INF FAKULTÄT FÜR INFORMATIK

Yelp Dataset analysis - Gilbert

Shivam Maurya, Mukul Salhotra, Raghav Singh, Daniel Franke, Gaurav Sharma

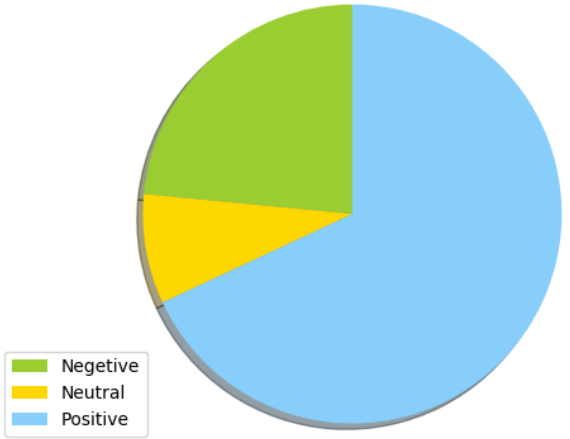
Outline

- Dataset Overview
- Pre-processing
- Features used for classifiers
- Classifiers selected
- Evaluation Criteria Selection and results
- Conclusion

Dataset Overview

- We would be analyzing for City Gilbert which has:
  - Total: 71015 reviews
  - Avg. Rating: 3.78
  - Std. Deviation: 1.5
- Approx. 2/3rd values are Positive Review, 1/4th values are negative.
- The business outlets have :
  - Avg. Rating: 3.58
  - Std. Deviation: 0.83
- Users tend to give :
  - Avg. Rating: 3.35
  - Std. Deviation: 0.92 .

Distribution of Ratings in Review dataset



Pre-processing

- Pre-processing on review text:
  - Total no of words, no. of positive words and negative words in review[1].
  - Stop word elimination (english).
  - Removal of punctuation marks (!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~)
  - Removal of digits
  - Tfidf vectorization of the review
- The target variable value range from 0-5, it converted into negative (<=2 rating), neutral (rating between 2 and 3.5) and positive (rating above 3.5) review.

[1] Opinion lexicon, University of Illinois at Chicago.

Feature Vector

- From User dataset
  - Average rating given by the user
  - No. of reviews given
- From Business dataset
  - Average rating of the business outlet
  - No. of reviews received.
- From Review dataset rating and review text. The review is further processed to fetch following features:
  - Total no of words in review, no of positive and negative words present in review.
  - Tfidf vector for each document

Classifiers Used

- Logistic Regression
- kNN with 25 neighbors
- Random forest classifier with 250 trees

Evaluating Classifiers and Results

- Data is first randomized and then cross validation with 5 folds is used to fetch the below metrics. Further the results will be averaged from the 5 fold cross validation
  - Accuracy
  - Precision
  - Recall

Classifiers	Accuracy	Precision	Recall
Logistic Regression	0.8075	0.77	0.81
kNN	0.7299	0.68	0.73
Random Forest	0.8045	0.76	0.80

Table: Average of parameters after cross validation of dataset with 5 folds.

Conclusion

- kNN can be easily extended to stream mining and other algorithms can be also extended but not so intuitively.
- Feature selection was important aspect and information about user and business helped in boosting accuracy of the learning models.
- Its important to handle the text as reviews tend to have lot of internet lingos which our. It would be interesting to remove words with low and very high frequency. This could help in dimensionality reduction and a better model.
- Handling text will be quite challenging in streaming scenario.