OVERVIEW

Large data sets have been integrated into this presentation, which can be helpful for every organization's operational aspect where decision-makers can monitor the prospective effects of its features and operations. Data may be transformed into valuable information, and when that information is consolidated, it can be used to develop concrete analyses and answers on how to operate a particular feature more effectively and efficiently.

These days, we have a lot of open source resources available, and our company uses the cost of the mechanisms as a sign to provide our clients with our solutions more quickly. Manual data interpretation will be labor-intensive and have a finite scope. There are literally hundreds of data points that cannot all be understood at once. Data science has thus made a significant contribution to technology that makes interpretation simple.

For instance, this analysis is particularly about reviews from different kinds of businesses that can be found on yelp, a website that publishes crowd-sourced reviews about businesses.

About this Data

This is a list of 500 yelp reviews provided by yelp themselves on their website. Their dataset includes business name, rating, review data, title, username, date, and more.

OBJECTIVES

What you can do with this Data

Using natural language processing and machine learning techniques, you can use this data to classify customer reviews into positive or negative sentiments. The review data lets you correlate keywords in the review text with ratings and aims the objective which is to save time for an individual while making a decision about a product solely allowing for checking the ratings instead of textual reviews.
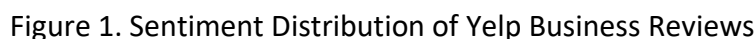
PROCESS AND METHOD

This analysis uses Sentiment Analysis, Word Cloud, Multinomial Naive Bayes, and confusion Matrix to interpret the data sets. This process also obtain research and survey as to the use of 500 yelp reviews and the following methods were used:

- NLTK: the most famous python module for NLP techniques
- Numpy: offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more
- Scikit-learn: the most used python machine learning library
- TensorFlow: is an end-to-end open source platform for machine learning.
- Panda: is used to analyze data

- Word Cloud: this is a technique to show which words are the most frequent in the given text.
- Seaborn: data visualization library based on matplotlib
- Matplotlib: is a comprehensive library for creating static, animated, and interactive visualizations
- Collections: implements specialized container datatypes providing alternatives to Python's general purpose built-in containers, dict , list , set , and tuple .
- BS4: library for pulling data out of HTML and XML files.

RESULT AND DISCUSSION

Yelp Business Reviews Database

B. Confusion Matrix Result

Figure 1. Sentiment Distribution of Yelp Business Reviews

This sentiment distribution represented the total number of positive and negative reviews from the sites. Seeing the figure, there is a fair sentiment from the proponents were negative and positive.

B. Confusion Matrix Result

Figure 2. Frequent Used words in the review

```
              precision    recall  f1-score   support

           0       1.00      0.97      0.99        34
           1       0.98      1.00      0.99        56

    accuracy                           0.99        90
   macro avg       0.99      0.99      0.99        90
weighted avg       0.99      0.99      0.99        90
```

Figure 3. Confusion Matrix from the reviews

From the figure presented above, the confusion matrix has reached an accuracy of 99%and a precision of 100% in which these representations can be described with anequal sentiment frequency.
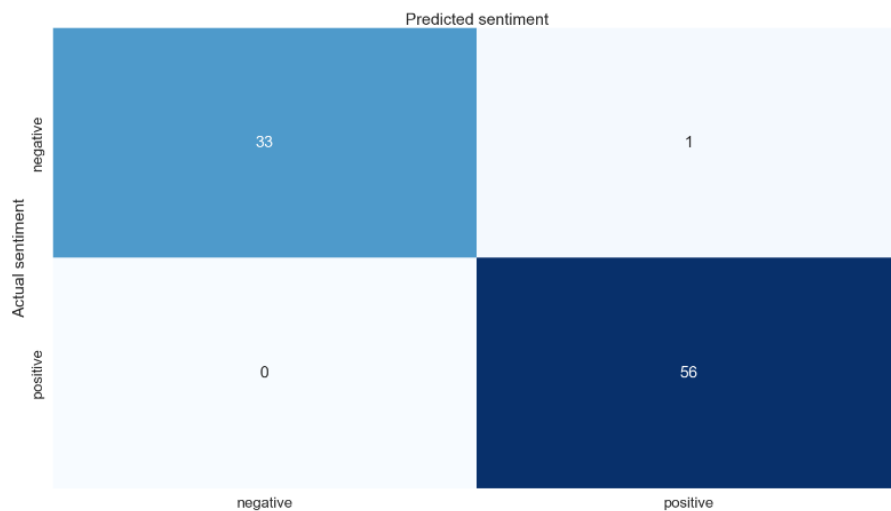

Figure 4. Predicted Sentiment vs Actual Sentiment from the reviews