain) license.

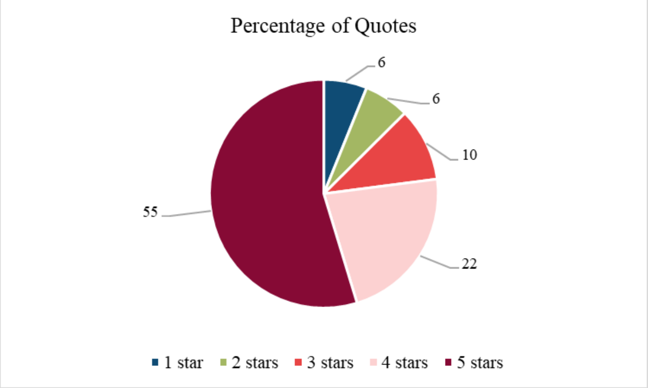

Figure 1. Dataset description

B. Content Analysis

With a primary focus on the Yelp dataset, this study uses a multifaceted methodology to do sentiment analysis on restaurant reviews. The method is implemented in a step-by-step manner, starting with importing the dataset into a Pandas DataFrame and utilizing the TextBlob package to calculate sentiment polarity scores. After that, based on the polarity score, sentiments are categorized as either positive, negative, or neutral, which serves as the basis for additional research. Many sentiment analysis techniques are used, combining text preprocessing, feature engineering, and machine learning model predictions. The analysis begins by preparing the text data. It applies denoising techniques to remove HTML tags and expand contractions, enhancing the quality of the text for analysis. Text normalization steps, such as converting text to lowercase, removing punctuation, and optionally removing stopwords, are performed. Tokenization is carried out using the Natural Language Toolkit (NLTK). After that, I explored the effectiveness of the Bag-of-Words (BoW) and Averaged Word Embeddings approaches in predicting sentiment using the Logistic regression model. When working on the BoW technique I split the dataset into training and testing sets and trained a Logistic Regression model on the training data. Additionally, the code employs spaCy's 'en_core_web_md' model to create averaged word embeddings for each review text. Pearson correlation analyses between TextBlob sentiment scores and user ratings are also included in the study, along with a comparison with the VADER sentiment analysis tool and the creation of word clouds and visualizations. I have used the WordCloud python library to generate word clouds for different categories of reviews based on user ratings. Word clouds were created for 'positive,' 'neutral,' and 'negative' reviews, displaying the most frequently occurring words in each category.

The methodology also explores the evaluation of language quality, the extraction of attribute-related words, and the creation and assessment of machine learning models for sentiment categorization. I also extracted adjective and adverb phrases associated with specific attributes within a dataset of reviews. It begins by preparing the environment, including the download of essential NLTK resources for tokenization and stopwords. A set of keywords is defined for attributes like 'Cost/Price,' 'Food,' 'Location,' and 'Service.' Subsequently, a custom function is created to tokenize and tag words in each review text, identifying and extracting relevant adjective and adverb phrases linked to the specified attributes. I was also interested in exploring gauging the linguistic precision of reviews and their potential impact on user satisfaction, shedding light on the connection between language and sentiment in the context of user feedback To achieve this, first I ensured the availability of WordNet data through NLTK and defines a function to assesses the percentage of correctly worded content in each review. It tokenizes the text, identifies valid words with WordNet synsets, and calculates the proportion of valid words to the total number of words in the review. Subsequently, this function is applied to the review texts in the dataset. To gain insights into the correlation between the quality of wording and user sentiment, I computed Pearson correlation coefficients for both negative and positive reviews, which are categorized based on user ratings. Finally, build multiple machine learning models to predict user ratings. The process starts by removing missing values and encoding the user's rating. The data is then split into training and testing sets. Text data is converted into numerical features using TF-IDF vectorization. Several machine learning models are trained and evaluated on the dataset such as Linear Support Vector Classification (Linear SVC), Random Forest, Multinomial Naive Bayes, Logistic Regression, and Decision Tree. Moreover, I performed hyperparameter tuning for each model using GridSearchCV, optimizing their performance. Finally, the accuracy, classification reports, and confusion matrices for each model to assess their effectiveness in sentiment classification, providing a comprehensive approach to text sentiment analysis. This thorough and organized approach guarantees an in-depth investigation of sentiment nuances in restaurant reviews, providing insightful information about customer preferences and satisfaction.

III. Results

A. Identifying optimal business phases for Yelp

In my study, I was interested in exploring which year of phase was good for Yelp's business according to online reviews from the customers. To achieve this, I have used the Python TextBlob library to calculate the polarity of the reviews. My finding suggests that the years 2018 to 2020 were the most reputed for the company as there are all positive sentiments from the reviews.

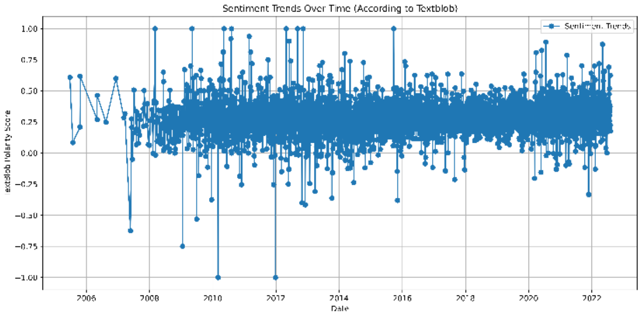

Figure 2: Sentiment trends over time

B. Effectiveness of the BOW and Averaged Word Embeddings approaches in predicting sentiment

In my dataset, I applied two more alternative sentiment analyzer tools (bag-of-words + logistic regression) and (word embedding + logistic regression) techniques. I wish to explore the effectiveness of the two models. My findings suggested that the BOW technique demonstrates remarkable performance in predicting sentiment. It achieves great accuracy, precision, recall, and F1 Score, all of which are around 96% illustrated in Figure 3. This implies that BOW can efficiently classify reviews into 'positive,' 'negative,' and 'neutral' categories, correlating closely with the provided numeric ratings. While the Averaged Word Embeddings approach still generates respectable sentiment prediction results, it falls behind the BOW approach. The accuracy, precision, recall, and F1 Score for Averaged Word Embeddings are roughly 85%. Although these results are lower compared to BOW, they are still decent and can efficiently divide reviews into feelings.

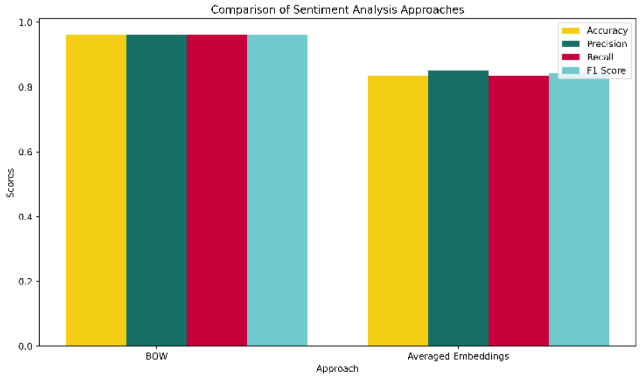

Figure 3. Comparison of sentiment prediction using BOW and Average Embeddings

C. Model Effectiveness and User Sentiments

In my study, I wanted to explore whether there is any statistical relationship between user ratings and predicted ratings from TextBlob, BOW + Logistic Model Predicted Rating, and Average Embedding + Logistic Regression. In Table 1 I have represented the findings of the study. It highlights that the BOW + Logistic model exhibits the strongest correlation with user ratings, closely reflecting the sentiments expressed in user reviews. The Averaged Word Embeddings + Logistic regression model also shows a substantial correlation, indicating its effectiveness in sentiment prediction. However, TextBlob, while moderately correlated, appears to be less effective in capturing the nuances of user sentiments compared to the other two models. These findings emphasize the importance of selecting an appropriate sentiment analysis model when analyzing user-generated content to gain more accurate insights into user sentiment and preferences.

TABLE I. CORRELATION AND SIGNIFICANCE SCORES FOR USER RATINGS AND PREDICTED RATINGS USING DIFFERENT MODELS.

	Textblob	**BOW+Logistic**	**AverageEmbedding+Logistic**
Pearson Corelation score	0.520	0.939	0.733
P-value	0.0001	0.0001	0.0001

D. Comparison of sentiment distributions predicted from TextBlob and VADAR

I want to interpret the user's rating from the sentiment polarity perspective. For this purpose, I have assumed user rating 4-5 belongs to positive sentiment, 3 for neutral and 1-2 for negative sentiment. In the following section, I have presented a sentiment distribution comparison between Textblob-predicted sentiment and VADAR-predicted sentiment in Figure 3. The finding shows that the distributions of positive and negative feelings provided by TextBlob and VADER are remarkably comparable. The permissible range of modest deviations in percentages can be explained by the intrinsic disparities in the implementation of different sentiment analysis techniques. Notably, positive and negative sentiment distributions are the dominating categories, suggesting that TextBlob and VADER are equally effective at capturing strongly held attitude polarities. There is a more pronounced disparity in the percentage of neutral thoughts; VADER classifies a slightly higher proportion of reviews as neutral. This implies that when classifying reviews as neutral, VADER might be more cautious. But even in this instance, the change is not noticeable.

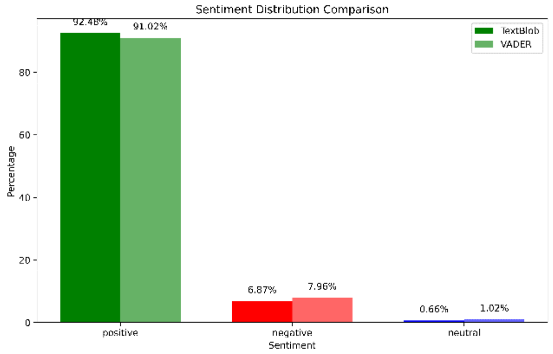

Figure 4. Sentiment distribution comparison between TextBlob and VADAR.

E. Words associated with individual sentiment through WordCloud

Understanding the words associated with individual sentiment not only offers valuable insights into human emotional expression but also serves as a foundation for the development of sentiment classification models. The finding has revealed fascinating insights into the word's association with the individual sentiment. Before analyzing WordCloud, I preprocessed the data including denoise text, removed ASCII, converted all characters to lowercase, removed punctuation, replaced all integers, removed stopwords, and stem words, lemmatize verbs, normalized the text, and tokenized them.

WordCloud of positive and negative sentiment in figure 5 illustrates the word associated with positive reviews (user rating 4,5) and figure 6 shows the negative reviews (user rating 1,2). It is shown that ice cream, good, flavour, love, sweet, place, doughnut and cookies are the most frequent words for both the sentiments. This means there is no such difference between the word but there is a difference in terms of frequency. The reason may be because of the stop words removal from the text.

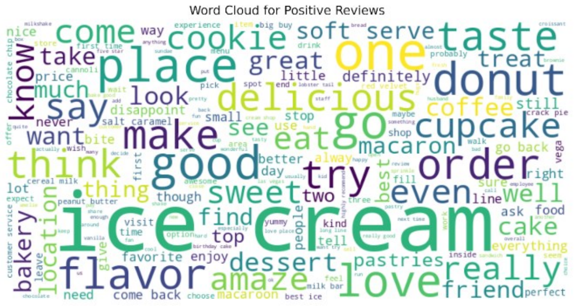

Figure 5. WordCloud representation of users' positive

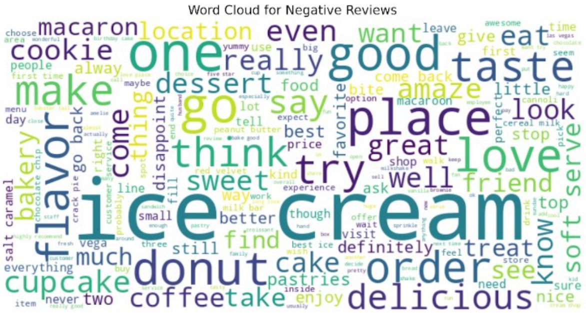

Figure 6. WordCloud representation of users' negative sentiment

F. Correlation between the percentage of correct wording in reviews

In my analysis, I studied the association between the percentage of right phrasing in reviews and the related ratings, concentrating on both negative and positive reviews. For unfavorable evaluations, I discovered a statistically significant but extremely weak negative connection (Pearson connection = -0.0401). This implies that if the percentage of proper wording in unfavorable reviews grows slightly, the related negative scores tend to decline slightly. However, the amplitude of this association is low, indicating that it is not practically significant. On the other hand, for favorable evaluations, the correlation was not statistically significant (Pearson Correlation = 0.0084). This means that there is essentially no linear link between the percentage of proper phrasing and good evaluations, and the apparent association is likely due to random chance. These data underline that the choice of words and precise wording in reviews has a limited impact on ratings, particularly for favorable evaluations. The significance tests (P-values) verify the conclusions, with a P-value of 0.0451 for negative reviews (statistically significant) and a P-value of 0.5003 for good reviews (not statistically significant).

TABLE II. CORRELATION BETWEEN CORRECT WORDING IN REVIEWS.

	Negative reviews	**Positive reviews**
Pearson Corelation score	-0.0401	0.0084
P-value	0.0451	0.5003

G. *Machine learning model to predict ratings*

This work outlines significant advancements in feature engineering and model refining toward the goal of building an effective machine-learning model for the prediction of user ratings, leading to increased predictive effectiveness. A core component of the process is feature engineering, in which prominent features are extracted from textual data by carefully using the Bag-of-Words (BoW) methodology. BoW is especially well-suited for text analysis since it records the frequency distribution of words in user evaluations, making it easier to identify observable relationships between certain lexical items and user ratings. Furthermore, a crucial feature selection procedure is carried out, utilizing the chi-squared test to identify the top 1000 qualities that exhibit the strongest correlations with the goal variable—user ratings. The model's predictive precision is much enhanced by this careful feature curation. Based on a logistic regression framework, the model is subjected to fine-grained hyperparameter optimization using Bayesian optimization. One of the most important steps in optimizing the model's prediction performance is this careful hyperparameter tuning. A thorough examination of the model's robustness and generalizability is carried out using a 10-fold cross-validation process with the optimized hyperparameters, providing a thorough assessment of predictive accuracy on untested data—an essential step in determining the model's dependability and relevance.

TABLE III. PERFORMANCE OF THE LOGISTIC REGRESSION MODEL

Rating	**Precision (%)**	**Recall (%)**	**F1 score (%)**
1 star	58	50	54
2 stars	40	31	35
3 stars	43	36	39
4 stars	51	34	41
5 stars	74	90	81
Accuracy	65		
Cross-validation score	66		

My thorough approach to creating a user rating prediction model has shown encouraging outcomes. After rigorous feature engineering, including the deployment of Bag-of-Words (BoW), I further refined the model through feature selection and hyperparameter tweaking using Bayesian Optimization. The hyperparameter selection led to a Best C value of 0.3571, and 10-fold cross-validation consistently provided an average score of 0.66, confirming the model's reliability. In test set evaluation, the model attained an accuracy of 65%, signifying its competence to effectively anticipate user ratings.

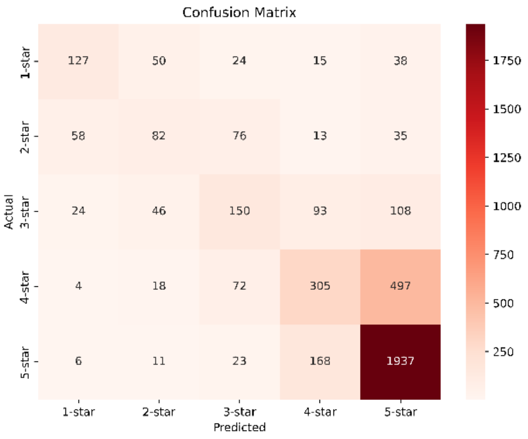

Figure 7. Confusion matrix of the model

Notably, the Classification Report shows the model's remarkable accuracy in predicting 5-star ratings, sporting an impressive F1 score of 0.81. However, the model's efficacy differs across different rating categories, as evidenced by differing F1 scores for 1-star, 2-star, 3-star, and 4-star ratings. While my efforts in feature engineering, selection, and hyperparameter tuning have produced a strong model, future work is geared toward boosting its performance for other rating categories. These findings give useful insights and constitute the basis for further modifications and optimizations.

IV. DISCUSSION

My study highlights the superior performance of logistic regression over more complex deep learning models such as gradient boosting and neural networks in sentiment and user rating prediction from textual data. This is in line with the findings of a related study conducted by Liu [8], who noted that logistic regression performed well in sentiment analysis while not predicting user ratings as this study did. Notably, a cutting-edge work by Alamoudi & Alghamdi [9], focusing on two and three-class sentiment categorizations, obtained an astounding 98.30% accuracy in multi-class sentiment prediction utilizing their ALBERT model. Furthermore, the BRET model was introduced by the Google team in a groundbreaking language model study [12], which also examined feature-based techniques with BERT, pre-training tasks, and the implications of model size on performance. Even though my study did not address these issues, they offer fascinating directions for future investigation. This comparison places my study in the context of the changing field of sentiment analysis research and highlights the relative advantages of logistic regression. It also provides possible directions for future research.

V. LIMITATIONS & FUTURE WORKS

Notwithstanding the extensive scope of this study, it is important to recognize its limitations. Our rating prediction model could not perform well in terms of 1-star, 2-star, 3-star, and 4-star prediction as there was less data from 5-star reviews. It means that I need more data to produce a robust machine-learning model. The performance could be increased if I use pre-defined models.

By expanding on the knowledge gathered from this study, future research may improve and expand on a number of the areas of sentiment analysis in restaurant reviews. To capture complex contextual links and increase sentiment prediction accuracy, one line of investigation is the introduction of more sophisticated sentiment analysis models, such as transformer-based models like BERT. Furthermore, improving sentiment analysis performance may require investigating deep learning architectures other than the CNN model, perhaps combining recurrent neural networks (RNNs) or attention processes. Moreover, a thorough examination of the influence of sentiment on particular facets of the customer experience, such as pinpointing characteristics like price, cuisine, location, and assistance, may offer more sophisticated commercial insights. Finally, there may be opportunities to increase the robustness and accuracy of sentiment analysis by utilizing current developments in natural language processing, such as pre-trained language models and unsupervised learning techniques. With the purpose of better understanding user sentiments in restaurant reviews, these suggested directions hope to further the continuous development of sentiment analysis approaches and their useful applications.

VI. Conclusion

This study uses a variety of approaches and techniques to provide a comprehensive examination of user review sentiment analysis. The study examines the relationships between user ratings and sentiment prediction models, such as Averaged Embeddings, Bag of Words (BoW), TextBlob, and VADER. The study shows that the BoW approach outperforms TextBlob and Averaged Word Embeddings in sentiment prediction. Nuanced insights are offered via sentiment distribution comparisons, WordCloud representations, and the investigation of the relationship between language precision and ratings. The creation of a machine-learning model highlights the need for more improvement across rating categories while showcasing encouraging outcomes for 5-star ratings. All things considered, this study provides insightful information that will direct future developments in sentiment analysis algorithms and increase prediction accuracy for various user rating scenarios. This study offers a solid framework for further research, suggesting avenues for enhancing sentiment analysis models and raising predicted accuracy in a range of user rating categories

Reference

[1] L. Singh, S. Singh, and N. Aggarwal, Two-stage text feature selection method for human emotion recognition. Singapore: Springer Singapore, 2019, p. 531–538.

[2] R. Ezhilarasi and R. I. Minu, “Automatic emotion recognition and classification,” Procedia Eng., vol. 38, pp. 21–26, 2012.

[3] J. Praveen Gujjar and H. Prasanna Kumar, “Sentiment analysis: Textblob for decision making,” Int. J. Sci. Res. Eng. Trends, vol. 7, no. 2, pp. 1097–1099, 2021.

[4] I. R. Putri and R. Kusumaningrum, “Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia,” J. Phys. Conf. Ser., vol. 801, p. 012073, 2017.

[5] M. Parmar, B. Maturi, J. M. Dutt, and H. Phate, “Sentiment analysis on interview transcripts: An application of NLP for quantitative analysis,” in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.

[6] M. I. Sajib, S. Mahmud Shargo, and M. A. Hossain, “Comparison of the efficiency of Machine Learning algorithms on Twitter Sentiment Analysis of Pathao,” in 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019

[7] A. Hossain, M. Karimuzzaman, M. M. Hossain, and A. Rahman, “Text mining and sentiment analysis of newspaper headlines,” Information (Basel), vol. 12, no. 10, p. 414, 2021.

[8] S. Liu, “Sentiment analysis of yelp reviews: A comparison of techniques and models,” arXiv.org, 2020.

[9] E. S. Alamoudi and N. S. Alghamdi, “Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings,” J. Decis. Syst., vol. 30, no. 2–3, pp. 259–281, 2021.

[10] Hemalatha and R. Ramathmika, “Sentiment analysis of yelp reviews by machine learning,” in 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019.

[11] M. F. Alam, “Yelp Restaurant Reviews.” 31-Dec-2022.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of the 2019 Conference of the North, 2019.