Yelp Reviews Text Analysis based on NLP Models

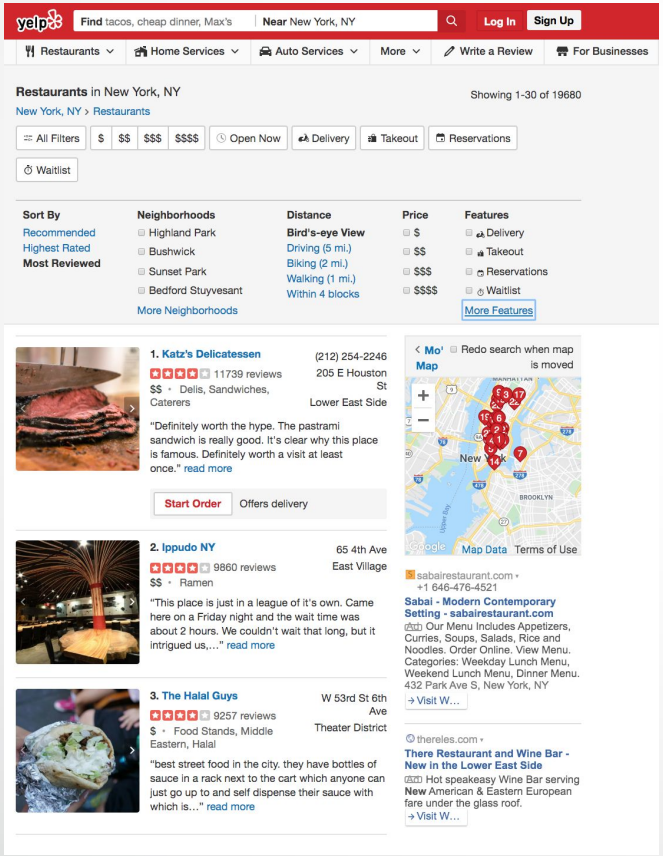
Group K

Names:
Bowen Zhang, Emily Luo,
Shangxian Liu, Tony Zheng

CONTENTS

01 Introduction

1.1 Motivation



For users/businesses:

- For potential customers to view based on the reviews
- Positive feedback from customers may prosper the store businesses
- Tremendously many data about businesses, reviews, and users

For YELP:

- Classify reviews into proper ratings for empowering its recommendation system
- Detect anomaly reviews to protect businesses from malicious competitions
- Assign rating to texts automatically

Training Model

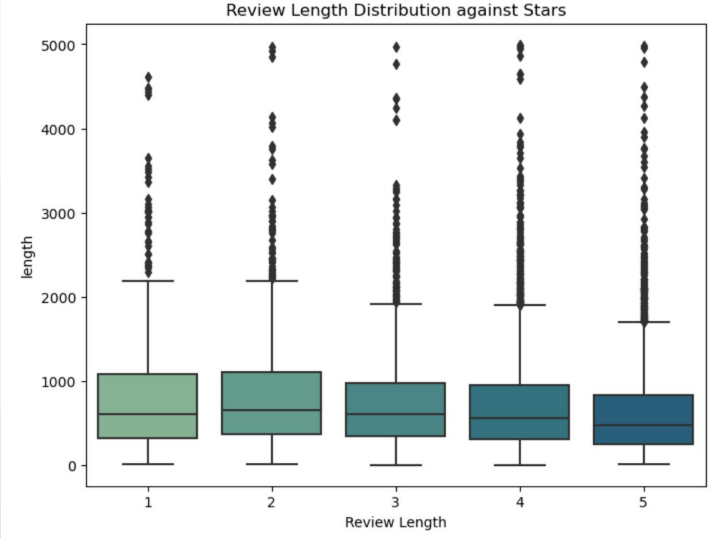
ML/DL
Classification

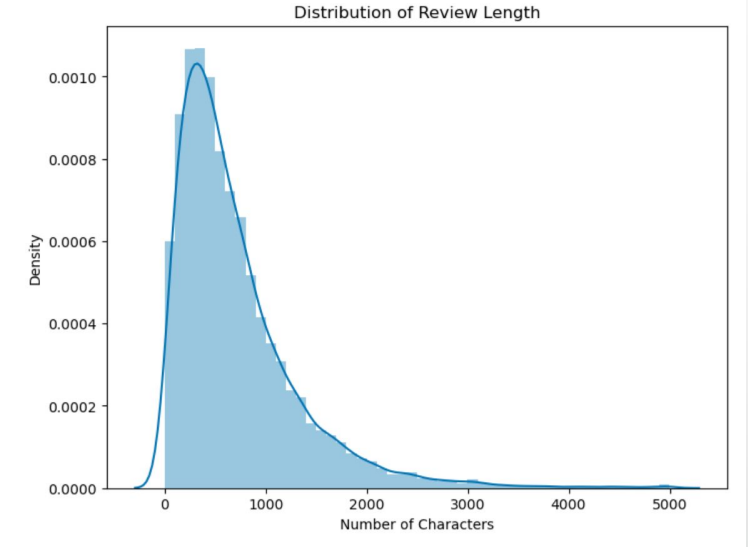
Text data (reviews) → Predicted Classes (1-5 stars)

Data Source: https://www.kaggle.com/datasets/vivekhn/yelp-reviews

1.2 Exploratory Data Analysis

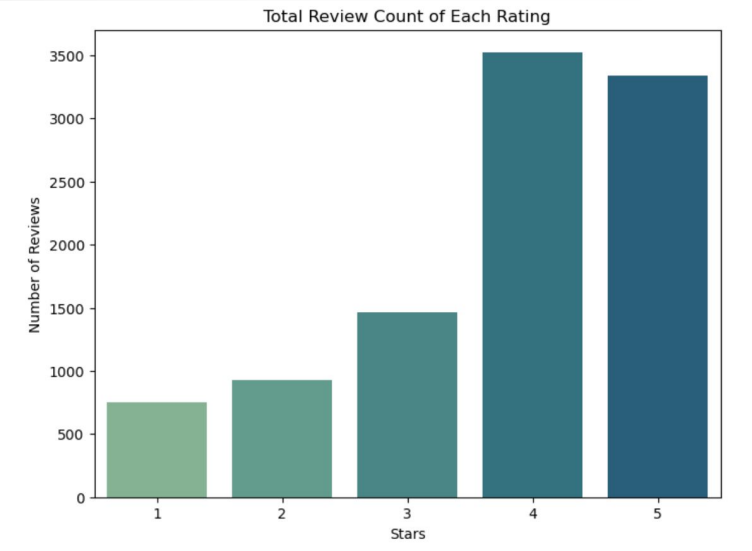
	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny	length
0	9yKzy9PApeiPPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtl8ZkDX5vH5nAx9C3q5Q	2	5	0	889
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbIuQ	0	0	0	1345
2	6oRAC4uyJCsJl1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtfLiobPvh6cDC8JQg	0	1	0	76
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGaISbqqaMHlNnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg	1	2	0	419
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw	0	0	0	469

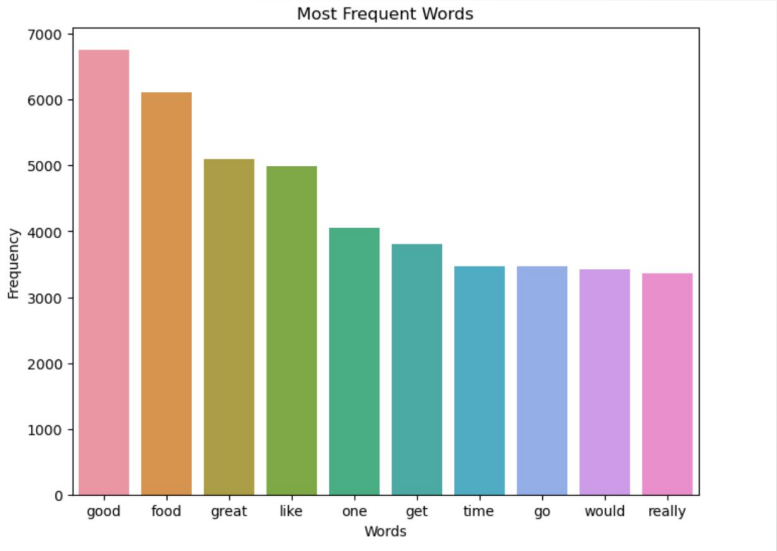




1.2 Exploratory Data Analysis

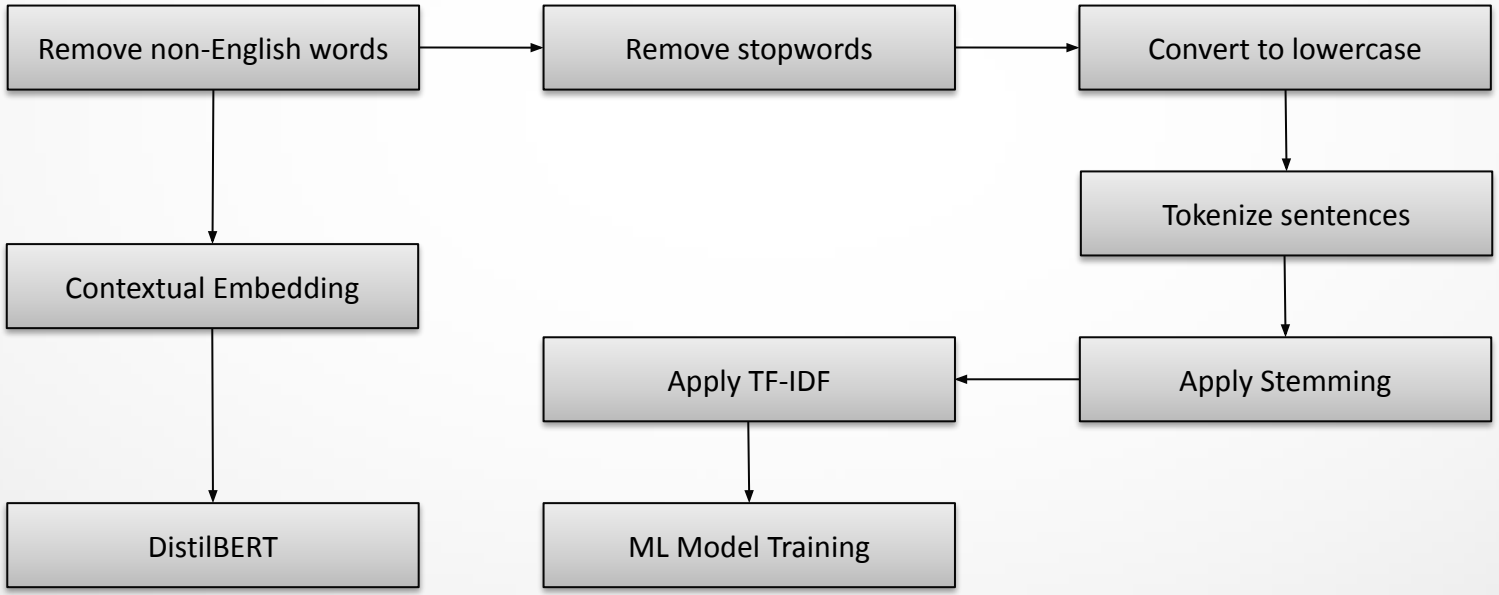
	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny	length
0	9yKzy9PApeiPPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtl8ZkDX5vH5nAx9C3q5Q	2	5	0	889
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbIuQ	0	0	0	1345
2	6oRAC4uyJCsJl1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtfLiobPvh6cDC8JQg	0	1	0	76
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGaISbqqaMHlNnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg	1	2	0	419
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw	0	0	0	469





02 Data Processing

2.1 Text Cleaning



2.2 What is TF-IDF?

- A vectorization method that measures how important a term is to a specific document in the context of the entire corpus that contain the term

- penalizes words that appear frequently in all documents

- gives credits to words that appear frequently in a few documents

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

D1: He is a lazy boy. She is also lazy.

D2: Neeraj is a lazy person.

	He	She	lazy	boy	Neeraj	person
D1	0.06	0.06	0	0.06	0	0
D2	0	0	0	0	0.1	0.1

03 Topic & Sentiment Analysis

3.1 Topic Analysis

LDA (*Latent Dirichlet Allocation*):

- unsupervised clustering of documents
- calculates probability of words belonging to a topic
- iteratively improves assignments of words until converging to a stable state
- we generated three topics for two groups

Five Stars:

- good service
- delicious food
- nice staff

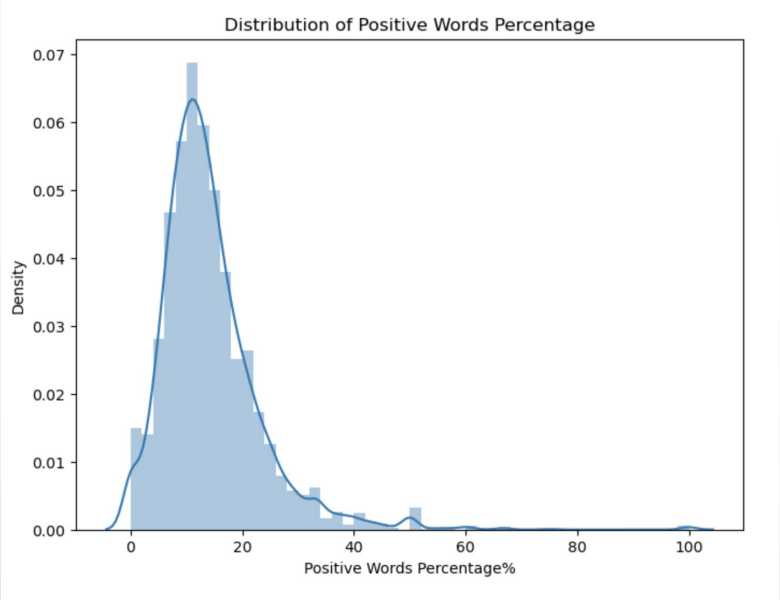
One Star:

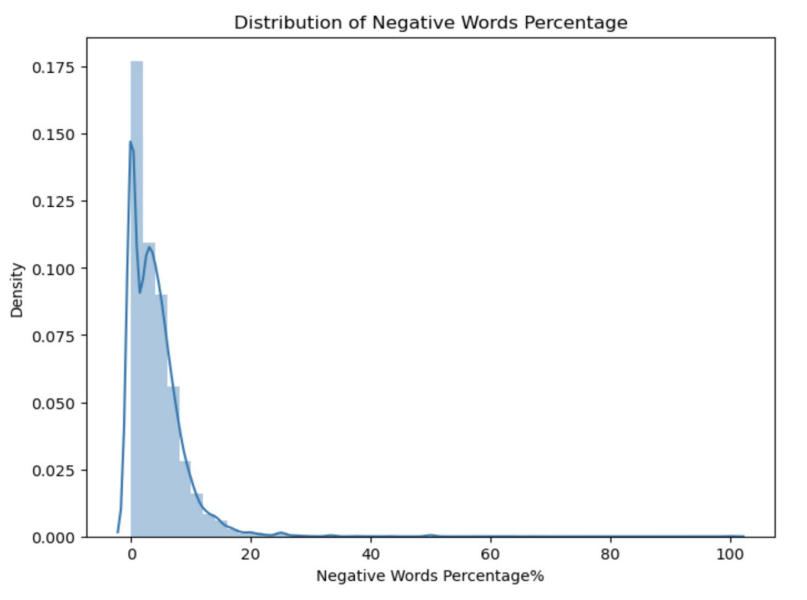
- long waiting time
- bad staff
- bad service

Most popular topic of two groups:

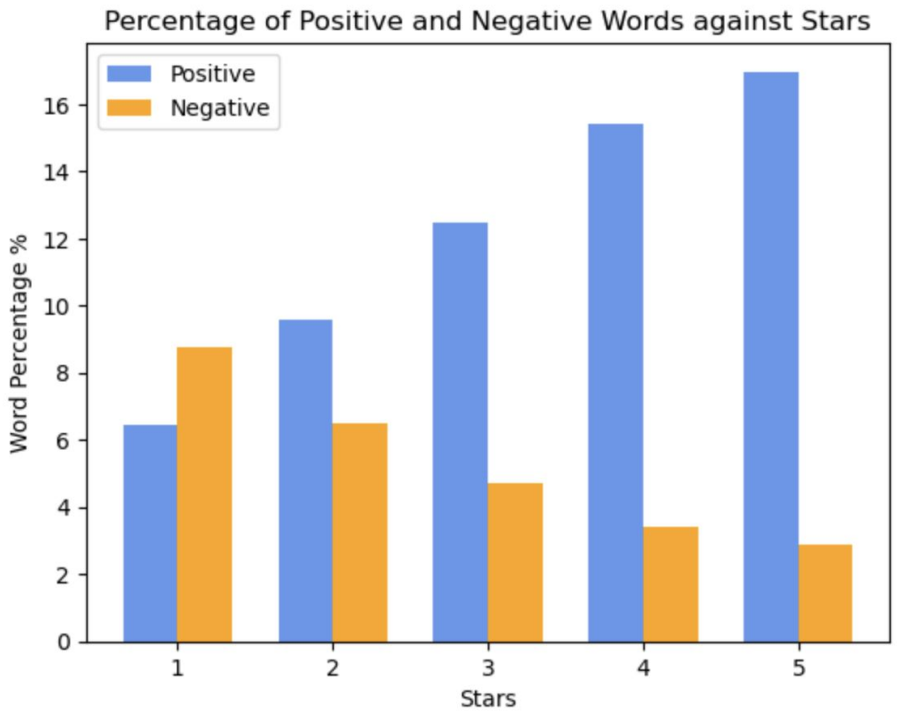
3.2 Sentiment Analysis

- Used pre-determined positive & negative wordlist from *www.ptrckprry.com*
- Calculated each review's positive & negative words percentage
- Customers tend to be friendly even in bad reviews!





3.2 Sentiment Analysis



A simple logistic regression based two variables "positive words %" and "negative words %":

	precision	recall	f1-score	support
1	0.44	0.49	0.46	218
2	0.47	0.06	0.10	265
3	0.10	0.00	0.00	442
4	0.38	0.64	0.48	1087
5	0.48	0.43	0.45	988
accuracy			0.42	3000
macro avg	0.37	0.32	0.30	3000
weighted avg	0.38	0.42	0.37	3000

Accuracy: 0.4156666666666667

04 Prediction Models

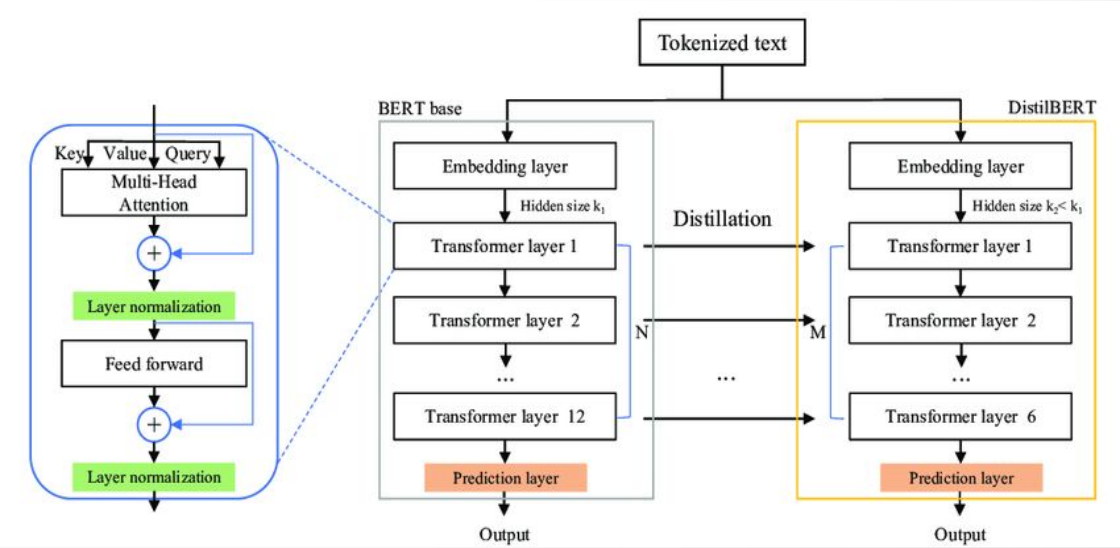
4.1 Transformer-Based Models

BERT (Bidirectional Encoder Representations from Transformers)

- Pre-trained deep learning model
- Transformer architecture
- Bidirectional context
- Fine-tuning for specific tasks

DistilBERT

- Smaller, faster variant of BERT
- Knowledge distillation process
- Student-teacher model
- Retains 97% of BERT's performance



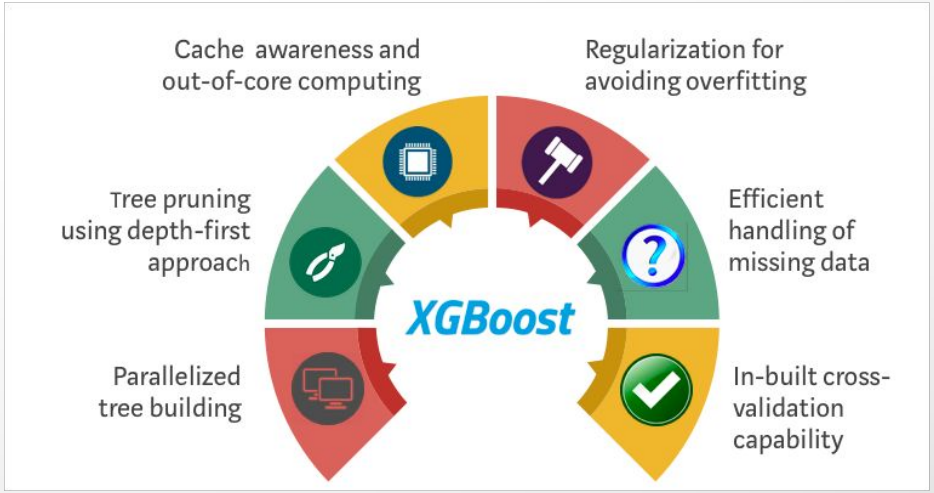
We choose to only use DistilBert in this project.

4.2 XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful, scalable, and efficient machine learning algorithm that utilizes gradient boosted decision trees to perform classification tasks. It works by iteratively combining weak learners (shallow decision trees) to create a strong, predictive model.

Key Features

1. Regularization
2. Parallel Processing
3. Early Stopping
4. Pruning
5. Handling Missing Values
6. Customizable Loss Functions



4.3 Other Machine Learning Models

- Support Vector Machine
- Logistic Regression
- Naive Bayes
- Random Forest

4.4 Comparison of Machine Learning Models

Hyperparameter Tuning criteria: 5-fold Cross Validation (GridSearchCV)

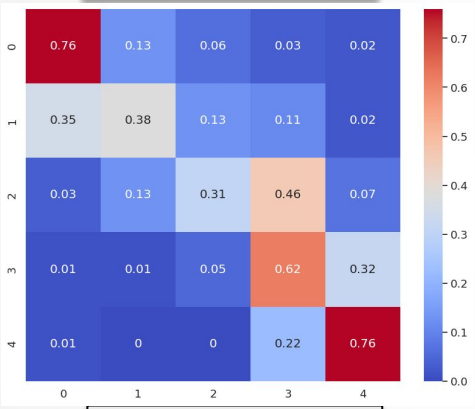
Models	Best Hyperparameter	Accuracy Score	F1 Score
DistilBert	Batch size = 4, # of Epochs = 2	**0.609**	**0.597**
XGBoosting	Max_depth = 4 N_estimators = 300 Learning_rate = 0.1	0.517	0.498
Random Forest	n_estimators = 400 max_depth = 40 min_samples_split = 8 min_samples_leaf = 3	0.467	0.384
SVM	C=1 Gamma = 1 kernel = linear	**0.552**	**0.542**
Logistic Regression	C=2 solver = sag	0.532	0.519
Naive Bayes	-	0.434	0.339

Best Model

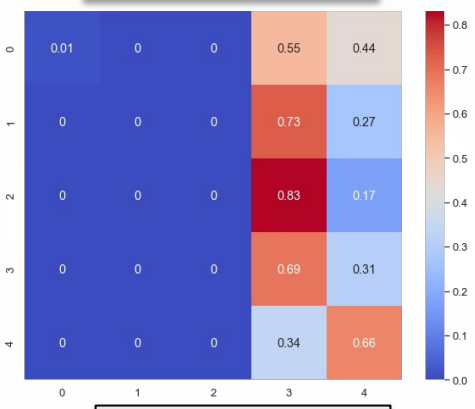
Second Best Model

Confusion Matrix of machine learning models

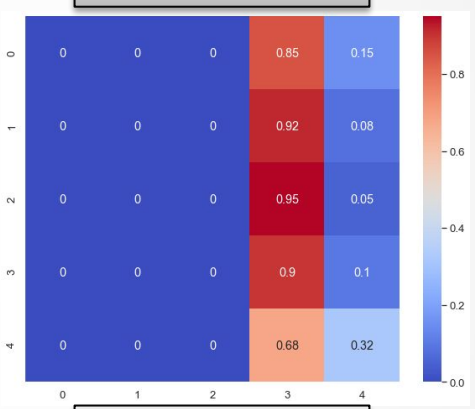
DistilBert



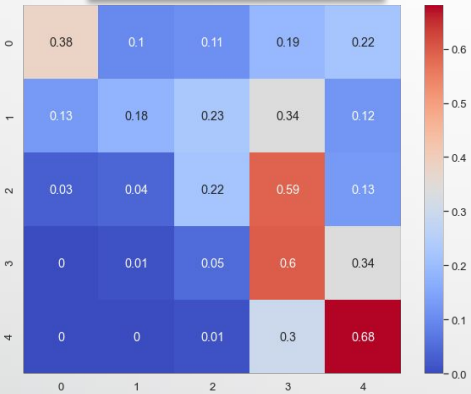
Random Forest



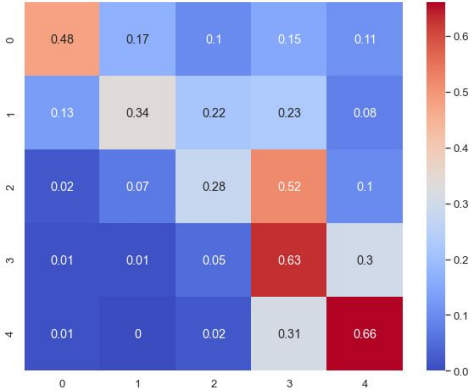
Naive Bayes



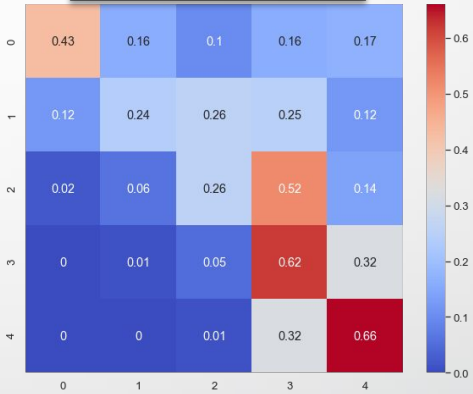
XGBoost



SVM



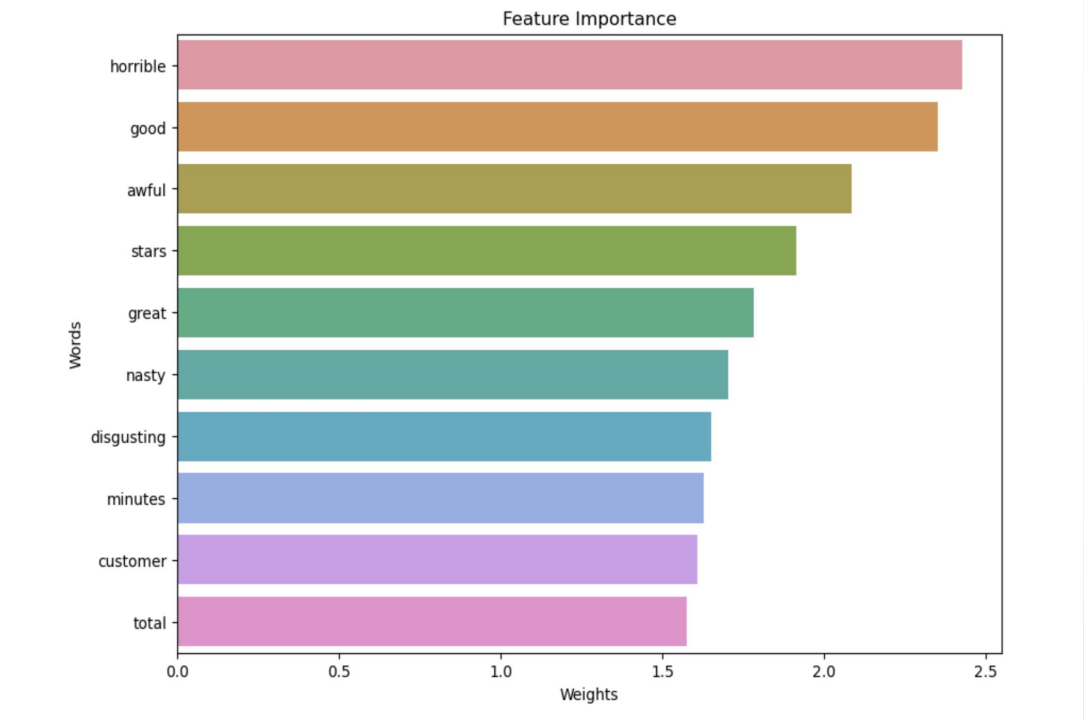
Logistic Regression



05 Conclusion

5.1 Feature Importance

	Words	Weights
0	horrible	2.426076
1	good	2.349551
2	awful	2.085384
3	stars	1.912846
4	great	1.783368
5	nasty	1.703266
6	disgusting	1.650621
7	minutes	1.628923
8	customer	1.608882
9	total	1.573600



5.2 Conclusions and Limitations

Summary of Model Fit

From this point, the NLP driven DistilBERT model has the best fit for the Yelp text data. The accuracy 0.61 looks fine for this 5-classes classification problem.

Overall, transformer-based model is more advanced than machine learning models in this question. We hope our work could give some insights for further work in Yelp review rating predictions.

Future Improvement

Limitations:

- Some machine learning models do not work well
- Significant amount of time & memory requirement

Solutions:

- Handle data imbalance
- Increase the size of the dataset
- Use more powerful GPU (eg. RTX 4090 ti)

Contribution

Bowen Zhang: found the dataset, data visualization, text preprocessing, feature importance calculation, slides/report editing

Emily Luo: LDA topic analysis, sentiment analysis, hyperparameter tuning, data visualization, script compiling, slides/report editing

Shangxian Liu: created ppt template, hyperparameter tuning, metric calculation, script compiling, slides/report editing

Tony Zheng: DistilBERT model building, hyperparameter tuning, ML model building, slides/report editing

Reference

1. Liu, Z. (2020). Yelp review rating prediction: Machine learning and deep learning models. arXiv preprint arXiv:2012.06690.

2. Siqi Liu. Sentiment analysis of yelp reviews: A comparison of techniques and models, 2020.

3. Boya Yu, Jiaxu Zhou, Yi Zhang, and Yunong Cao. Identifying restaurant features via sentiment analysis on yelp reviews, 2017.