# CMPE 255

# Analyzing User Feedback on Yelp Reviews: A Natural Language Processing and Data Mining Approach

Made By: TRISHALA M (SJSU ID: 015219646)
AKANKSHA GUPTA (SJSU ID: 014707784)
VANDANA PATEL (SJSU ID: 016114852)
CHIRAG ARORA (SJSU ID: 016726567)

# MOTIVATION

Customer feedback is essential for businesses as it provides insights into customer experiences, opinions, and preferences, which can help improve services and retain customers.

# OBJECTIVES

- To develop models that can classify user comments as positive or negative sentiments using natural language processing and data mining techniques.
- To analyze the sentiment in greater detail to understand the reasoning behind positive or negative reviews.

# ALGORITHMS USED

We used several classification algorithms, including

- **Naive Bayes** is a probabilistic algorithm that assumes independence among the features.
- **Decision Tree** algorithm splits the data into branches based on the most significant attribute.
- **Support Vector Classifier** tries to find the best hyperplane that separates the data into two classes.
- **Random Forest** is an ensemble learning algorithm that creates multiple decision trees and combines their results.

# DATASETS USED

- **Yelp Academic Dataset Business:** 160585 rows and 14 columns (1.21 GB) (business_id, name, address, city, state, postal code, latitude, longitude, stars, review_count, is open, attributes, categories, hours)
- **Yelp Academic Dataset Review:** 879878 rows and 9 columns (6.25 GB) (review_id, user_id, business_id, stars, useful, funny, cool, text, date)

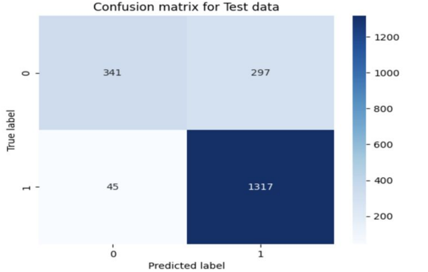
# Methodology – Part-I Predicting user sentiment

1. Loading and Merging Datasets
2. Preprocessing
3. Filtering and Adding a New Column
4. Assigning Sentiment Values
5. Normalization: The next step was to perform tf-idf normalization of the rows to ensure that the dataset was scaled properly.
6. Splitting the Dataset: The dataset was split into an 80:20 ratio.
7. Finally, several machine learning models were used to predict the sentiment of the review.

# RESULTS

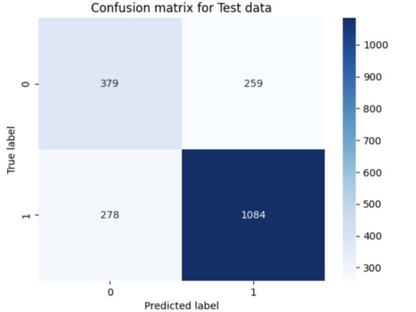
**Naive Bayes**

Precision Score of the model: 81.59851301115242
Recall Score of the model: 96.69603524229075
Accuracy score of the model: 82.89999999999999
F1 score of the model: 88.50806451612902



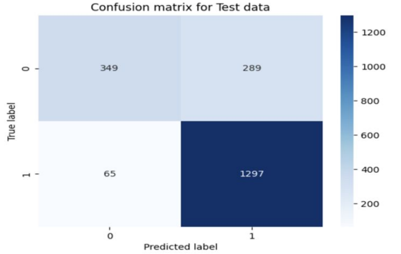

**Decision Tree**

Precision Score of the model: 80.71481757259866
Recall Score of the model: 79.58883994126285
Accuracy score of the model: 73.15
F1 score of the model: 80.1478743068392



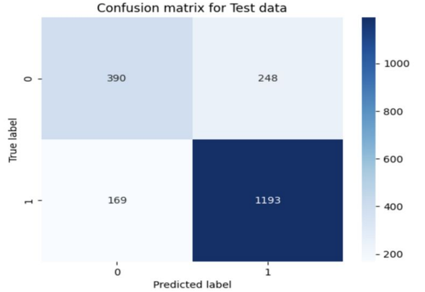

**Support vector classifier**

Precision Score of the model: 81.7780580075662
Recall Score of the model: 95.22760646108664
Accuracy score of the model: 82.3
F1 score of the model: 87.99185888738127




**Random forest**

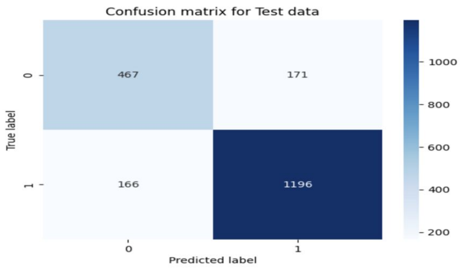
Precision Score of the model: 82.78972935461485
Recall Score of the model: 87.59177679882526
Accuracy score of the model: 79.14999999999999
F1 score of the model: 85.12308241170174



# To improve the results, we performed oversampling and hypertuning grid search cross   validation

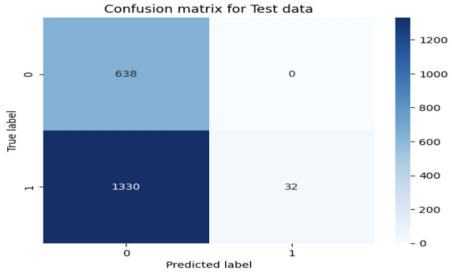
**Naive Bayes**

Best parameters for the algorithm - Alpha = 0.001
Precision Score of the model: 87.4908558888076
Recall Score of the model: 87.81204111600587
Accuracy score of the model: 83.15
F1 score of the model: 87.651154268963



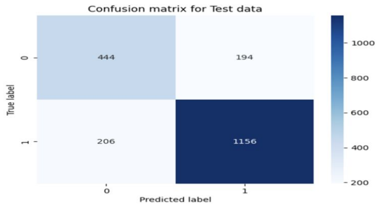

**Decision Tree**

Best parameters for the algorithm DecisionTreeClassifier (max_depth=8, max_features=5,min_samples_leaf=5, min_samples_split=3)
Precision Score of the model: 100.0
Recall Score of the model: 2.3494860499265786
Accuracy score of the model: 33.5
F1 score of the model: 4.591104734576758




**Support vector classifier**

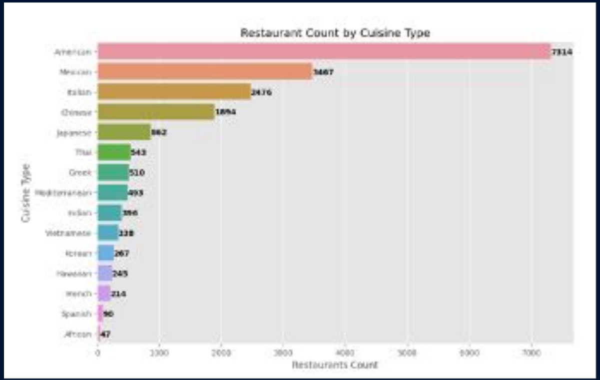
Best parameters for the algorithm SGDClassifier(alpha=0.0001, max_iter=20)
Precision Score of the model: 85.62962962962963
Recall Score of the model: 84.87518355359765
Accuracy score of the model: 80.0
F1 score of the model: 85.25073746312685

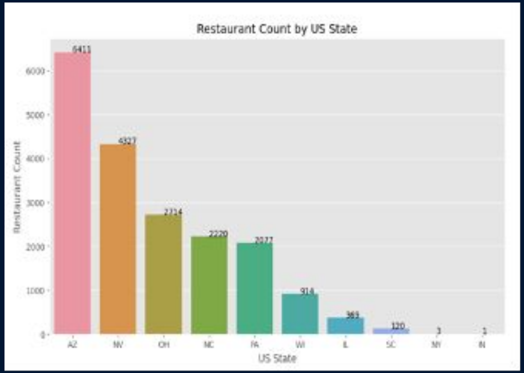


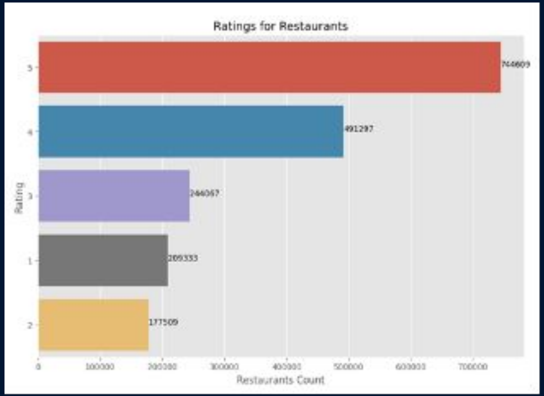
# Part 2 - Analyzing the polarity of positive and negative sentiment

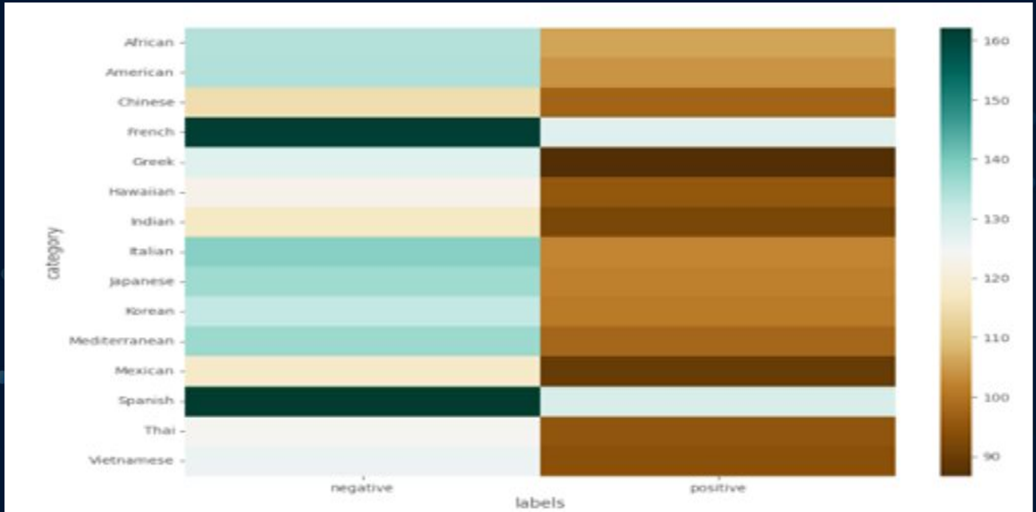
1. We analyzed Yelp restaurant reviews to understand factors contributing to positive or negative reviews.
2. We used Count Vectorizer and SVC classifier to obtain the score of each word.
3. Polarity score of a word indicates its contribution to the sentiment of that review.
4. We dropped obvious polarity words and identified top 10 words contributing to each cuisine type.
5. We reinforced our analysis through exploratory data analysis.

# RESULTS


Restaurant Count by Cuisine Type
American
Mexican
Italian
Chinese
Japanese
Thai
Greek
Mediterranean
Indian
Vietnamese
Korean
Hawaiian
French
Spanish
African
7314
2476
1894
862
543
510
493
396
267
245
214
90
47
Cuisine Type
Restaurants Count

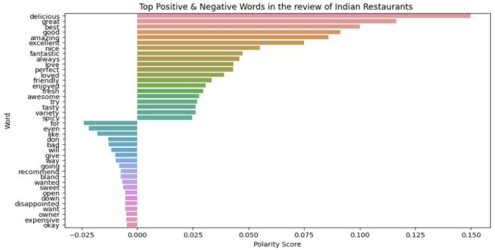

Restaurant Count by US State
6411
4327
2714
2220
2077
914
363
120
3
1
Restaurant Count
AZ
NV
OH
NC
PA
WI
IL
SC
NY
IN
US State


Ratings for Restaurants
491297
244067
209333
177509
Rating
Restaurants Count

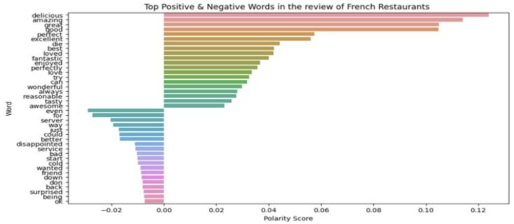

African
American
Chinese
French
Greek
Hawaiian
Indian
Italian
Japanese
Korean
Mediterranean
Mexican
Spanish
Thai
Vietnamese
category
negative
positive
labels
160
150
140
130
120
110
100
90

# POLARITY RESULTS

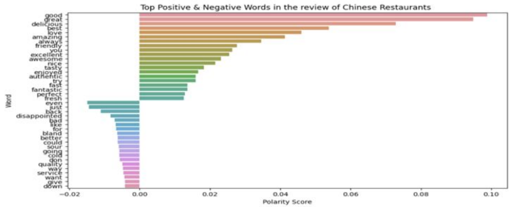
## Indian



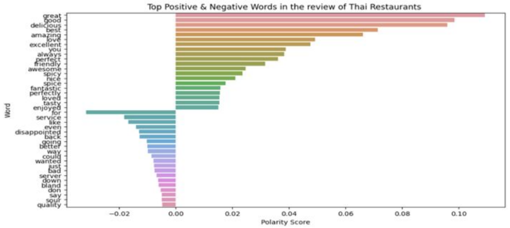

## French



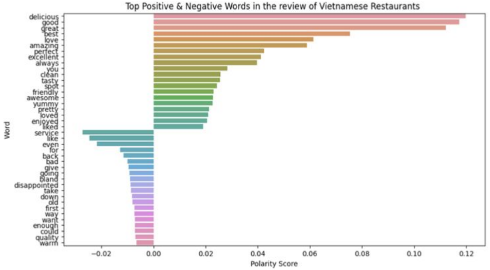

## Chinese



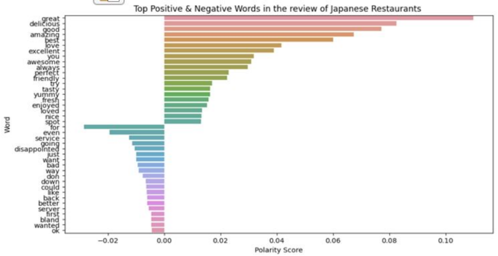

## Thai




## Vietnamese




## Japanese



# Positive and Negative sentiment polarity

| cuisine | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Japanese | friendly | fresh | recommend | fun | reasonable | creative | clean | variety | attentive | tasty |
| Chinese | friendly | fresh | authentic | reasonable | hot | fun | fast | tender | recommend | yummy |
| Vietnamese | friendly | fresh | recommend | reasonable | variety | attentive | fast | comfortable | yummy | authentic |
| Thai | fresh | clean | fast | recommend | reasonable | tender | fancy | refreshing | generous | yummy |
| French | sweet | tender | impeccable | recommend | rich | attentive | romantic | perfection | incredible | friendly |

| cuisine | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Japanese | hard | cold | wrong | slow | bland | dark | expensive | rude | overpriced |
| Chinese | sour | bland | cold | greasy | hard | slow | wrong | rude | overpriced |
| Vietnamese | bland | greasy | expensive | weird | wrong | slow | hard | cold | sour |
| Thai | bland | wrong | hard | slow | expensive | rude | greasy | dirty | weird |
| French | cold | expensive | slow | bland | overpriced | mediocre | wrong | poor | squash |

# RESULTS/OUTCOMES

- The developed model achieved an accuracy of 84% in classifying user comments as positive or negative sentiments.
- The analysis of sentiment revealed that customer service, food quality, ambiance, and pricing were the factors contributing to positive and negative reviews.

# CHALLENGES FACED AND OVERCOMING THEM

- ❖ One of the significant challenges we faced was the imbalance in the dataset, with a majority of reviews being positive.
- ❖ To overcome this, we used techniques such as under-sampling, over-sampling, and SMOTE to balance the dataset.

# CONCLUSION

- The study highlights the importance of customer feedback in the food industry and how businesses can use it to improve their services.
- By developing models to classify customer sentiment and analyzing the factors contributing to positive and negative reviews, businesses can gain a better understanding of their customers'.

# THANK YOU!