Northeastern University
CS 6220 Data Mining
Spring 2018

RESTAURANT RECOMMENDATIONS USING YELP RATINGS AND REVIEWS

Submitted By
Karan Tyagi and Shraddha Shah

CS6220 - DATA MINING

PROJECT VISION

- Yelp contains review data of various restaurants in a city and helps user in choosing a restaurant. But, it doesn't recommend any restaurant to an user.
- So, in this project we have used review text for recommending restaurants to the users
- We have investigated features of Yelp data to build models for rating prediction and recommendation tasks

DATASET

The dataset used for this task was obtained from the Yelp dataset challenge, which consists of 1.6M reviews and 61k businesses.

- **FEATURES ARE AS FOLLOWS:**

- Business

business_id	*categories*	*name*	*city*	*state*	*postal_code*	*latitude*	*longitude*	*stars*	*review_count*

- Review

review_id	*business_id*	*user_id*	*text*	*stars*

DATASET VIEW

- Business

	business_id	categories	name	city	state	postal_code	latitude	longitude	stars	review_count
0	b'FYWN1wneV18bWNgQjJ2GNg'	['Dentists', 'General Dentistry', 'Health & Me...	b'Dental by Design'	b'Ahwatukee'	b'AZ'	b'85044'	33.330690	-111.978599	4.0	22
1	b'He-G7vWjzVUysIKrfNbPUQ'	['Hair Stylists', 'Hair Salons', 'Men's Hair S...	b'Stephen Szabo Salon'	b'McMurray'	b'PA'	b'15317'	40.291685	-80.104900	3.0	11
2	b'KQPW8lFf1y5BT2MxiSZ3QA'	['Departments of Motor Vehicles', 'Public Serv...	b'Western Motor Vehicle'	b'Phoenix'	b'AZ'	b'85017'	33.524903	-112.115310	1.5	18

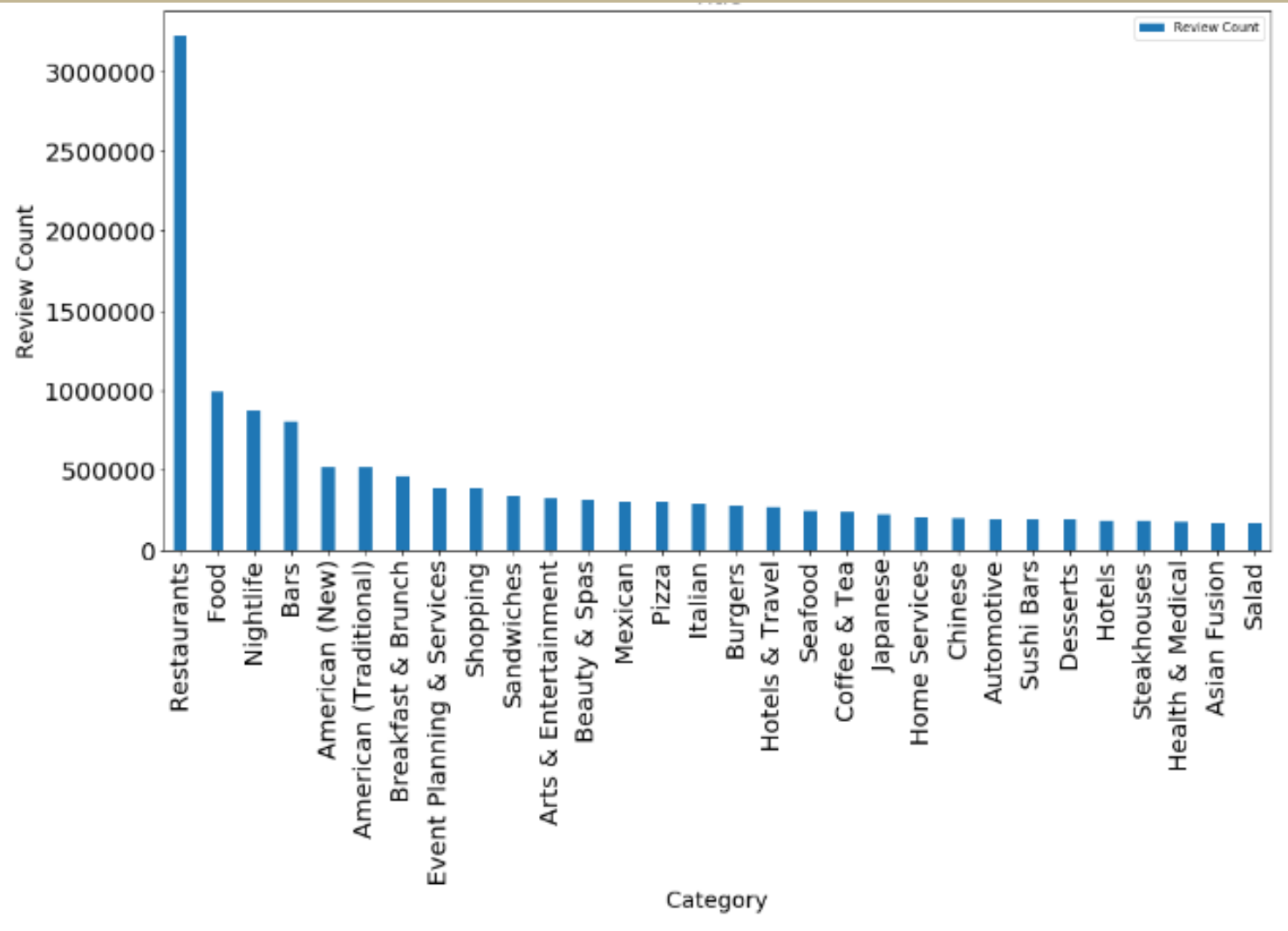
- Review

	review_id	business_id	user_id	text	stars
0	QgSf2JvYz-M4PU2yuJjxNQ	9Jc3W0aR9Xf2gcHI0rEXsw	nOTI4aPC4tKHK35T3bNauQ	b'After being scared away from Rock & Rita\'s,...	1
1	gN6GARS_BRr5UX2D3WAH0w	xVEtGucSRLk5pxxN0t4i6g	nOTI4aPC4tKHK35T3bNauQ	b"We got recommendations for this place from m...	5
2	t4oXDPN4S4USIhBGpuSD8A	2LZGeJy8qByYKB71ML-jcw	nOTI4aPC4tKHK35T3bNauQ	b"We got a coupon to eat here when we checked ...	2

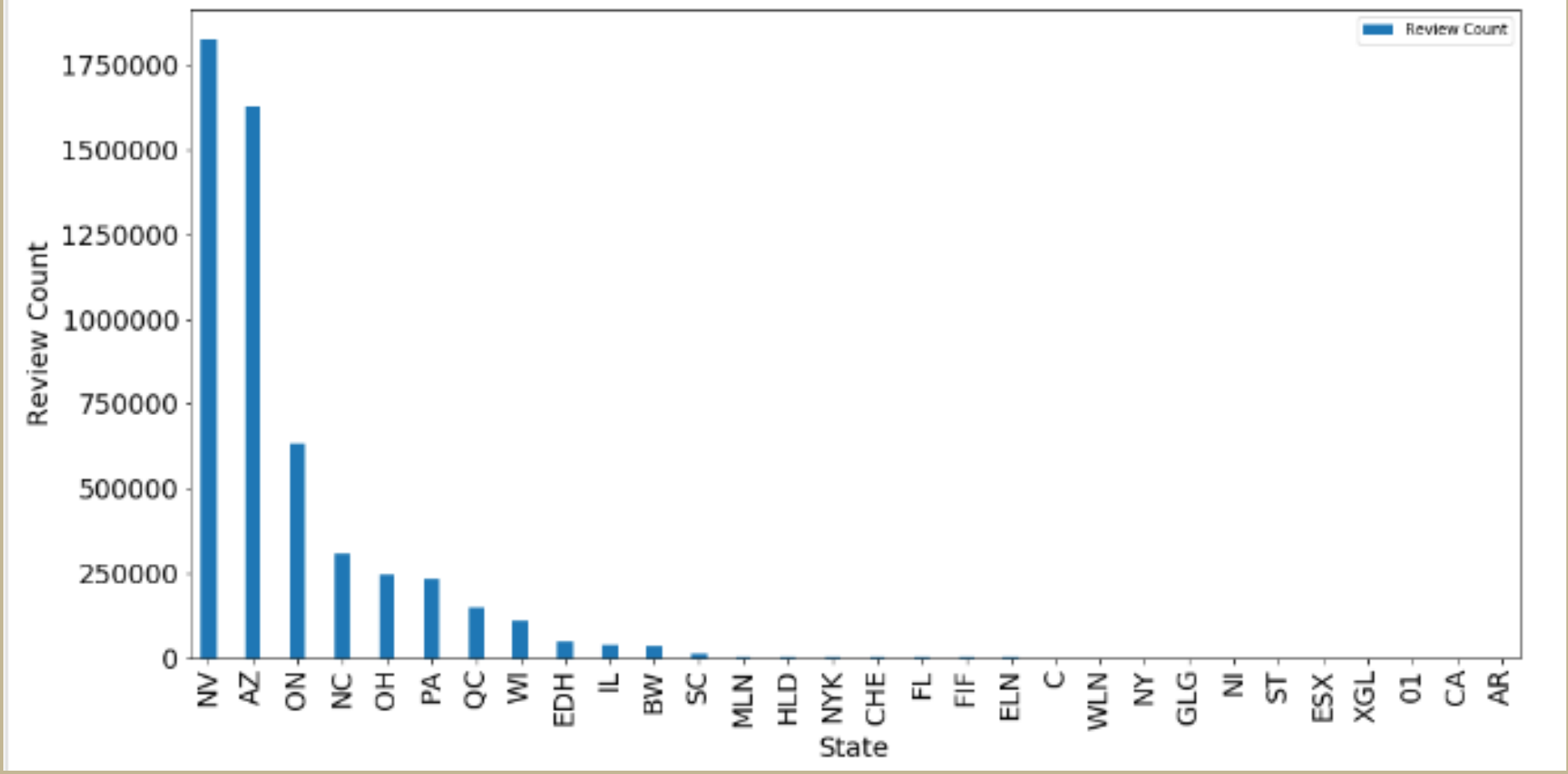
EXPLORATORY ANALYSIS

- Number of reviews for the category ‘Restaurants’ was the highest.
- Number of reviews for the state of ‘Nevada’ was the highest.
- Number of reviews for the city of ‘Las Vegas’ was the highest.
- Majority of the reviews are of length 100-200 words.
- Range of 100 to 200 words is a stable range of length of words in a review for rating prediction

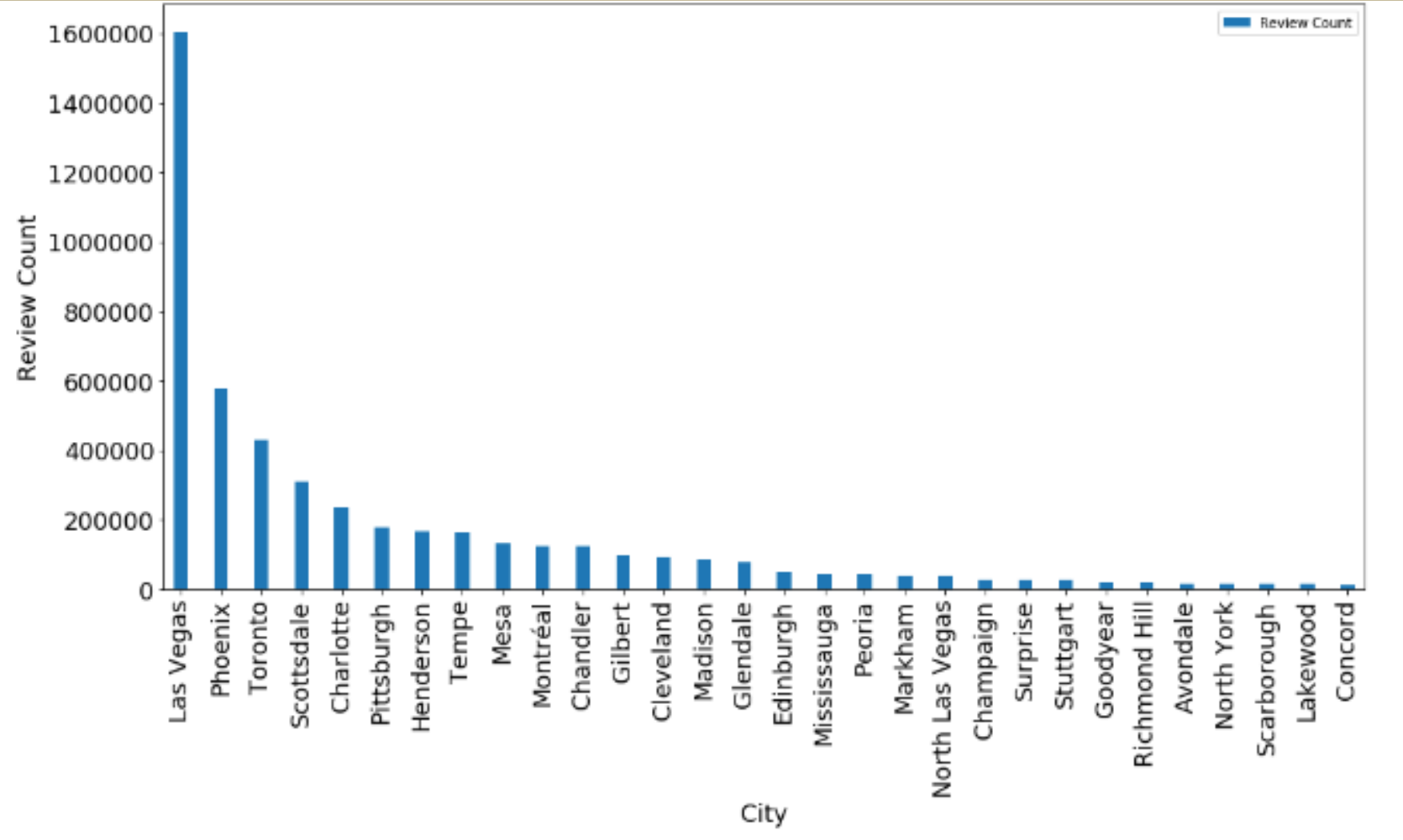
EXPLORATORY ANALYSIS - BY CATEGORY



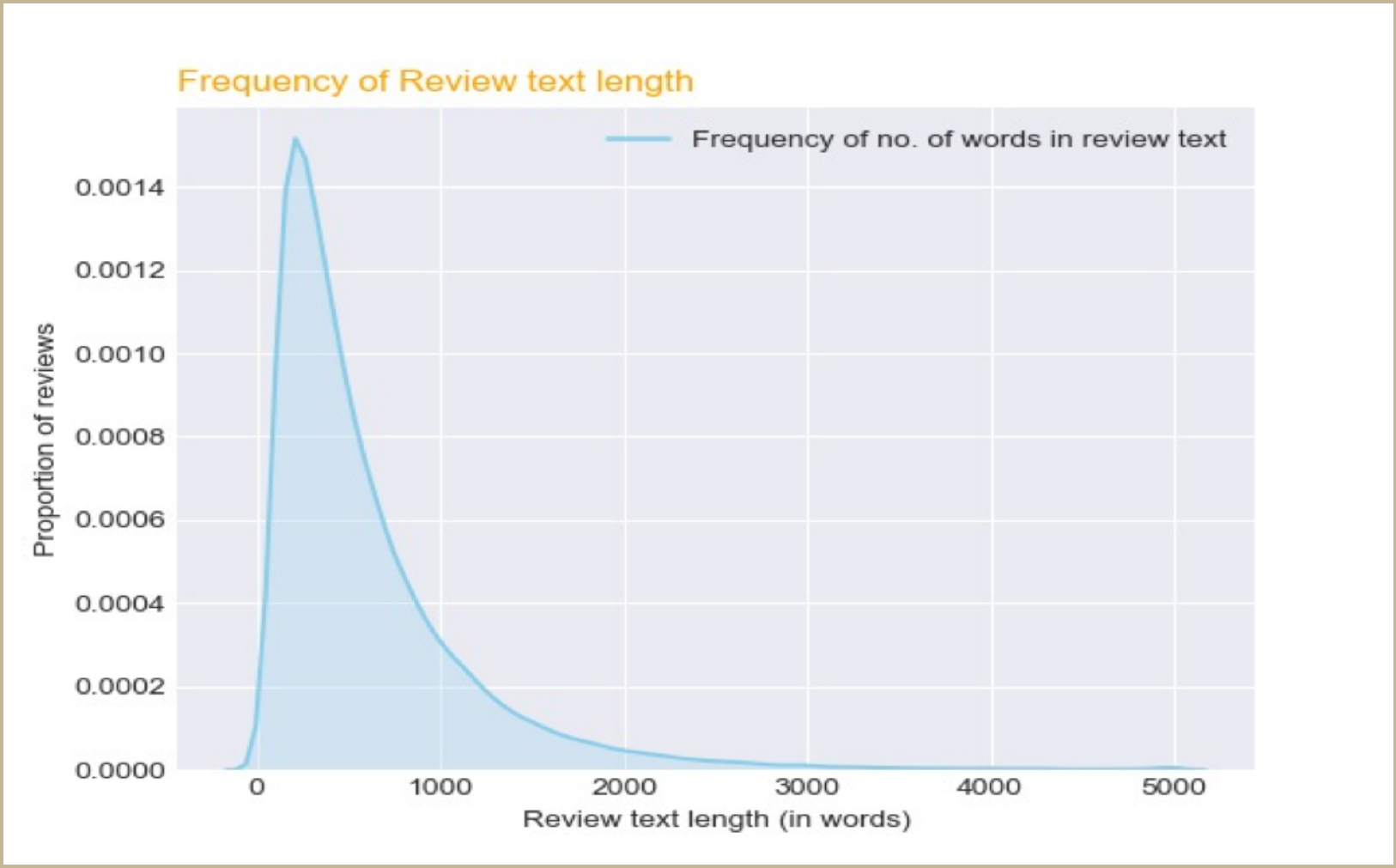
EXPLORATORY ANALYSIS - BY STATE



EXPLORATORY ANALYSIS - BY CITY



EXPLORATORY ANALYSIS - BY REVIEW LENGTH



DATASET REDUCTION

AFTER EXPLORATORY ANALYSIS, WE TRIMMED OUR DATASET TO OVERCOME SPARSITY PROBLEM OF YELP DATASET.

WE SELECTED INSTANCES WITH :

- Business category as ‘Restaurants’
- State as ‘Nevada’
- City as ‘Las Vegas’
- Length of reviews between 100 and 200 words

TEXT PREPROCESSING

- Text Preprocessing:
  - Stopping (using Stoplist form NLTK)
  - Stemming
  - case normalization
  - de-punctuation (removing punctuations)
- Tokenization
  - unigrams
  - bigrams (to consider terms like "not good")
  - trigrams (to consider terms like "not so great")

FEATURE SELECTION

- For converting words to vectors we used the following approaches:

  - Bag of words approach (Term frequencies)

  - tf.idf

- Models with tf.idf performed better;

FEATURE REDUCTION

- To improve prediction accuracy we selected the following as our feature set:
  - top 1000 most commonly occurring unigrams
  - top 1500 most commonly occurring bigrams
- tf.idf of 1500 most common bigrams gave better results.

PREDICTIVE TASKS

THERE ARE TWO MAJOR TASKS IN OUR PROJECT

- Predict rating from review text alone
- Clustering users based on similarity and recommending restaurants to them

TASK 1 : PREDICT RATING FROM REVIEW TEXT

WE IMPLEMENTED AND COMPARED THE FOLLOWING TWO MODELS:

- Naive Bayes Classifier
- Linear Support Vector Machine Classifier

NAIVE BAYES CLASSIFIER

TEXT PRE-PROCESSING

- Removed punctuations, stop words and tokenized the reviews
- Converted each review into a vector using the bags-of-words approach

TRAINING THE MODEL

- Split the dataset into training and test set by 80:20 ratio
- Build a Multinomial Naive Bayes model and fit it to our training set

TESTING AND EVALUATING THE MODEL

- Tested the model for 5 classes(1,2,3,4,5 rating)
- Tested the model for 2 classes(1 and 5 rating)- converted 2,3 to 1 and 4 to 5

NAIVE BAYES CLASSIFIER

USING 2 CLASSES (1 AND 5 STAR RATING)

	precision	recall	f1-score	support
1	0.89	0.88	0.88	3381
5	0.88	0.89	0.88	3432
avg / total	0.88	0.88	0.88	6813

USING 5 CLASSES (1,2,3,4 & 5 STAR RATING)

	precision	recall	f1-score	support
1	0.67	0.80	0.73	2395
2	0.39	0.12	0.19	1388
3	0.49	0.58	0.53	2387
4	0.50	0.48	0.49	2417
5	0.67	0.70	0.68	2413
avg / total	0.56	0.58	0.56	11000

LINEAR SUPPORT VECTOR MACHINE CLASSIFIER

TEXT PRE-PROCESSING

- Removed punctuations, stop words and tokenized the reviews
- Converted each review into a vector using tf-idf

TRAINING THE MODEL

- Split the dataset into training and test set by 80:20 ratio
  Build a Multiclass SVM classifier and fit it to our training set

TESTING AND EVALUATING THE MODEL

- Tested the model for 5 classes(1,2,3,4,5 rating) and then for binary sentiment;
- SVM performed better than Naive Bayes Classifier in terms of accuracy, precision and recall metrics.

LINEAR SUPPORT VECTOR MACHINE CLASSIFIER

USING 2 CLASSES (1 AND 5 STAR RATING)

	precision	recall	f1-score	support
n	0.89	0.77	0.83	7032
p	0.94	0.97	0.95	23916
avg / total	0.93	0.93	0.92	30948

USING 5 CLASSES (1,2,3,4 & 5 STAR RATING)

	precision	recall	f1-score	support
1	0.71	0.80	0.75	2833
2	0.49	0.17	0.26	1650
3	0.51	0.32	0.39	2549
4	0.50	0.29	0.37	6841
5	0.74	0.93	0.83	17075
avg / total	0.65	0.69	0.65	30948

COMPARISON

Comparison of models for predicting the rating

Model	*Feature*	*Accuracy*	*Number of Classes*
Naive Bayes	Unigram + TF	57.756%	5
Naive Bayes	Unigram + TF	88.272%	2
Linear SVM	Bigram + TF-IDF	68.938%	5
Linear SVM	Bigram + TF-IDF	92.639%	2

We also tried to evaluate our model for binary sentiment analysis (positive and negative reviews)

5 classes: 1,2,3,4 and 5 rating
2 classes: 1 and 5 rating (converting 2,3 to 1 and 4 to 5)

TASK 2: RECOMMENDING RESTAURANTS TO THE USER

We are implementing the following recommendation models

- CONTENT BASED FILTERING APPROACH
- COLLABORATIVE FILTERING APPROACH

CONTENT BASED FILTERING

Content-based filtering recommends new restaurants based on the similarity of a business' characteristics to a user's profile. Each restaurant is characterized by its

- Average review score
- Category (e.g. restaurant, Chinese, Italian)
- Attributes (e.g. non-smoking, accepts credit card)

A user's profile is a weighted average of the features of the restaurants she reviewed, weighted by her rating of the restaurant. Finally, the algorithm recommends the restaurants that are closest in cosine distance to the user's profile.

COLLABORATIVE FILTERING

- We are using user-based and item-based collaborative filtering approach.
- Similarity between two users is calculated based on their ratings.
- Cosine similarity is used to calculate similarity between users.
- We calculated similarity using the following 3 ratings:
  * Biased rating from the dataset
  * Unbiased rating from the task 1 using Linear SVM
  * Unbiased rating from the task 1 using Naive Bayes

RESULTS

Model	RMSE - Testing Data	RMSE - Training data	MAE - Testing Data	MAE - Training Data	Rating used
User based collaborative filtering	3.422425	3.405979	3.101183	3.086098	Biased rating (rating from the dataset)
Item based collaborative filtering	3.424497	3.406809	3.103325	3.086804	Biased rating (rating from the dataset)
User based collaborative filtering	3.393075	3.413332	3.073444	3.093057	Unbiased rating (rating from the rating prediction task using Linear SVM)
Item based collaborative filtering	3.395080	3.414164	3.075493	3.093765	Unbiased rating (rating from the rating prediction task using Linear SVM)
User based collaborative filtering	3.572904	3.567457	3.226867	3.219094	Unbiased rating(rating from the rating prediction task using Naive Bayes)
Item based collaborative filtering	3.575060	3.568335	3.229085	3.219826	Unbiased rating(rating from the rating prediction task using Naive Bayes)

FUTURE WORK

- To use parts-of-speech in feature selection process
- To build a novel approach that combines content-based filtering with collaborative filtering to arrive at more relevant recommendations
- To build a group recommendation system that considers the information of all individuals in a group and recommend restaurants that satisfies the group of users

THANKS !!