

# Yelp Reviews Text Analysis based on NLP Models



**Group K**

Names:

Bowen Zhang, Emily Luo,  
Shangxian Liu, Tony Zheng

# CONTENTS



## Introduction



## Data Processing



## Sentiment & Topic Analysis



## NLP Models



## Conclusion

1

- Motivation
- Exploratory data analysis

2

- Remove stopwords & non-English words
- Apply Stemming
- Convert cases
- Apply TF-IDF

3

- Sentiment Analysis
- Topic Analysis (LDA)

4

- DistilBERT
- XGBoost
- SVM
- Random Forest
- Logistic Regression
- Naive Bayes

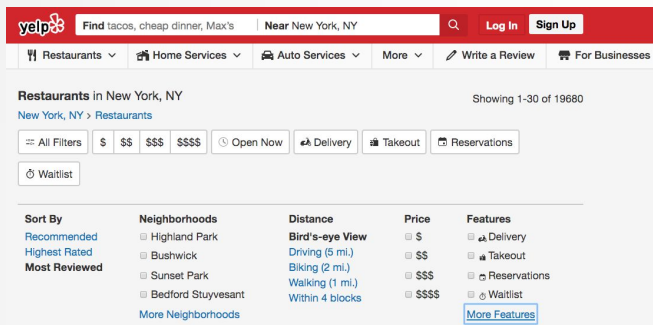
5

- Conclusions & limitations

**01**

**Introduction**

# 1.1 Motivation



Restaurants in New York, NY

Showing 1-30 of 19680

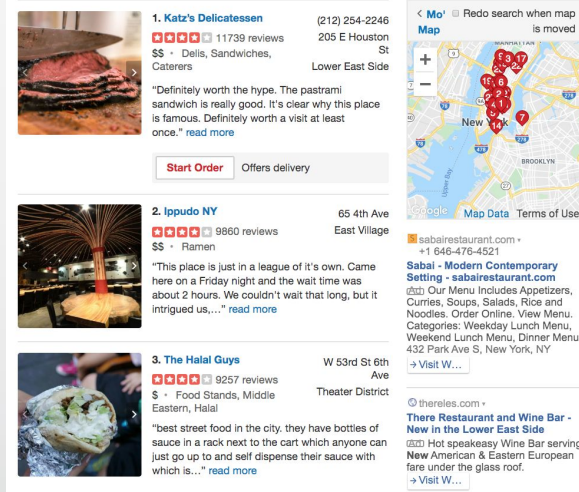
Sort By: Recommended, Highest Rated, Most Reviewed

Neighborhoods: Highland Park, Bushwick, Sunset Park, Bedford Stuyvesant, More Neighborhoods

Distance: Bird's-eye View, Driving (5 mi.), Biking (2 mi.), Walking (1 mi.), Within 4 blocks

Price: \$, \$\$, \$\$\$, \$\$\$\$

Features: Delivery, Takeout, Reservations, Waitlist



1. **Katz's Delicatessen** (212) 254-2246  
11739 reviews  
205 E Houston St  
Lower East Side  
"Definitely worth the hype. The pastrami sandwich is really good. It's clear why this place is famous. Definitely worth a visit at least once." [read more](#)  
[Start Order](#) [Offers delivery](#)

2. **Ippudo NY** 65 4th Ave  
9860 reviews  
East Village  
"This place is just in a league of it's own. Came here on a Friday night and the wait time was about 2 hours. We couldn't wait that long, but it intrigued us...." [read more](#)

3. **The Halal Guys** W 53rd St 6th Ave  
9257 reviews  
Theater District  
"best street food in the city. they have bottles of sauce in a rack next to the cart which anyone can just go up to and self dispense their sauce with which is..." [read more](#)

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)

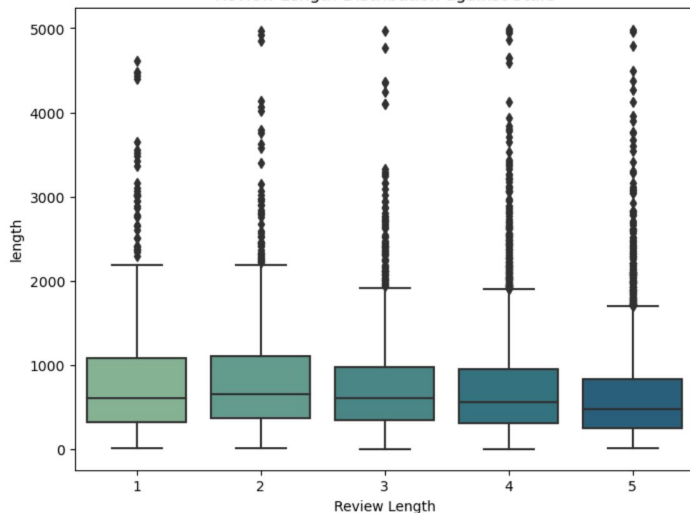
Predicted Classes (1-5 stars)

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

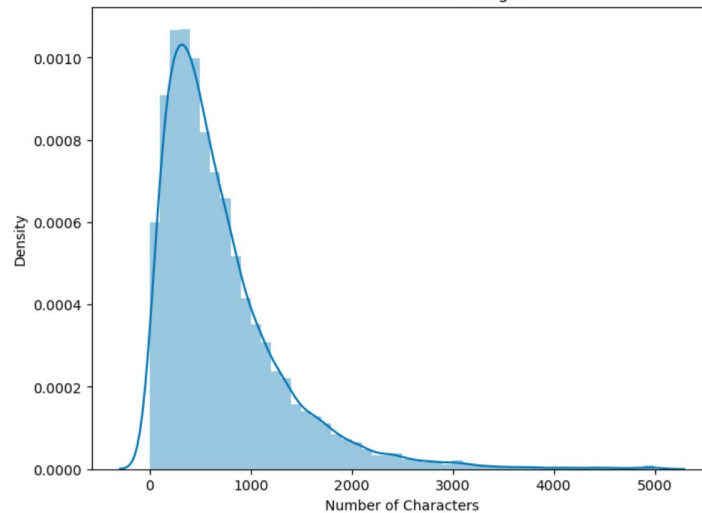
## 1.2 Exploratory Data Analysis

	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny	length
0	9yKzy9PApeiPPOUJEtnvgk	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtI8ZkDX5vH5nAx9C3q5Q	2	5	0	889
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbluQ	0	0	0	1345
2	6oRAC4uyJCslJ1X0WZpVSA	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-WvGalSbqqaMHINnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetI9T0NcROGOyFfughhg	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

Review Length Distribution against Stars



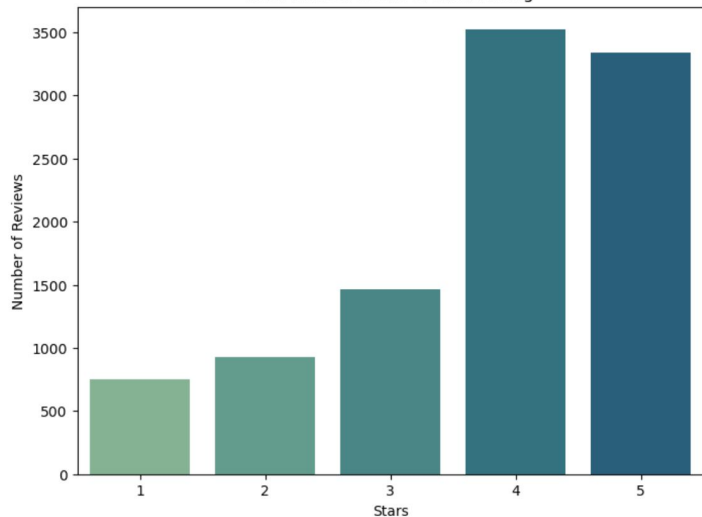
Distribution of Review Length



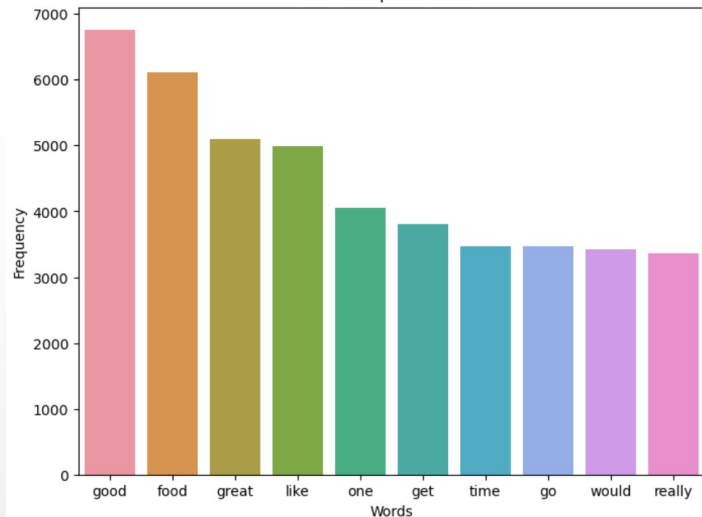
## 1.2 Exploratory Data Analysis

	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	rLtI8ZkDX5vH5nAx9C3q5Q	2	5	0	889
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbluQ	0	0	0	1345
2	6oRAC4uyJCsjI1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtfLiobPvh6cDC8JQg	0	1	0	76
3	_1QQZuf4zZ0yFCvXc0o6Vg	2010-05-27	G-WvGalSbqqaMHINnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetI9T0NcROGOyFfughhg	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

Total Review Count of Each Rating



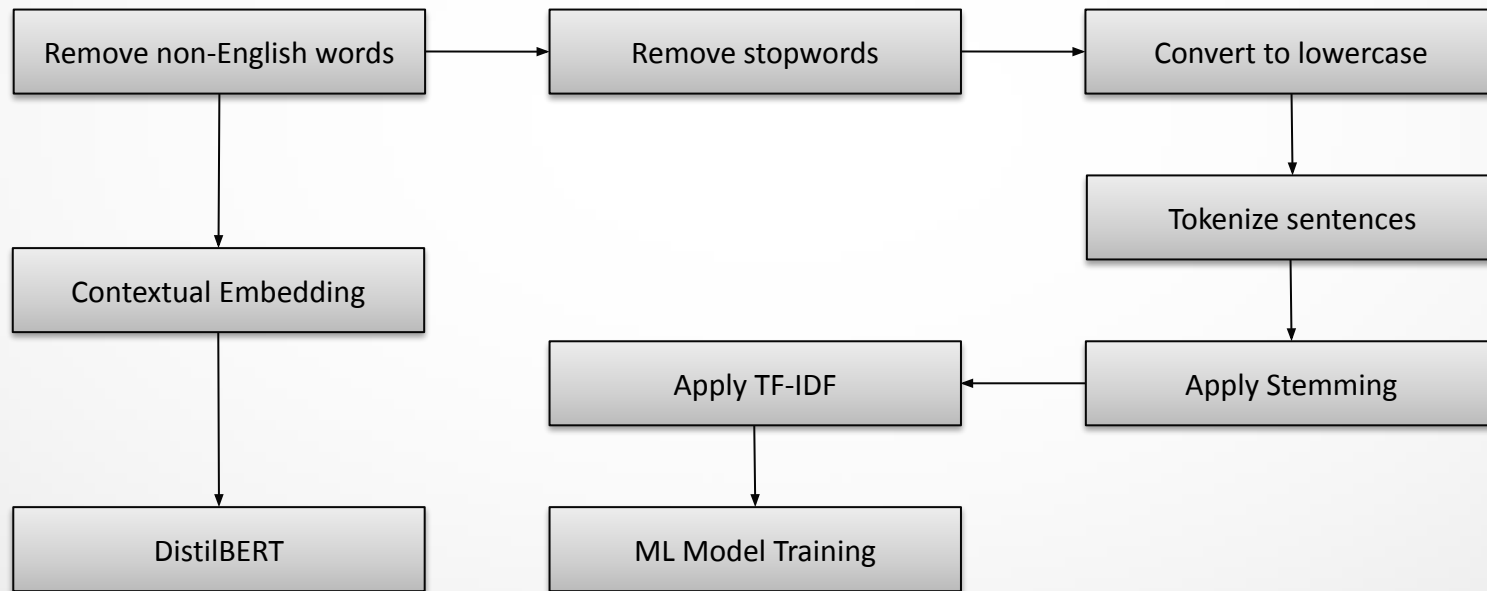
Most Frequent Words



**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

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

$tf_{ij}$  = 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:

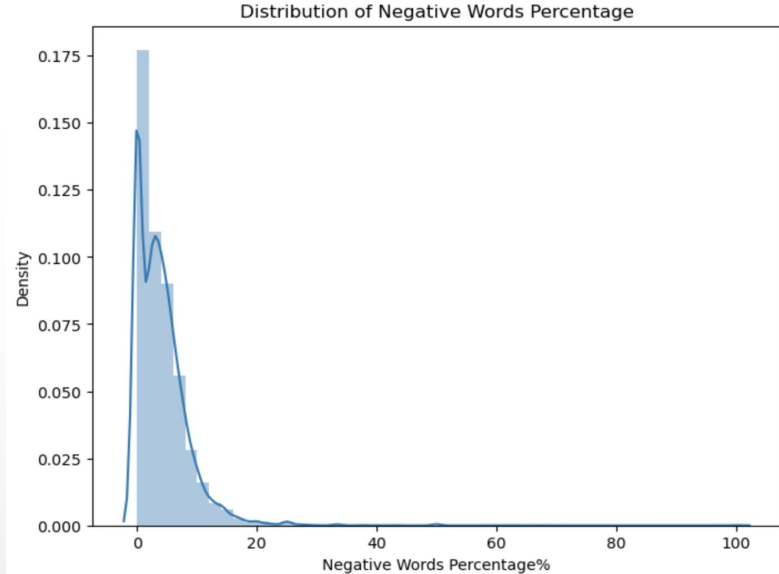
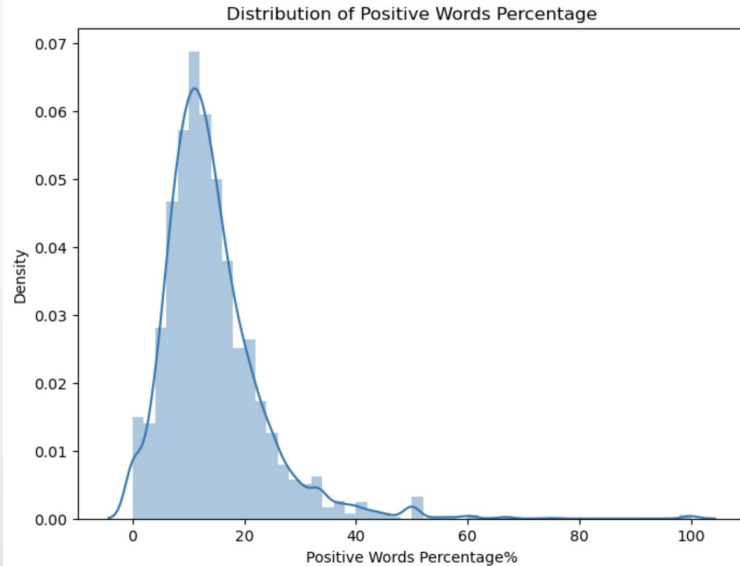
- long waiting time
- bad staff
- bad service

Most popular topic of two groups:

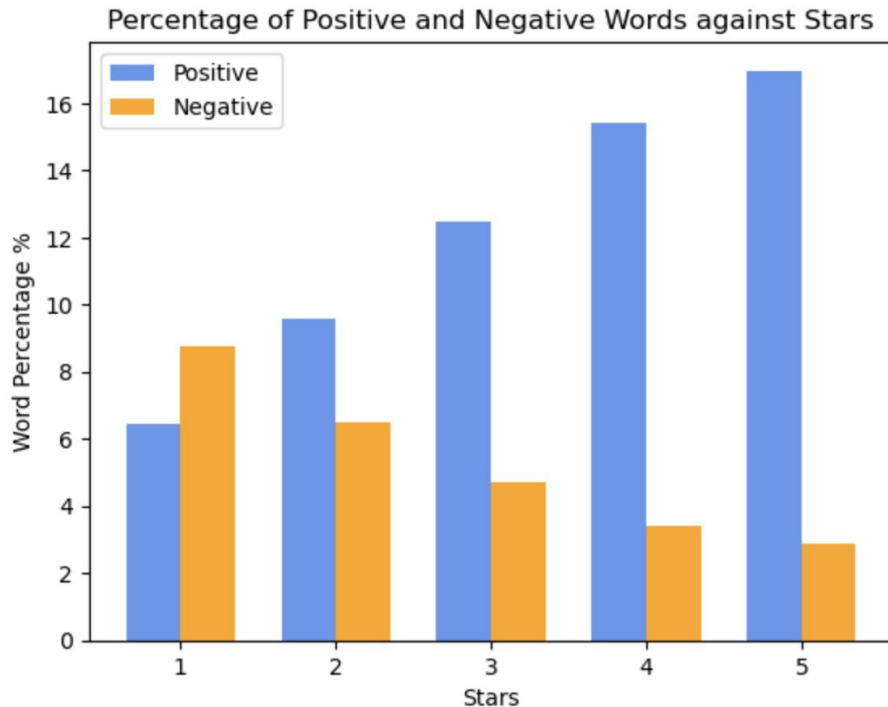


## 3.2 Sentiment Analysis

- Used pre-determined positive & negative wordlist from [www.ptreckprry.com](http://www.ptreckprry.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**

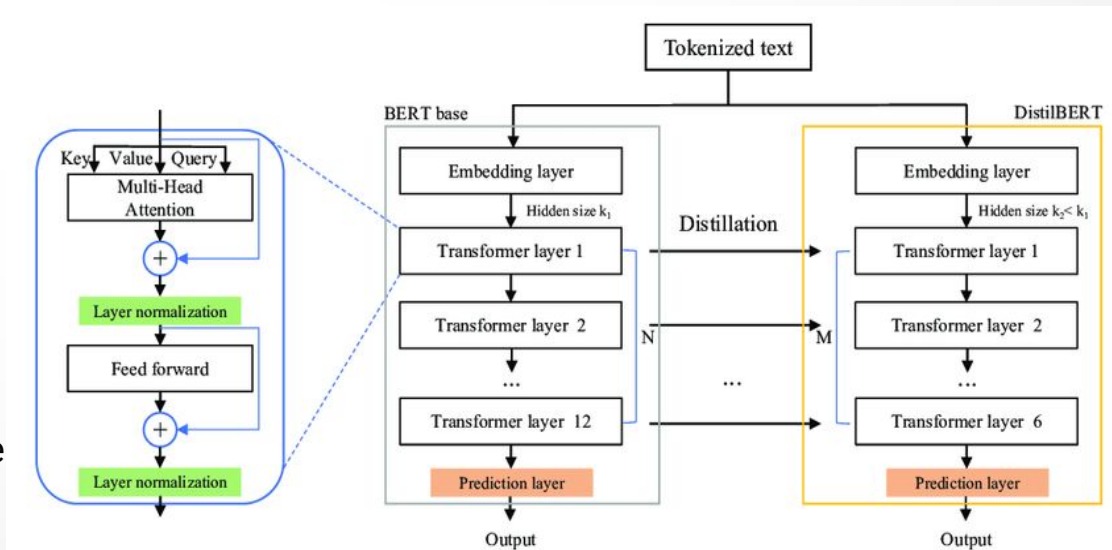
## 4.1 Transformer-Based Models

### 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



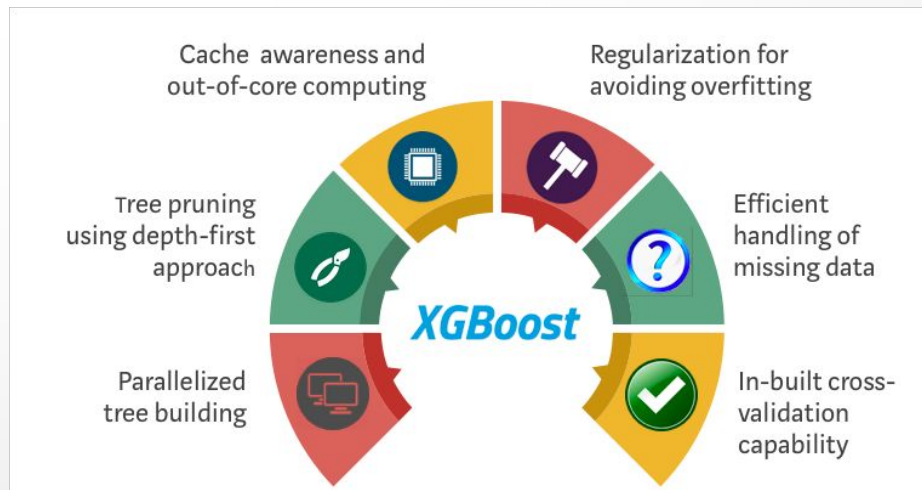
We choose to only use DistilBert in this project.

## 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

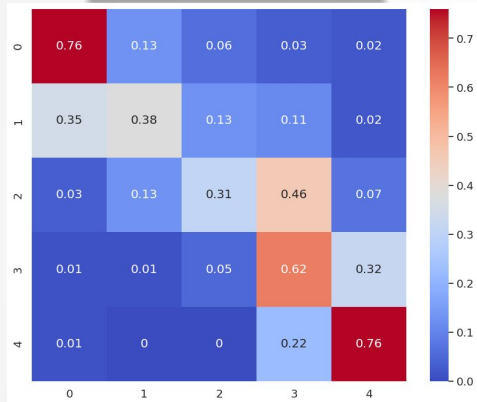
## 4.4 Comparison of Machine Learning Models

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

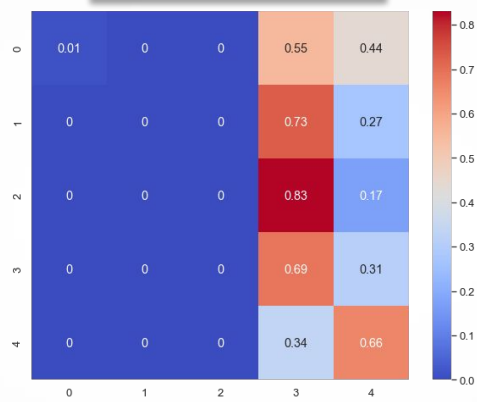
Models	Best Hyperparameter	Accuracy Score	F1 Score	
DistilBert	Batch size = 4, # of Epochs = 2	<b>0.609</b>	<b>0.597</b>	→ Best Model
XGBoosting	Max_depth = 4 N_estimators = 300 Learning_rate = 0.1	0.517	0.498	
Random Forest	n_estimators = 400 max_depth = 40 min_samples_split = 8 min_samples_leaf = 3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

# Confusion Matrix of machine learning models

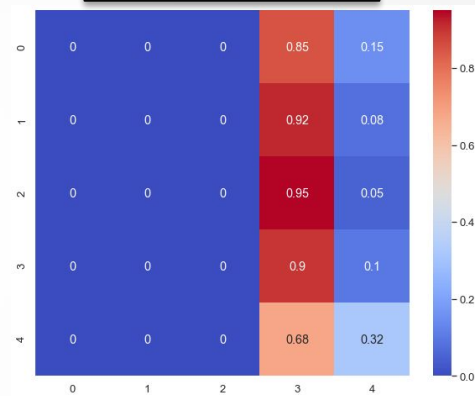
DistilBert



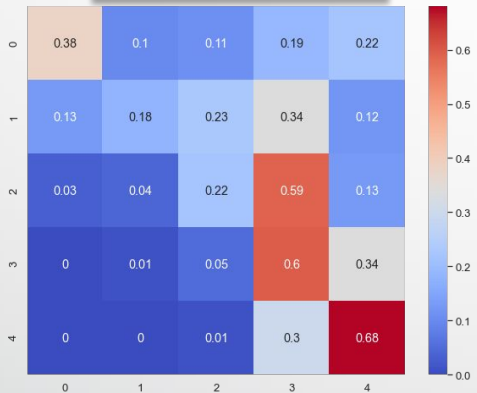
Random Forest



Naive Bayes



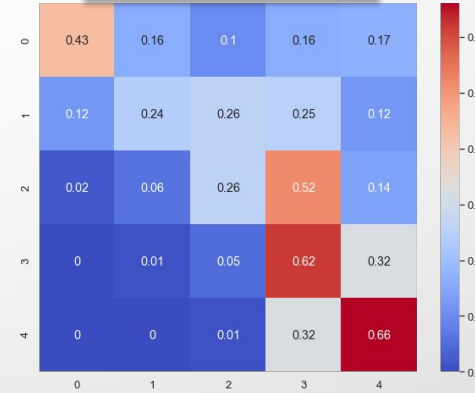
XGBoost



SVM



Logistic Regression

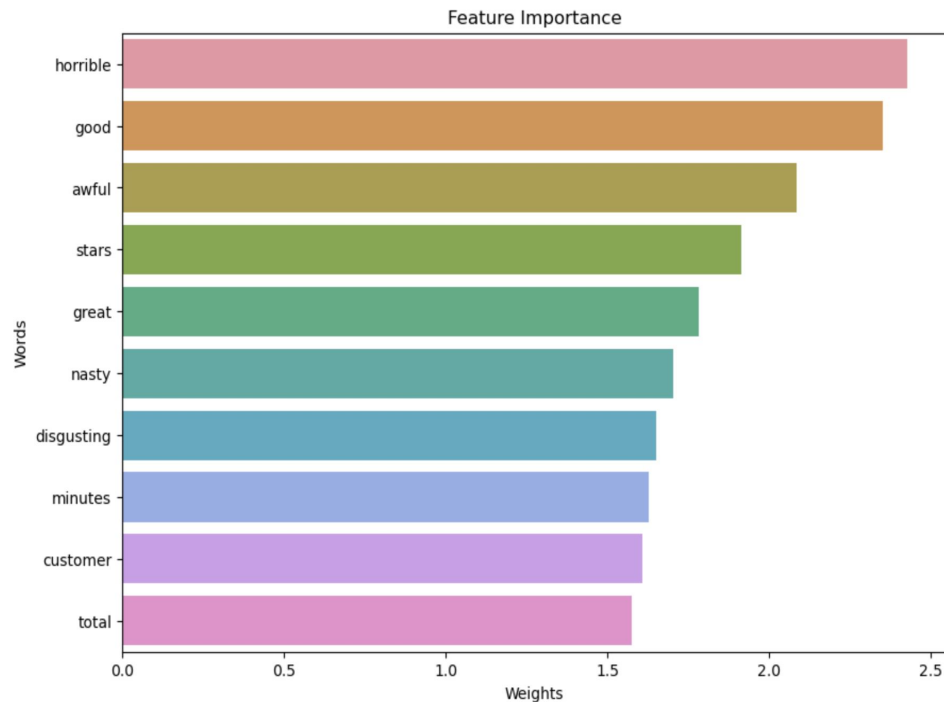


**05**

**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

The background of the slide shows the silhouettes of people walking in a transit station, possibly a subway or train station, with a red rectangular bar at the top center containing the title.

# 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.

“