PREDICTING STAR RATINGS FROM TEXT REVIEWS

—

BY NEELA AND SNEHA

OUTLINE

- MOTIVATION
- INTRODUCTION
- DATASET
- ASPECT BASED SENTIMENT ANALYSIS
- MODELS
- CONCLUSION

MOTIVATION

- Too many reviews to read.
- Review text might have some really helpful information which is lost.
- Our rating will be summary of the review text.
- Summary could be used for wide range of applications like
  - Predicting what features a product is good/bad at
  - What feature has improved/ deteriorated with time
  - What features should a product work on to get more 5 star ratings
  - A metric to rank products based on different aspects

INTRODUCTION

- Problem Statement: Predicting star rating from only the review text.
- Plain sentiment analysis which only involves detecting positive and negative adjectives for deciding star ratings may not always work as it does not take it into account the context of the sentiment.
  - Eg. “The pizza was great but the service was bad. Also the ambience needs some improvement. The lights were particularly unpleasant.”

DATASET

- Obtained data from yelp website which had 5 files containing business info, reviews, checking info, tip info and user info.
- Used Business info and reviews files to extract reviews only related to restaurants.
- We had total 200k records. Out of that we have used only 1k records. We have used 900 for training and 100 for testing.
- Use records voted as useful and not funny
- Review text is the only feature and star rating is the response variable.

ASPECT BASED SENTIMENT ANALYSIS

- Sentiment analysis is extracting opinions, sentiments, evaluations and emotions from text.
- In Aspect Based Sentiment Analysis the goal is to identify the aspects of given target entities and the sentiment expressed towards each aspect.
- ABSA attempts to detect the main aspects (features) e.g., ‘pizza’, ‘parking’ , of an entity e.g., ‘restaurant’ and estimate the average sentiment of the aspect.

ASPECT BASED SENTIMENT ANALYSIS

ASPECT	DEGREE OF POLARITY	
PIZZA	(PERFECT)	0.625
PARKING	(BUSY)	-0.125
WIFI	(SLOW)	-0.45
SUSHI	(GOOD)	0.875

ABSA FOR RATING PREDICTION

- We Perform ABSA for feature extraction.
- We use 5 features i.e. food, service, ambience, price and miscellaneous
- After performing ABSA we get scores for these features based on review text.
- We then apply Support Vector Regression for predicting the star rating using these features.

ABSA STEPS

- Aspect term extraction
- Aspect category detection
- Aspect sentiment analysis
- Aspect category sentiment scores

ASPECT TERM EXTRACTION

- An aspect term names a particular aspect of the target entity
- Nouns relevant to the predefined identity, i.e Restaurants are extracted
  - “ the fish was not good, but the staff greatly helped the perception”
    Aspects : fish, staff
- Used “Parts of speech” tagging to identify nouns
- The relevant nouns were extracted after identifying the category they belonged to
- Used nltk library in python

ASPECT CATEGORY DETECTION

- Detect the categories of the aspect terms from predefined categories
  - food, service, ambience, price and miscellaneous
- Used lin similarity to categorize words into the categories
- Similarity is measured with all words in every categories and the word goes to the category with word of maximum similarity
- Irrelevent aspects are not categorized

ASPECT SENTIMENT ANALYSIS

- For the aspect terms, find the degree of positivity or negativity
- Used sentiwordnet in nltk python to detect values and scaled them to range -1 to 1
  - “The manager was unprofessional” {unprofessional (manager): -0.675}
- Used a list of positive and negative words in reviews by Hu and Liu, KDD-2004 to enhance performance.
  - “top” is a neutral word in sentiwordnet, but contained in positive list

ASPECT CATEGORY SENTIMENT SCORES

- Need to associate scores with aspects
- After the aspects are categorized and scores assigned to adjectives, the next step is to detect the aspects the adjectives describe
- Regular expressions and distance metrics are not very efficient
- Used Stanford CoreNLP for better performance

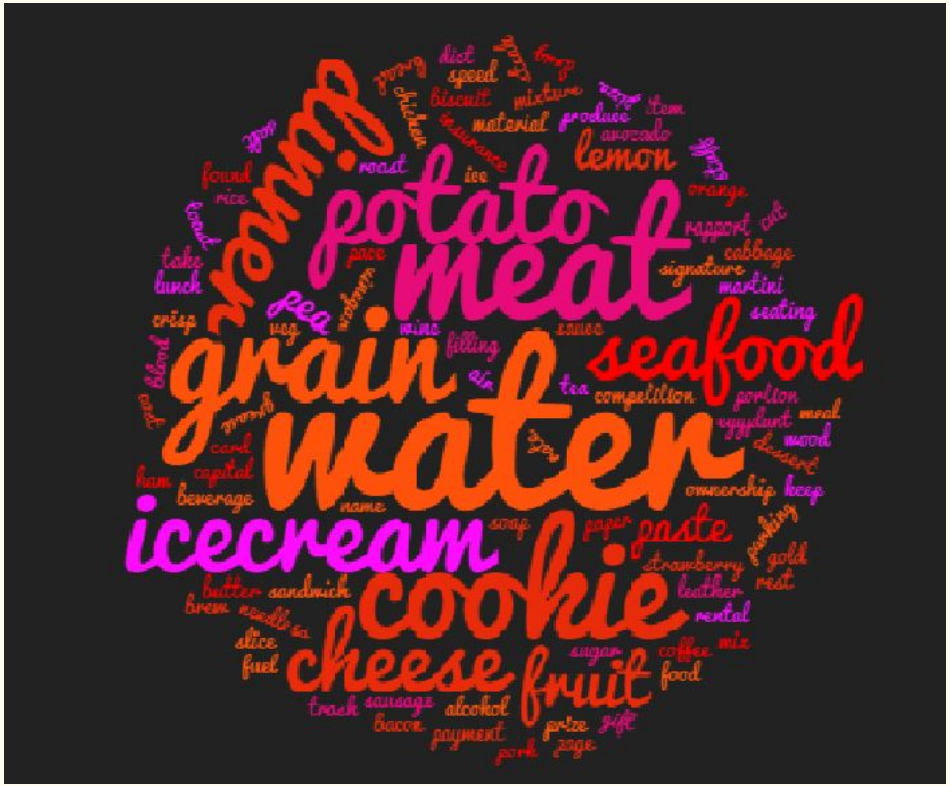
ASPECT CATEGORY SENTIMENT SCORES

- Stanford typed dependencies representation provides a simple description of grammatical relationship by extracting textual relations.
- Example : The pizza and salad were delicious

Universal dependencies, enhanced

```
det(pizza-2, The-1)
nsubj(delicious-6, pizza-2)
cc(pizza-2, and-3)
conj:and(pizza-2, salad-4)
nsubj(delicious-6, salad-4)
cop(delicious-6, were-5)
root(ROOT-0, delicious-6)
```

FEATURE: FOOD



FEATURE: AMBIENCE

SENTIMENTS: POSITIVE

SENTIMENTS: NEGATIVE

FEATURES

id	sentimen	review
1	5	My wife took me here on my birthday for breakfast and it was excellent.The weather was perfect which
2	5	I have no idea why some people give bad reviews about this place. It goes to show you you can please e

reviewnum	food	service	place	price	others	stars
1	0.6250000	0.6250000	0.0000000	0.6250	3.7500000	5
2	0.6250000	0.8333333	0.0000000	0.6250	3.7500000	5
3	0.6250000	0.8333333	0.0000000	0.6250	3.7500000	5

LINEAR REGRESSION

- Performed feature scaling. Scaled values to range -1 to 1.
- The Linear Regression finds best fitting Linear model through data points
- We got RMSE of 1.289
- The test set error was 1.41
- Thus we could see from linear regression coefficients that there is no strong linear relationship between features.

SUPPORT VECTOR REGRESSION

- We realized after performing linear regression on the data, that there is no linear relationship between the features as we got high RMSE.
- Thus we have to use a non linear model for prediction.We used support vector regressor.
- SVR with radial kernel function gave best results.

SUPPORT VECTOR REGRESSION

- Calculated the error using 10 fold cross validation.
- We got the training error to be 1.21
- The test set error was 1.16
- The difference between the training set error and test set error is very low.Thus the model is a good fit.

CONCLUSION

- We have built a model with Aspect Based sentiment Analysis and Machine Learning Model for predicting the star ratings
- We considered 2 methods Linear Regression and Support Vector Regressor.
- SVR had an error of 1.16 and Linear Regression had an error of 1.41 on test set.
- Thus SVR performs better than Linear Regression.


Thank You!