**Yelp Review Classification Using NLP**

**Milestone 4:** Project – Implement Embedding Methods

**Student:** Urvashi Dube

+12368633298

dube.u @northeastern.edu

**Percentage of Effort Contributed by Student :** 100%

**Signature of Student :\_\_\_** A close up of a word

Description automatically generated\_\_\_\_

**Submission Date:\_\_\_\_\_\_\_** 5th November, 2023\_**\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Introduction**

This project focuses on analyzing Yelp reviews, categorizing them into relevant labels, and building a predictive model to streamline the review categorization process. The main objective is to develop a powerful deep learning model that can automatically assign meaningful labels to Yelp reviews. This has significant implications as it can provide valuable insights from user-generated content, enabling businesses to make data-driven decisions and offer users more relevant, categorized reviews.

In the realm of deep learning, feature extraction and embedding techniques are crucial, especially for natural language processing tasks like text classification. These methods transform unstructured text data into numerical representations, allowing deep learning models to effectively understand patterns, relationships, and semantics within the text. This is vital for accurately categorizing Yelp reviews by context, location, and themes. The choice of embedding method, whether it's Word2Vec, GloVe or Fasttext, significantly influences the model's performance. Selecting the right embedding method and seamlessly integrating it with deep learning architectures is pivotal to achieving the project's objectives and improving Yelp review categorization quality.

**Problem Statement**

This project aims to improve the Yelp review categorization process by creating a powerful deep learning model, addressing the challenge of manually assigning categories to 500 reviews and using deep learning methods to automatically categorize the remaining and future review. By automating this process, the project seeks to enhance the accuracy and efficiency of review categorization, benefiting businesses and users with valuable insights from Yelp's dynamic user-generated content.

**Objectives**

* Conduct custom Named Entity Recognition (NER) analysis and data preprocessing to enhance the quality and informativeness of the dataset.
* Explore various embedding techniques, such as Word2Vec, GloVe, Fasttext to convert raw text data into numerical representations.
* Develop a deep learning model capable of automatically assigning relevant labels or categories to Yelp reviews.
* Implement different deep learning architectures, including Long Short-Term Memory (LSTM) networks, to build and fine-tune the predictive model.
* Evaluate the model's performance through comprehensive testing, including accuracy, precision, recall, and F1-score metrics.
* Enable the model for practical use (real reviews) on the Yelp platform, streamlining the review categorization process and providing businesses with insights from user-generated content.

**Data Used**

**Dataset Overview**

The "yelp\_polarity" dataset is a comprehensive collection of Yelp reviews specifically curated for sentiment classification tasks by Xiang Zhang. It comprises a total of 560,000 Yelp reviews for training and an additional 38,000 reviews for testing. This dataset originates from the Yelp Dataset Challenge 2015. Its primary purpose is to facilitate sentiment analysis, a task where each review is categorized into one of two classes: negative or positive sentiment.

To establish these sentiment classes, reviews with ratings of 1 or 2 stars are labeled as negative (class 0), while those with 3 or 4 stars are classified as positive (class 1). The dataset is organized into two CSV files, namely "train.csv" and "test.csv," with each containing two columns: one denoting the class index (0 for negative and 1 for positive) and the other containing the actual review text. The review text is enclosed within double quotes, and any internal double quotes are escaped with double quotes as well. Due to hardware complexity barriers, I will be using 100000 rows of train data and 10000 rows of test data for analysis and prediction. My goal is to utilize NER (Named Entity Recognition) analysis and deep learning techniques to process the dataset effectively. By doing so, I aim to train the system to predict custom labels for these reviews. This will significantly improve the efficiency of review categorization, ultimately contributing to the growth of businesses.

**Data Dependency**

* Effectively handling these dependencies will guarantee the reliable and precise classification of reviews.
* Class Imbalance Dependency: A balanced distribution of negative and positive sentiment reviews is necessary for unbiased model training.
* Rating-Sentiment Mapping Dependency: Accurate mapping of Yelp ratings to sentiment classes (1, 2 stars as negative; 3, 4 stars as positive).
* Textual Quality Dependency: High-quality, clean review text is crucial for accurate sentiment analysis.
* Feature Engineering Dependency: Choice of text features based on review text characteristics is crucial for better model performance.
* Evaluation Metric Selection Dependency: Selection of appropriate evaluation metrics aligned with project objectives.
* Data Preprocessing Dependency: Properly preprocess text data (tokenization, stop-word removal, normalization).
* Temporal Considerations: Considering temporal trends or seasonality in sentiments
* Training and Test Set Split Dependency: Ensuring a representative split of data for reliable model evaluation.
* Ethical Considerations: Addressing privacy, bias, and fairness in sentiment analysis to respect rights and avoid biases.

**Analysis**

**Explanatory Data Analysis**

Initially, a basic EDA was conducted to visualize the distribution of text lengths in the dataset, which is crucial for understanding the text data's characteristics as shown in Fig.1**.** Text length histograms were generated to show the distribution. This graph clearly shows that the text length of reviews in the dataset are not constant, and the distribution is rightly- skewed.

Next, a Named Entity Recognition (NER) analysis was performed using spaCy and a custom NER model. The NER analysis involved loading the spaCy model, processing text data, and rendering named entities using spaCy's "displacy" module. This was done to analyze the entities labelled by spaCy model for reference. The result of the SpaCy model is shown in Fig. 2. Additionally, the custom NER model was utilized to label 500 reviews for training data using [8].

A graph of a number of text

Description automatically generated

**Fig. 1. Text Length Distribution**

**A screenshot of a chat

Description automatically generated**

**Fig. 2. Results of SpaCy Model**

Following the NER analysis, a new spaCy model for NER was trained using the labels that were generated using [8]in .json file. The custom entities in whole data are shown in Fig. 3. These entities will surely help in better categorization of yelp reviews. The model was trained for 150 iterations, and the loss for each iteration was monitored. The trained model was saved for future use.

After undergoing 150 training iterations (epochs), there has been a noticeable reduction in the loss value. This loss value, approximately 662.6732557820573, is indicative of the degree of dissimilarity between the model's predicted entities and the true entities present in the training data. The progressive decrease in the loss value across the 150 iterations suggests that the model is continually improving its ability to predict and classify entities accurately. The last iterations are shown in Fig.4.

A computer code with text

Description automatically generated with medium confidence

**Fig. 3 Custom Entities**

A black and white striped background

Description automatically generated

**Fig. 4. Iterations of training NER Model**

Subsequently, the saved model was loaded to perform NER on a text file containing 500 rows of train data. The output of the custom labelled data is shown in Fig. 5.

A screenshot of a chat

Description automatically generated

**Fig. 5. Custom NER Labels on the text data**

In the deep EDA section, a pre-trained spaCy NER model was loaded to extract named entities from a list of text documents. The top entities and entity count for each label was computed. Visualizations and tables were created to display the top entities and their frequencies for different labels. This analysis provided insights into the most frequently occurring named entities in the text data.

The explanatory data analysis is performed in three parts overall EDA, positive EDA, and negative EDA. These parts each explain the different analysis conducted on the yelp dataset using the custom NER analysis.

Plots and tables were generated for various entity labels, both overall and separately for positive and negative reviews, offering a comprehensive understanding of the dataset.

The graphs of EDA of positive and negative reviews combined are show in Fig. 6.

A graph with blue and black bars

Description automatically generated

**Fig. 6 (a). Top 30 Organization entities and their frequencies for overall data**

A blue and black squares

Description automatically generated

**Fig. 6 (b). Top 30 Dentist entities and their frequencies for overall data**

A graph with blue and white bars

Description automatically generated

**Fig. 6 (c). Top 30 Product entities and their frequencies for overall data**

A graph with blue bars

Description automatically generated with medium confidence

**Fig. 6 (d). Top 30 Person entities and their frequencies for overall data**

A graph with blue and white bars

Description automatically generated

**Fig. 6 (e). Top 30 Time entities and their frequencies for overall data**

A bar graph with numbers and text

Description automatically generated

**Fig. 6 (f). Top 30 Restaurant entities and their frequencies for overall data**

A graph showing a number of numbers

Description automatically generated with medium confidence

**Fig. 6 (g). Top 30 Quantity entities and their frequencies for overall data**

A graph with blue and white bars

Description automatically generated

**Fig. 6 (h). Top 30 Cuisine entities and their frequencies for overall data**

A graph with blue and white bars

Description automatically generated

**Fig. 6 (i). Top 30 Prize entities and their frequencies for overall data**

A graph with blue and white bars

Description automatically generated

**Fig. 6 (j). Top 30 Location entities and their frequencies for overall data**

A graph of a service

Description automatically generated

**Fig. 6 (k). Top 30 Service entities and their frequencies for overall data**

A graph with blue and white stripes

Description automatically generated

**Fig. 6 (l). Top 30 Disease entities and their frequencies for overall data**

**Findings and Business Insights from overall reviews EDA:**

Yelp's Prominence: The term "Yelp" and its variations are frequently mentioned, indicating the prominence and brand recognition of the platform among users, with a high frequency of 171 mentions.

Popular Fast-Food Chains: Fast food chains like "McDonald's" and "Taco Bell" are highly mentioned in reviews, suggesting their popularity among customers, with frequencies of 57 and 30, respectively.

Customer Reviews of Hotels The word "hotel" appears frequently, highlighting the significance of accommodation services within the reviews, with a frequency of 548.

Diverse Restaurant Types: Different types of restaurants, including "Chinese," "Italian," "American," and "Mexican," are mentioned, showing the diverse culinary options available to customers.

Pricing Sentiments: Terms like "cheap," "expensive," and "overpriced" are used to express pricing sentiments, providing valuable insights into customer perceptions of cost.

Regional Insights: Locations like "Pittsburgh" and "Charlotte" are frequently mentioned, indicating areas of high activity and possibly strong user engagement.

Service Quality: Adjectives such as "friendly" and "great" are used to describe service quality, suggesting that positive customer service experiences

Doctor and Healthcare References: The presence of terms like "dentist" and "doctors" highlights the relevance of healthcare services within the reviews, with a frequency of 14 and 13, respectively.

Food Preferences Popular food items such as "pizza," "coffee," and "chicken" are mentioned frequently, reflecting customers prefer these items

Temporal References Time-related mentions, such as "morning," "evening," and "Sundays," provide insights into customers are likely to visit or share their experiences.

The graphs of EDA of positive reviews are show in Fig. 7.

A screenshot of a computer

Description automatically generated

**Fig. 7 (a). Top 30 Organization entities and their frequencies for positive reviews**

A graph with numbers and text

Description automatically generated with medium confidence

**Fig. 7 (b). Top 30 Dentist entities and their frequencies for positive reviews**

A graph with blue bars

Description automatically generated with medium confidence

**Fig. 7 (c). Top 30 Product entities and their frequencies for positive reviews**

A graph with blue bars

Description automatically generated with medium confidence

**Fig. 7 (d). Top 30 Person entities and their frequencies for positive reviews**

A graph with blue bars

Description automatically generated with medium confidence

**Fig. 7 (e). Top 30 Entities and their frequencies for positive reviews**

A bar graph with numbers and text

Description automatically generated

**Fig. 7 (d). Top 30 Restaurant entities and their frequencies for positive reviews**

A graph with blue and white bars

Description automatically generated

**Fig. 7 (e). Top 30 Quantity entities and their frequencies for positive reviews**

A bar graph with numbers and text

Description automatically generated

**Fig. 7 (f). Top 30 Cuisine entities and their frequencies for positive reviews**

A graph showing a number of blue and white bars

Description automatically generated with medium confidence

**Fig. 7 (g). Top 30 Price entities and their frequencies for positive reviews**

A graph showing a number of locations

Description automatically generated with medium confidence

**Fig. 7 (h). Top 30 Location entities and their frequencies for positive reviews**

A graph with blue and white bars

Description automatically generated

**Fig. 7 (i). Top 30 Service entities and their frequencies for positive reviews**

A graph of a disease

Description automatically generated with medium confidence

**Fig. 7 (j). Top 30 Disease entities and their frequencies for positive reviews**

**Findings and Business Insights from positive reviews EDA:**

Yelp's Positive Impact: The brand "Yelp" and its variations are frequently mentioned in positive reviews, indicating that users are actively engaged and sharing their positive experiences, with a high frequency of 941.

Local Business Promotion: Local businesses such as "Subway," "McDonald's," and "Taco Bell" receive substantial positive attention, suggesting that these establishments are satisfying their customers and receiving word-of-mouth marketing.

Company Acknowledgment: "Company" is a common term, demonstrating that businesses are recognized and appreciated in positive reviews, with a frequency of 101.

Dining in Paradise: The term "Paradise" is associated with positive sentiments, potentially indicating that businesses in idyllic locations are well-regarded by customers, with a frequency of 84.

Customer Appreciation for Walmart: "Walmart" is a recognizable brand in positive reviews, suggesting that the retail giant is delivering satisfactory customer experiences, with a frequency of 77.

Individual Recognition: The name "Steve" appears with a high frequency, potentially indicating that specific employees or individuals within businesses are making a positive impact, with a frequency of 74.

Culinary Satisfaction: Positive reviews often mention popular food items like "pizza," "chicken," and "coffee," showcasing customer satisfaction with these products.

Temporal References: Time-related mentions such as "Sunday" and "morning" may indicate when customers prefer to visit, providing insights for scheduling and staffing.

High Praise for Friendliness: The term "friendly" is widely used, suggesting that excellent customer service and friendly staff are highly valued by customers, with a frequency of 4139.

Diverse Cuisine: Multiple cuisines, including "Mexican," "Chinese," and "Italian," are associated with positive reviews, signifying a diverse culinary landscape that satisfies customers.

The graphs of EDA of negative reviews are show in Fig. 8.

A graph with blue bars

Description automatically generated

**Fig. 8 (a). Top 15 Organization entities and their frequencies for negative reviews**

A graph with blue and white bars

Description automatically generated with medium confidence

**Fig. 8 (b). Top 15 Dentist entities and their frequencies for negative reviews**

A graph with blue bars

Description automatically generated

**Fig. 8 (c). Top 15 Product entities and their frequencies for negative reviews**

A graph with blue bars

Description automatically generated with medium confidence

**Fig. 8 (d). Top 15 Organization entities and their frequencies for negative reviews**

A graph with blue bars

Description automatically generated

**Fig. 8 (e). Top 15 Time entities and their frequencies for negative reviews**

A bar graph with numbers and text

Description automatically generated

**Fig. 8 (f). Top 15 Restaurant entities and their frequencies for negative reviews**

A graph with blue bars

Description automatically generated

**Fig. 8 (g). Top 15 Quantity entities and their frequencies for negative reviews**

A bar graph with numbers and a number of blue bars

Description automatically generated with medium confidence

**Fig. 8 (h). Top 15 Cuisine entities and their frequencies for negative reviews**

A graph with blue and white bars

Description automatically generated

**Fig. 8 (i). Top 15 Price entities and their frequencies for negative reviews**

A graph with blue bars

Description automatically generated

**Fig. 8 (j). Top 15 Location entities and their frequencies for negative reviews**

A graph of a service

Description automatically generated with medium confidence

**Fig. 8 (k). Top 15 Service entities and their frequencies for negative reviews**

A graph with blue bars

Description automatically generated

**Fig. 8 (l). Top 15 Disease entities and their frequencies for negative reviews**

This EDA is a valuable foundation for further analysis and model development, providing insights into the significant named entities within the text data.

**Findings and Business Insights from negative reviews EDA:**

Challenges for Yelp: "Yelp" and its variations appear frequently in negative reviews, highlighting that users actively voice their dissatisfaction, but the positive reviews show that ‘Yelp’ is used maximum there too. Hence, the review matters on the service of the restaurants and places. Addressing these concerns could help Yelp improve its service, with a frequency of 1046.

Corporate Critique: Negative reviews mention "company" and its variations significantly, indicating areas where businesses may be falling short, with a frequency of 466.

Concerns About Fast Food Chains: Fast food restaurants like "McDonald's," "Taco Bell," and "Subway" receive criticism in negative reviews, showing areas where these chains can improve, with a frequency of 185 to 254.

Dissatisfaction with Medical Professionals: Negative reviews often mention "doctors" and "dentists," which might indicate challenges in the healthcare industry, with a frequency of 290 and 113.

Hotel Issues: "Hotel" is a term frequently associated with negative feedback, suggesting that there might be concerns related to accommodations, with a frequency of 2420.

Management Matters: "Manager" and "owner" are common in negative reviews, indicating potential issues with leadership or management, with a frequency of 6013 and 2437.

Time-Related Complaints: Negative reviews mention time-related issues, such as long wait times and poor scheduling, highlighting areas where businesses can make improvements, with a frequency of 1128 to 350.

Mixed Culinary Reviews: While negative reviews mention popular food items like "pizza," "chicken," and "coffee," they also signal areas where food quality and service can be enhanced, with a frequency of 4080 to 1272.

Displeasure with Prices: Negative reviews often criticize "expensive" and "overpriced" items, pointing to potential pricing concerns that businesses should address, with a frequency of 614 to 611.

Location Concerns: "Phoenix" and "street" appear frequently in negative reviews, suggesting issues related to specific locations, which businesses in these areas should consider addressing, with a frequency of 944 and 1146.

These insights provide valuable information about areas where businesses and services can improve and address customer concerns, ultimately leading to enhanced customer satisfaction and brand loyalty.

**Data Preprocessing and Transformation**

In the data preprocessing and transformation phase, several key steps were taken to prepare the text data for further analysis and modeling. The maximum text length was determined by calculating the maximum number of words in a document, which is essential for understanding the data's structure.

The text data underwent a series of transformations to make it more suitable for analysis. First, all text was converted to lowercase to ensure consistent text processing. Next, punctuation symbols were removed from the text, eliminating potential noise in the data. Additionally, numerical digits were stripped from the text, as they may not carry significant meaning for the analysis.

To improve the quality of the text data, common stopwords in the English language were removed. Stopwords are non-informative words that do not contribute significantly to the text's meaning. The removal of stopwords helps focus the analysis on the most relevant content.

Furthermore, lemmatization was applied to the text. Lemmatization is a linguistic process that reduces words to their base or root forms, ensuring that variations of a word are standardized. This step aids in the recognition of common word forms and reduces the dimensionality of the data.

The processed text was then reconstructed by joining the individual words back into a single string, resulting in a more structured and cleaned text data. These preprocessing and transformation steps were applied to both the training and test datasets to ensure consistency. This clean and preprocessed text data is now ready for subsequent analysis and modeling, setting the stage for effective natural language processing tasks. 10 rows of the preprocessed data is shown in Fig. 9.

A screenshot of a computer

Description automatically generated

**Fig. 9. Preprocessed Data**

**Feature Engineering and Feature Selection**

In the feature engineering process, several techniques were applied to extract valuable insights and create meaningful features from the text data. First, three key text statistics features were generated, namely 'words\_count' (representing the number of words in a document), 'character\_count' (indicating the total character count in a document), and 'average\_word\_length' (calculated as the ratio of character count to word count). These features were designed to capture the structural and linguistic characteristics of the text. A correlation analysis was conducted to evaluate the relationships between these text statistics features and the target variable 'label,' but the correlation values were found to be minimal (shown in Fig. 11), indicating that these features may not be strongly correlated with sentiment. The train data with custom made columns are shown in Fig. 10.

A screenshot of a computer

Description automatically generated

**Fig. 10. Train data with label, words\_count. Average\_word\_length**

A screenshot of a graph

Description automatically generated

**Fig. 11. Correlation Matrix**

Visualizations, including histograms, were employed to showcase the distribution of these text statistics features for 'label' 0 and 'label' 1, providing a comparative view. While the initial text statistics features did not exhibit a strong correlation with sentiment, they helped understand the underlying data structure.

A graph with a blue and green line

Description automatically generated

**Fig. 12 (a). Histogram of word count of positive and negative reviews**

**A graph with a green and red line

Description automatically generated**

**Fig. 12 (b). Histogram of character count of positive and negative reviews**

A graph with numbers and symbols

Description automatically generated with medium confidence

**Fig. 12 (c). Histogram of average word length of positive and negative reviews**

Visualizations from Fig. 12 clearly shows that words count of both positive reviews are between 20-40, character count between 100-300 and average word length 6-8.

Furthermore, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique was applied to the text data, transforming it into a numerical format suitable for machine learning. The TF-IDF matrix was computed, and the TF-IDF scores for individual words were summed to determine their importance. The top 30 words with the highest TF-IDF scores were identified, shedding light on the most informative and discriminative terms in the text data. This implicit feature selection method allowed for the reduction of dimensionality and the focus on the most relevant terms in subsequent analysis. The top 30 words having maximum frequency is calculated using the TF-IDF matrix and is shown in Fig. 13.

A screenshot of a computer

Description automatically generated

**Fig. 13. Top 30 words with maximum frequency**

Overall, these feature engineering techniques have provided a foundation for understanding the text data's structure and identifying key terms that may influence classification.

**Implementing Embedding Layer**

**Evaluate the Glove, Fasttext, and Word2Vec Pretrained Models**

In the process of implementing word embeddings, three popular word embedding techniques—Word2Vec (shown in Fig. 14), GloVe (shown in Fig. 15), and FastText (shown in Fig. 16)—were utilized to create word vectors from the text data. First, a Word2Vec model was trained on the tokenized text from the 'train\_df' dataset. The model generated word vectors with a dimension of 100, capturing semantic relationships among words. The trained Word2Vec model was saved for later use, providing a rich representation of words in the text.

A screenshot of a computer

Description automatically generated

**Fig. 14. Similar words using Word2vec model**

A screenshot of a computer

Description automatically generated

**Fig. 15. Similar words using GloVe model**

A screenshot of a computer

Description automatically generated

**Fig. 16. Similar words using Fasttext model**

Subsequently, GloVe embeddings were incorporated into the analysis. A pre-trained GloVe model, specifically the 'glove.6B.100d' model, was loaded. This allowed the utilization of GloVe word vectors. The GloVe model provided additional word vectors to compare with the Word2Vec embeddings.

FastText, another powerful word embedding technique, was also employed. A FastText model was trained on the tokenized text data, similar to Word2Vec, with a vector dimension of 100. FastText, known for its ability to capture subword information, was a valuable addition to the analysis.

For all three embedding models (Word2Vec, GloVe, and FastText), similarity analysis was conducted to identify words similar to the top 30 words with the highest TF-IDF scores. This analysis provided insights into the semantic relationships among words in the data. The word embeddings will surely help in the classification of reviews effectively.

**Selecting Best Embedding Method**

The Word2Vec similarity analysis provides valuable insights into the associations between specific keywords and the context in which they appear. The results show that the model has successfully captured semantic relationships among words. For example, the similarity between "food" and terms like "ntaste," "quickness," and "nsushi" reflects how customers often discuss the quality of food in their reviews. Similarly, the strong association between "place" and words like "restaurant" and "joint" highlights the tendency of reviewers to refer to dining establishments when discussing a place. The model's ability to identify synonyms and related terms for words like "good," "great," "service," and "friendly" demonstrates its capability to grasp sentiment and customer experiences. This analysis showcases the potential of Word2Vec in extracting meaningful patterns and relationships from text data, which can be valuable for sentiment analysis and understanding customer preferences in the restaurant industry.

The GloVe word embeddings reveal interesting patterns in word similarity. It's clear that the embeddings capture semantic relationships, with words like "food" being like "foods," "supplies," and "meat." Similarly, "place" is associated with terms like "time," "next," and "spot," indicating its role in describing locations and moments. Words like "good" are closely related to "better," reflecting the concept of improvement. Additionally, the embeddings capture social and emotional aspects, as "love" is like "passion" and "dream," while "friendly" is linked to "cooperative" and "warm." Overall, these similarities illustrate the richness and complexity of language as expressed in Yelp reviews, and demonstrate how GloVe embeddings can provide valuable insights into the semantics and context of words.

The Fasttext word embeddings provides some insights, words like “place” are “pplace”, “pplaces”, “nplace”, “townplace”, “placebo”, etc. Some of these similarities are helpful but maximum number of words have the parent word inside them. Overall, these texts show good similarity but lacks in the meanings and emotional aspects.

Cosine similarity analysis was performed to measure the similarity between word vectors from each embedding model for the top words. The results highlighted the varying degrees of similarity between the embedding models, shedding light on the strengths and nuances of each technique. The embedding vectors of the word “good” and its similar word’s embedding vectors are calculated. Cosine similarity between embedding model vectors of word, “good” and the similar word using the embedding model is shown in Fig. 17.



**Fig. 17. Cosine Similarity of Word2vec, GloVe and Fasttext Model**

In terms of cosine similarity analysis, the Fasttext embedding model outperform all the other models, but to select the best embedding model, GloVe is the best embedding model where the cosine similarity ranks second but the in terms of similarity between words it stands first. This is because, words found by the gloVe model are similar in terms of emotional aspects and dictionary too. Hence, gloVe model is the best embedding model among Word2Vec, GloVe and Fasttext.

These embedding techniques, along with the similarity analysis, enriched the understanding of word relationships within the text data and are essential for subsequent natural language processing and deep learning tasks.

**Conclusion**

In conclusion, the key findings from my project involve the analysis and processing of Yelp review data, focusing on Named Entity Recognition (NER), text data distribution, feature engineering, and embedding methods. I conducted a thorough exploratory data analysis (EDA) to understand the characteristics of the text data, identified significant named entities, and provided business insights into customer preferences and concerns. The data was preprocessed, and relevant features were engineered to prepare it for further analysis. Additionally, three popular word embedding methods—Word2Vec, GloVe, and FastText—were evaluated for their ability to capture semantic relationships among words in the text data.

After a detailed analysis of the embedding methods, we found that GloVe word embeddings stood out as the most suitable choice for the project. This choice was made based on their ability to capture semantic relationships, including synonyms and context, effectively. The GloVe embeddings were shown to provide valuable insights into the language used in Yelp reviews, enhancing the understanding of customer reviews and preferences.

The implications of choosing GloVe embeddings for the project is significant. These embeddings will serve as a crucial foundation for subsequent natural language processing and deep learning tasks. With GloVe embeddings, I can better analyze and categorize Yelp reviews, identify customer preferences, and gain insights into areas for improvement in businesses and services. The richness and complexity of the GloVe embeddings will help to build more accurate and robust models for text classification, ultimately leading to improved customer satisfaction and brand loyalty. Therefore, the choice of GloVe embeddings is a key step towards achieving project goals of understanding and enhancing the customer experience in the industry.

**Reference**

1. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. [Dataset]. Papers with Code. Retrieved from <https://paperswithcode.com/dataset/yelp-review-polarity/>
2. Zhang, X., Zhao, J., & LeCun, Y. (2016). Character-level Convolutional Networks for Text Classification. arXiv preprint arXiv:1509.01626.
3. Ghotra, Amandeep & Kudupudi, Chaitanya & Konda, Komali & Bizel, Gulhan & Gilkey, Joseph. (2020). Extraction of Aspects and Opinion Indicators from Yelp Reviews Using Different Methods of Sentiment Analysis. SSRN Electronic Journal. 10.2139/ssrn.3873335.
4. Monigatti, L. (Aug 31, 2022). "Fundamental EDA Techniques for NLP." Towards Data Science. Retrieved from <https://towardsdatascience.com/fundamental-eda-techniques-for-nlp-f81a93696a75>.
5. Wei, Jason, and Kai Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks." EMNLP-IJCNLP 2019 short paper. arXiv:1901.11196 [cs.CL], 2019. DOI: 10.48550/arXiv.1901.11196.
6. Fang, X., Zhan, J. Sentiment analysis using product review data. Journal of Big Data 2, 5 (2015). <https://doi.org/10.1186/s40537-015-0015-2>
7. Barigou F (2018) Impact of instance selection on kNN-based text categorization. J Inf Process Syst 14(2):418–434
8. Tecoholic. (n.d.). NER Annotator. Tecoholic. <https://tecoholic.github.io/ner-annotator/>