SENTIMENT ANALYSIS OF YELP REVIEWS - FINAL REPORT

BACKGROUND AND LITERATURE REVIEW

Yelp is a widely used online review platform with a vast collection of over 200 million user-generated reviews covering various local businesses, such as restaurants and dog parks. Yelp's user-generated reviews and ratings can be helpful for both businesses and consumers, as businesses can use the feedback to improve their services, and consumers can make informed decisions. Sentiment analysis refers to the process of analyzing text to determine if the emotional tone of the message is positive, negative, or neutral. This project focuses on training a machine learning model to learn from a supervised yelp dataset which already has tone labels and predicting the tone of unseen yelp reviews.

The sentiment analysis of Yelp reviews has been a popular research topic in recent years. Many studies have focused on developing machine learning models that accurately classify the sentiment of Yelp reviews. One such study by Xiong et al. (2018) used deep learning models to classify Yelp reviews as positive or negative. Another study by Ghose and Ipeirotis (2011) focused on the impact of online reviews on the sales of local businesses. The study analyzed over 40,000 Yelp reviews and found that a one-star increase in Yelp rating leads to a 5-9% increase in revenue for a local business. Overall, the sentiment analysis of Yelp reviews has been a valuable research area, with potential applications in business, marketing, and customer experience. While there have been successes in developing accurate sentiment analysis models, there is still room for improvement in addressing the limitations of past research.

One of the main limitations of sentiment analysis of Yelp reviews in the past has been the accuracy of the machine learning models used. Previous studies have shown that achieving high accuracy in sentiment analysis is challenging due to the complexity of human language and nuances in sentiment expression. Bias in training data and difficulty in identifying context and tone can also impact the results. Moreover, studies like Hu et al. (2018) have shown that sentiment analysis models struggle to classify ambiguous reviews containing sarcasm or irony. Reviews can contain variations of language, such as slang, abbreviations, and misspellings, making it difficult for sentiment analysis models to interpret accurately. Another limitation is the subjective nature of sentiment analysis, as different people may interpret the sentiment of a review differently, leading to subjective judgments about the accuracy of sentiment analysis models. These findings suggest a need for advanced models that can accurately interpret the context of reviews.

OBJECTIVE

The objective of this sentiment analysis project is to develop a machine learning model that accurately classifies the sentiment of Yelp reviews into one of five categories: positive, somewhat positive, neutral, somewhat negative, or negative. The ultimate goal is to provide insights around brand perception for

businesses that rely on customer feedback, allowing them to identify areas for improvement and enhance the overall customer experience.

DATASET

This project is using the Yelp Customer Review dataset from Yelp, which is a collection of user-generated reviews, ratings, and other metadata for local businesses in various cities. It includes ~10000 reviews contributed by ~6,300 users covering ~4,100 businesses, such as restaurants, bars, salons, and more. The reviews have been labeled with sentiment ratings from 1 to 5 by Yelp's machine learning algorithm. The data includes the business ID, date, review ID, stars given by the reviewer, the text of the review, the type of review, user ID, and the number of cool, useful, and funny votes given by other users. The focus of the project will be on the star rating column which will be the target variable, and the text column containing reviews which will be used to create the features on which the model will be trained.

DATA CLEANING

The dataset was cleaned prior to training the machine learning models with it. The number of missing values in each column of the dataset was checked and found to be 0. Then, a check was performed to ensure that there were no duplicate entries in the dataset. Following this, the data type of each column was standardized. For example, the *date* column contains the date on which a review was written, which was originally a text column. This was converted to a datetime type column so that it becomes easier to perform date operations on this column such as extracting the month or year. Finally, the columns were evaluated for their qualitative relevance with respect to the use case. The columns type, business_id, review_id, and user_id were dropped as these columns were found to be irrelevant for sentiment analysis.

DATA PREPROCESSING AND FEATURE CREATION

After the dataset was cleaned, it was put through some text preprocessing to extract meaningful features out from the reviews. This was done using natural language processing techniques by leveraging NLTK, a python library for text processing. First, the reviews were tokenized, meaning they were split into individual words. In order to focus more on the key sentiment indicators rather than grammar, certain high frequency words such as “the”, “a”, “and”, etc were removed in the stopword removal process. To further make it easier for the model to make sense of words, every word was lemmatized, meaning that it was broken down to its base form. This means that both “swimming” and “swam” were replaced with their root word “swim”. To incorporate the semantic importance of words, a Word2Vec model was trained on the tokenized data to generate word embeddings that encode the semantics. Finally, the processed text was vectorized to convert it to a format that classification models can work with.

The vectorized reviews were the primary feature for training the classification model. In addition to this, some secondary features were generated to help the model get a better signal for classification of reviews. Polarity and Subjectivity scores of the review are such feature which measure the subjective opinion in a review. Polarity score measures the sentiment expressed in the text, i.e., whether the text expresses a positive or negative. It is a float value between -1.0 and 1.0, where -1.0 indicates a highly negative sentiment, 1.0 indicates a highly positive sentiment. Subjectivity score measures the degree to which the text expresses a personal opinion, feeling, or emotion rather than a factual statement. It is a float value between 0.0 and 1.0, where 0.0 indicates an objective text and 1.0 indicates a highly subjective text. Another feature is the negated sentiment score which calculates the sentiment while taking into account negation words. This is important because sometimes negative words like "not" can change the meaning of a sentence. For example, "I am happy" has a positive sentiment score, but "I am not happy" has a negative sentiment score.

EXPLORATORY DATA ANALYSIS

After the text was preprocessed and some meaningful features were extracted from the reviews, some exploratory data analysis was performed to get a better understanding of the data. Upon looking at the distribution of the number of reviews by the assigned star rating (Fig 1), it was found that the dataset had some inherent imbalance with there being a higher number of 4 and 5 star rated reviews and very few 1 and 2 star rated reviews. This was handled prior to training of the classification model.

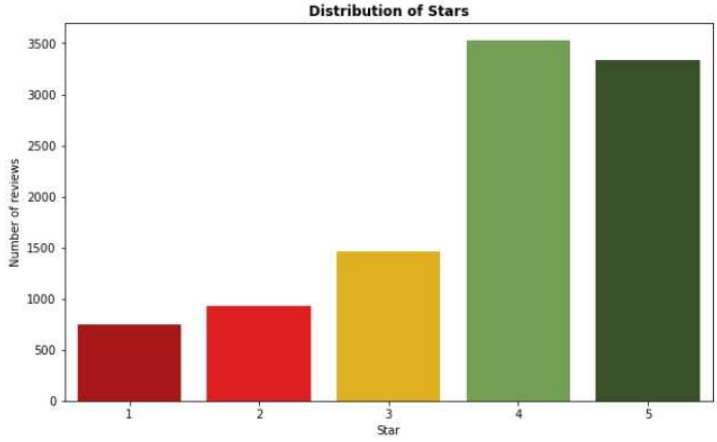

Fig 1: Distribution of reviews by their star rating

To determine if the words occurring in the reviews were congruent with the star rating of the review, the top 10 most common adjectives in each star rating class were plotted (Fig 2). It can be observed that as the star rating increases, the frequency of positive adjectives like “great”, “nice”, and “delicious” increases while that of negative adjectives like “bad”, and “small” decreases.

A simple histogram of the polarity score (Fig 3) showed that the reviews were more positive in tone overall, which is as expected since there are more 4 & 5 star rated reviews in the dataset. A histogram of subjectivity score showed a clear peak at 0.5, indicating that the reviews had a good balance of objective and subjective comments.

Fig 2: Most common adjectives by star rating

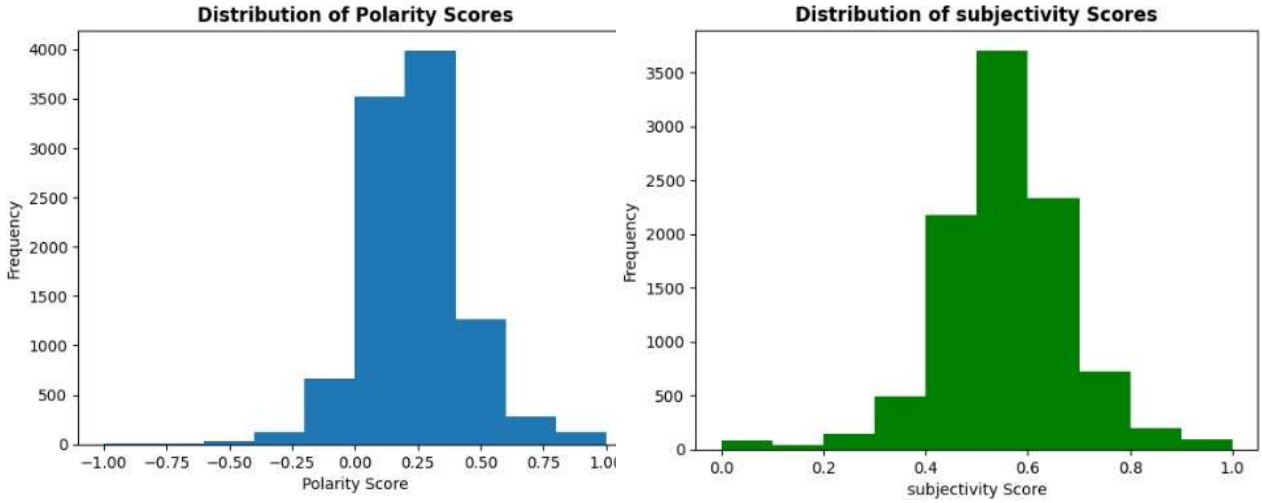

Fig 3: Histogram of polarity and subjectivity scores

MULTICLASS CLASSIFICATION & RESULTS

Prior to the training classification model, the imbalance in the dataset was handled using SMOTE Synthetic Minority Oversampling Technique. SMOTE is a popular technique used in machine learning to address the problem of class imbalance in a dataset. It is a type of oversampling method that creates synthetic samples of the minority class to increase its representation in the dataset.

The dataset was split into train and test dataset with a 70% to 30% ratio. The train dataset was used to train a Random Forest Classifier, Support Vector Machine and Multinomial Naive Bayes Classifier. The target variable had 5 classes - 1 to 5 representing Poor to Great. After training, the model was tested on the test dataset. The predictions were evaluated by looking at the classification report for the predictions made on the test dataset. To get a better understanding of the results, the precision, recall and F1-score from all the models were put together and conditionally formatted such that the darker the shade, the higher the value.

Fig 4 shows the result from the random forest classifier. The overall accuracy obtained is 60%. The model overall gives a better precision than recall and performs the best with the bad class with 69% precision and 80% recall, followed by the poor class with 73% precision and 57% recall.

Accuracy: 0.6055019852524106

	precision	recall	f1-score	support
bad	0.69	0.80	0.74	706
good	0.50	0.52	0.51	766
great	0.53	0.65	0.58	688
neutral	0.62	0.49	0.55	673
poor	0.73	0.57	0.64	693
accuracy			0.61	3526
macro avg	0.61	0.61	0.60	3526
weighted avg	0.61	0.61	0.60	3526

Fig 4: Classification report from Random Forest Classifier

Fig 5 shows the result from Support Vector Machine. This model is more balanced between precision and recall, overall. However, it gives a better recall than precision for the bad class, while for the other classes the gap is minimal. This model arguably performs at the same level for bad, poor and great classes, showing that it has a better discerning ability between classes as compared to random forest.

	precision	recall	f1-score	support
bad	0.60	0.83	0.69	1049
good	0.48	0.49	0.49	1126
great	0.55	0.53	0.54	1040
neutral	0.54	0.48	0.50	1038
poor	0.68	0.49	0.57	1036
accuracy			0.56	5289
macro avg	0.57	0.56	0.56	5289
weighted avg	0.57	0.56	0.56	5289

Fig 5: Classification report from Support Vector Machine

Fig 6 shows the result from the Multinomial Naive Bayes model. This model is even more balanced between precision and recall, overall. However, it tends towards a better recall than precision. The model gives the

best performance for Bad class with 79% precision and 48% recall. This model also appears to performs at the same level for bad, poor and great classes, showing that it has a better discerning ability between classes as compared to random forest, similar to Support Vector Machine. However, given that it has a much better overall performance as compared to the other models, Multinomial Naive Bayes is the final model chosen for this project.

	precision	recall	f1-score	support
bad	0.79	0.48	0.59	1049
good	0.44	0.65	0.53	1126
great	0.53	0.58	0.55	1040
neutral	0.52	0.41	0.46	1038
poor	0.51	0.51	0.51	1036
accuracy			0.53	5289
macro avg	0.56	0.53	0.53	5289
weighted avg	0.56	0.53	0.53	5289

Fig 6: Classification report from Multinomial Naive Bayes Classifier

LIMITATIONS

Sentiment analysis applications are generally limited by the very nature of the underlying data, i.e. the ambiguity in the text which can give rise to inaccurate predictions. For example, reviews could have text like “The food was amazing but the service was horrible, which is why I give a low rating” – If the reviews have equal number of sentences conveying positive and negative points, then the model could get confused while learning Also, Sarcasm and irony which are intricate levels of english language constructs would confuse the learning process. The Yelp review dataset is biased towards users who are more likely to leave reviews, which may not be representative of the overall population. Additionally, there may be biases in the reviews themselves, such as users who are more likely to leave negative reviews.

In terms of the scope of this project, there are several ways in which the current analysis can be improved. First and foremost, spending more time on hyperparameter tuning of the classifiers used could result in better scores. One way to do this is to use grid search with cross validation. Another area of improvement would be to try other ways of handling data imbalance such as specifying a class weight parameter in the models. Also, more sophisticated classifiers such as Multi layer perceptron can be used to get potentially better predictions. Finally, reducing the number of classes could result in a better performance as the model would be required to create less number of decision boundaries.

CONCLUSION

Sentiment analysis of Yelp reviews can be a valuable tool for businesses looking to better understand their customers and improve their overall reputation. The results achieved show a promising trend in terms of Precision, Recall and F1 scores. The various models were able to achieve performance metric scores in the range of 80% when SMOTE technique was applied. This gives us a strong direction to predict sentiment of user generated reviews. The work completed in this project can have wide ranging applications such as understanding customer satisfaction in airline and hotel industries, analyzing public opinion on political issues, conducting market research to understand consumer preferences and more.

REFERENCES

Xiong, C., Dai, Z., Bian, J., & Xiong, D. (2018). Sentiment analysis of Yelp's review corpus. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1026-1033). IEEE.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering, 23(10), 1498-1512.

Hu, S., Chen, Y., & Kim, S. (2018). Yelp review analysis: From simple to complex sentiment classification. Journal of Hospitality and Tourism Technology, 9(3), 311-326.

Sabnis, O. (2019). Yelp Reviews Dataset. Retrieved from https://www.kaggle.com/omkarsabnis/yelp-reviews-dataset.

Urytrayudu, S. (2021). Sentiment Analysis for Yelp Review Classification. Retrieved from https://urytrayudu1.medium.com/sentiment-analysis-for-yelp-review-classification-54b65c09ff7b.

Xu, Y., Wu, X., & Wang, Q. (2014). Sentiment analysis of Yelp's ratings based on text reviews. CS229 Project Report. Retrieved from https://cs229.stanford.edu/proj2014/Yun%20Xu,%20Xinhui%20Wu,%20Qinxia%20Wang,%20Sentiment%20Analysis%20of%20Yelp's%20Ratings%20Based%20on%20Text%20Reviews.pdf.