Project Final Presentation

...

Pradeep Ravilla
Naveen Jetty
Janak Bhalla
Jaini Vora
Sujith Shivaprakash

Task 1

...

Predict the Categories of each Business from Reviews

Tools & Technologies

- Python
- NLTK
- Sci-kit Learn
- Java
- MongoDB

Data Preprocessing

- Removed special characters, URLs,

(Several URL Links were present)

- Removed emoticons, smileys or other symbols.

(Example: :), :(, :O)

- Remove stopwords

- Used Lemmatizer to remove inflectional endings of the words in reviews.

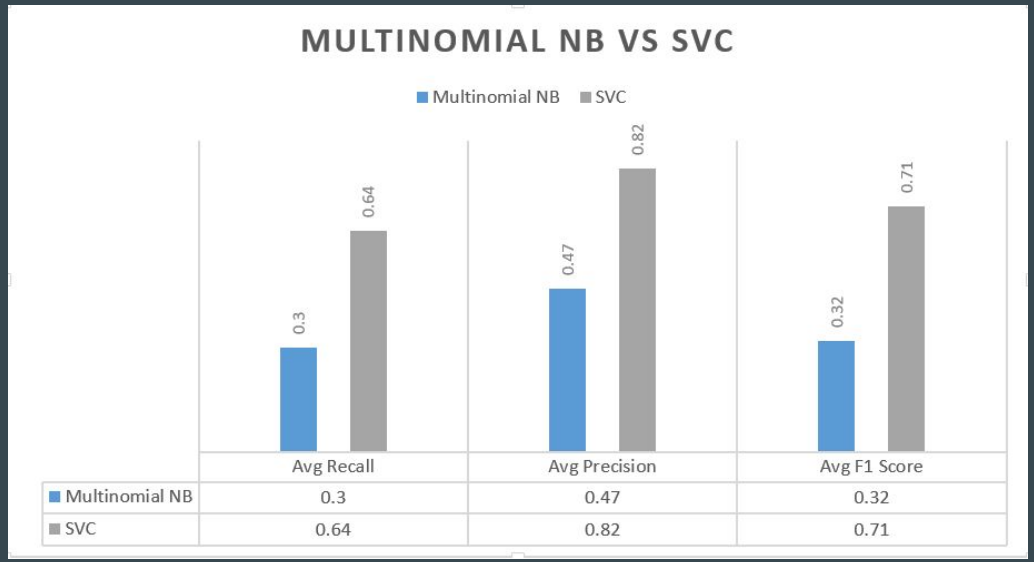
Features & Method

- Reviews and tips were vectorized using Tf-IDf scores.
- We used a filter to remove words which are infrequently used.
- We used a OneVsRest Classifier to create multi label classification results.
- Used SVC & MultiNomial NB methods to classify for each class label.

Evaluation

- Use Stratified Sampling to split data into 70:30 training & testing
- Compared the precision & recall for businesses using LinearSVC and Multinomial Naive Bayes
- Accuracy for Business Categories which are rare/unique.

Results



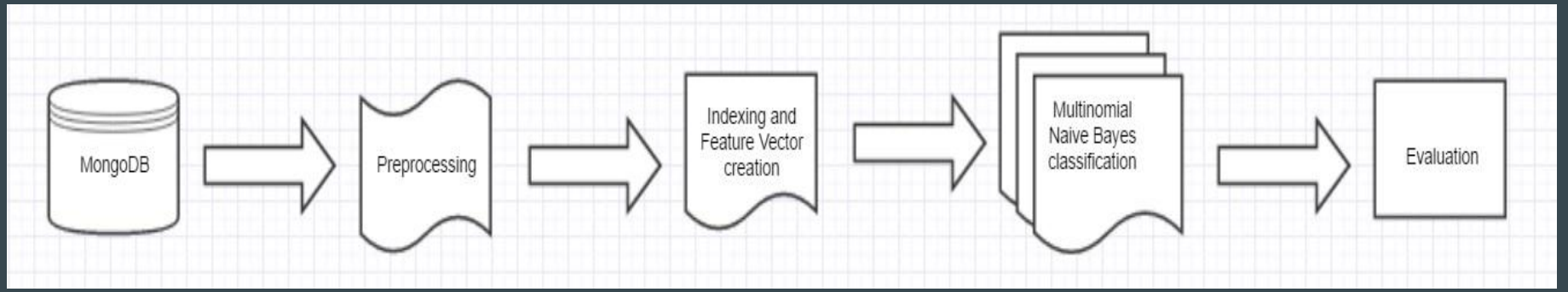
- Linear SVC performed significantly better than Multinomial NB.
- The approach resulted in having the accuracy of the business with few training examples very less.

Task 2

...

Predict Rating based on User Review

Design


MongoDB
Preprocessing
Indexing and Feature Vector creation
Multinomial Naive Bayes classification
Evaluation

DataSet

- Reviews were first stemmed and the stop-words were removed.
- The remaining reviews were indexed using Lucene where each review was considered as a document.
- The top 1000 frequent words were extracted from all reviews. This was considered as a feature set for our classifier.

Classification

- Constructed a feature vector out of each review using the bag-of-words model.
- Used multinomial logistic regression or soft-max regression(DatumBox) to classify data into different ratings

Reference

Prediction of rating based on review text of Yelp reviews - Sasank Channapragada, Ruchika Shivaswamy
[https://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/031.pdf]

Thank You!!!