Yelp Review Classification

Github Repository

1st Sri Datta Kiran Avasarala
Univ. of North Carolina at CharlotteCharlotte, NC
Savasar1@uncc.edu

2nd Sandeep Reddy Chalamalla
Univ. of North Carolina at Charlotte Charlotte, NC
schalama@uncc.edu

I. Introduction

Businesses that rely heavily on customer evaluations and ratings to draw in new clients and uphold their online reputations are called "review-based businesses." These companies operate in a wide range of sectors, such as hotels, cafes, restaurants, and beauty parlors, to name a few.

Customers may now share their experiences and opinions about companies more easily because of the growth of social media and online review sites like Yelp, TripAdvisor, and Google Reviews. In order to effectively manage their reviews and ratings, review-based businesses must pay close attention to their online reputation.

A company's reputation and success can be strongly impacted by online evaluations and ratings, both positively and adversely. Negative reviews and low ratings can turn away potential consumers and harm a company's reputation, whilst positive reviews and high ratings can draw in new clients and boost client loyalty.

Review-based businesses must deliver exceptional goods and services and skillfully manage their online reputations if they want to prosper in today's cutthroat marketplace. This includes replying to reviews, handling client grievances, and routinely monitoring their web presence.

Yelp is a website that allows users to rate and review local establishments like restaurants, cafes, hairdressers, and many others. Since its establishment in 2004, it has grown to be a well-liked platform for users to discuss their ideas and experiences with nearby companies.

User profiles, social networking, search options, and filtering are just a few of the tools that Yelp offers consumers to locate and assess local businesses. The platform also provides tools for businesses to monitor their operations, communicate with clients, and control their internet reputation.

Yelp has been more well-liked in recent years as a research and analytical tool, particularly in the areas of sentiment analysis and natural language processing. Machine learning models for sentiment analysis and classification as well as for obtaining insightful data about customer sentiment can be developed using Yelp reviews and ratings.

II. Approach

The approach for the Yelp review classification using sentimental analysis was focused on creating a machine learning model that could accurately predict the sentiment of a review based on its text content. The first step in this process was to perform data preprocessing on the Yelp review dataset. The data preprocessing included basic text cleaning techniques such as removing stop words, stemming, and converting all text to lowercase.

After the text cleaning, the next step was to perform feature extraction using TF-IDF vectorization. The TF-IDF vectorization [1] converts the textual data into numerical vectors, which can be used as input for the machine learning models. This process assigns weights to each term in the document based on its frequency and importance in the document. This helps to reduce the dimensionality of the data and improves the performance of the models.

Once the data was preprocessed and feature extraction was completed, several machine learning models were trained on the Yelp review dataset. The models chosen for this project were Linear SVC, Random Forest, Naive Bayes, LSTM, and GRU. These models were selected based on their ability to perform well on text classification tasks.

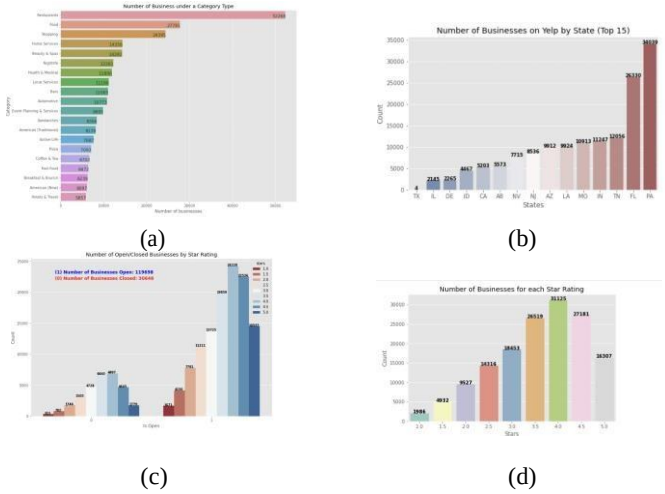
III. Dataset and Training Setup

A. Dataset and Data Cleaning

This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research or analysis on Yelp's data and share their discoveries [2]. In the most recent dataset you'll find information about businesses across 8 metropolitan areas in the USA and Canada. The review data set was taken from Kaggle.It consisted of around 699k reviews on total , where a subset of 10k reviews were used as an input for processing and training.

It consisted of different columns like Review ID, User ID, Business ID, Stars, Useful, Funny, Cool, Text and Date. The data has been used to generate visualizations which shows us the different organization categories that are available, classification of these organisations based on ratings, classification based on being opened and closed and then we had classification based on the states that the organization were available in.

Data pre-processing involved the basic text cleaning process wherein we initially converted the entire reviews into lower case characters. After this, the contractions words were removed. For example, the words like I'm, couldn't, weren't were abbreviated as I am, could not, were not etc,. After all the contraction words has been removed, the stop words like I, will, wont, could, go



(a) (b)

(c) (d)

Fig. 1: Data Exploration

, yes, no etc,. were removed. Once all of this data has been cleared, the lemmatization process has been performed on the data where a natural language processing technique used to convert words to their base or dictionary form, known as a lemma. This is achieved by applying morphological analysis to remove inflectional endings and reduce words to their base form.

The resulting lemmas are then used for text analysis and processing. Compared to stemming, which simply truncates words, lemmatization produces more accurate and meaningful results by preserving the grammatical structure of the text. However, lemmatization can be computationally expensive and may not always be necessary depending on the task at hand. Once the data has been further cleaned, the TFIDF vectorization was performed on the cleaned data. The TF and IDF scores are multiplied together to get the TF-IDF score for each word in each document. This results in a numerical vector for each document, with each dimension representing a different word. The values in the vector represent the importance of each word in the document relative to the entire dataset. The entirely cleaned data is then turned into numerical data using sparse data methodology.

(a) Original Text (b) Text after Cleaning

Fig. 2: Text Analysis and Cleaning

B. Model Training

For the task of sentiment analysis on Yelp reviews, several models were trained to classify reviews as positive or negative. The rating scale we tried to implement was 1,2 were considered negative reviews and 3,4 and 5 were considered positive reviews. The models we used for this task were Linear SVC, Random Forest, LSTM, Naive Bayes, and GRU. Every model had its own approach to training and classification.

When training the model for Linear SVC classifier, it was trained using the sklearn library. The training involved fitting the model to the training data and then using it to predict the sentiment of the validation data. The accuracy of the model was evaluated using sklearn metrics.

When training the model for Random Forest classifier, it was also trained using the sklearn library. The training involved creating a random forest of decision trees and fitting it to the training data. The accuracy of the model was evaluated using sklearn metrics.

When training the model for LSTM classifier, it was trained using the Keras library. The training involved defining the model architecture and compiling it with the appropriate loss function and optimizer. The model was then trained on the training data and its performance was evaluated using accuracy and loss metrics.

When training the model for Naïve Bayes classifier, it was trained using the sklearn library. The training involved fitting the model to the training data and using it to predict the sentiment of the validation data. The accuracy of the model was evaluated using sklearn metrics.

When training the model for GRU, it was also trained using the Keras library. The training involved defining the model architecture and compiling it with the appropriate loss function and optimizer. The model was then trained on the training data and its performance was evaluated using accuracy and loss metrics.

Overall, the training of each model involved defining the model architecture, compiling it with appropriate loss and optimization functions, and fitting it to the training data. The accuracy of each model was evaluated using validation data and appropriate metrics. The model with the highest accuracy was selected as the final model for sentiment analysis on Yelp reviews.

IV. Results and Analysis

To assess the performance of the Linear SVC model, a classification report was generated using the classification_report function from scikit-learn. The report compared the predicted labels (y_test_pred_clf) with the true labels (y_test) of the test dataset. It produced a dictionary containing precision, recall, F1-score, and support for each class. These metrics were then converted into a pandas data Frame for better readability.

After training the Random Forest model with the best hyperparameters, the model was used to make predictions on the test dataset, represented by tf_x_test. The predicted labels were stored in the y_test_pred_rf variable. The grid search identified the best hyperparameters for the Random

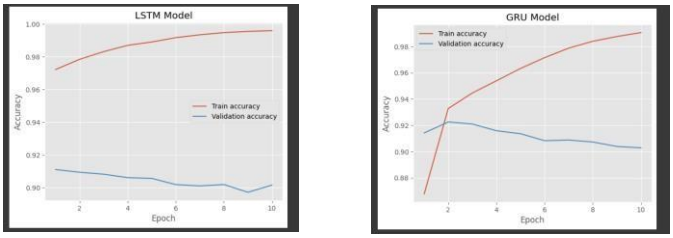
TABLE I: Model Comparison

Model	Accuracy(%)
Linear SVC	- 91.79
Random Forests	- 88.10
LSTM	- 91.42
GRU	- 90.28
Naive Bayes	- 87.31

Forest model were given as: The maximum depth was none, and the no. of estimators was given as 100. The best score achieved by the Random Forest model during the grid search was 0.9288, indicating the average cross-validated accuracy across the different hyperparameter combinations.

After training the Naive Bayes model with the best hyperparameters, the model was used to make predictions on the test dataset, represented by tf_x_test. The predicted labels were stored in the y_test_pred_nb variable.The grid search identified the best hyperparameter aplha for the Naive Bayes model was The best score achieved by the Naive Bayes model during the grid search was 0.9010, indicating the average cross-validated accuracy across the different alpha values.

The LSTM model and the GRU were trained on the training dataset (X_train and y_train) for a total of 10 epochs, with a batch size of 32. The validation_data parameter was set to the test dataset (X_test and y_test) to monitor the model's performance on unseen data during training. The number of hidden layers in these models were around 2 different layers: Embedding layer [3] with 32 output dimensions, LSTM layer with 64 units, The comparisons of the model accuracies is shown in table 1.



(a) LSTM (b) GRU

Fig. 3: Training and Validation Accuracy

V. Lessons Learned

There are several lessons that can be learned from performing review classification using sentiment analysis. Especially, building it using models like LSTM and GRU was way more complicated than we estimated it to be because of the classification that we had to perform and the pre processing as well. Different models have their strengths and weaknesses, and it is essential to try out multiple models to find the best-performing one for the given dataset. Text cleaning and pre-processing of data play a crucial role in sentiment analysis. Techniques such as removing stop words, stemming, and lemmatization can greatly improve the performance of sentiment analysis models. Understanding TF-IDF vectorization was difficult which was basically used to create features that could help the model differentiate between positive and negative reviews. Issues like Overfitting had to be overcome which basically occur when a model is too complex and fits the training data too closely, resulting in poor performance on the test data.

References

[1] D Alessia, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3), 2015.

[2] Nabiha Asghar. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*, 2016.

[3] Eman Saeed Alamoudi and Norah Saleh Alghamdi. 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, 2021.