Search – Yelp Dataset Challenge

Professor Xiaozhong Liu
Yang Yang, Huzefa Dargahwala, Kamal Adusumilli, Nipurn Doshi, Rama Raghava Reddy

DATA

Just kidding, you should already know what the data looks like :p

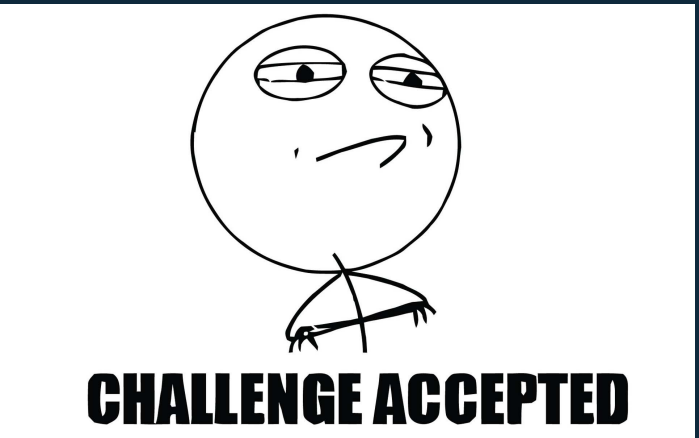
Task1

Predict the category/ies of each business from Yelp Dataset

1

Approaches

A. Machine Learning
B. Info Retrieval



A

Machine Learning Approach

1

A. Machine Learning Approach

Multilabel Classification Problem

- Inputs : All reviews and tips for all businesses combined
- Output : The categories to which the business belongs

1

A. Machine Learning Approach

Data Preprocessing

- Convert everything to lowercase
- Remove stopwords

Features

- tf-idf scores of the words present in the combined set of processed reviews and tips.
- Size of the Feature Set - **27163**

1

A. Machine Learning Approach

Algorithms

- Linear SVC
- Multinomial NB

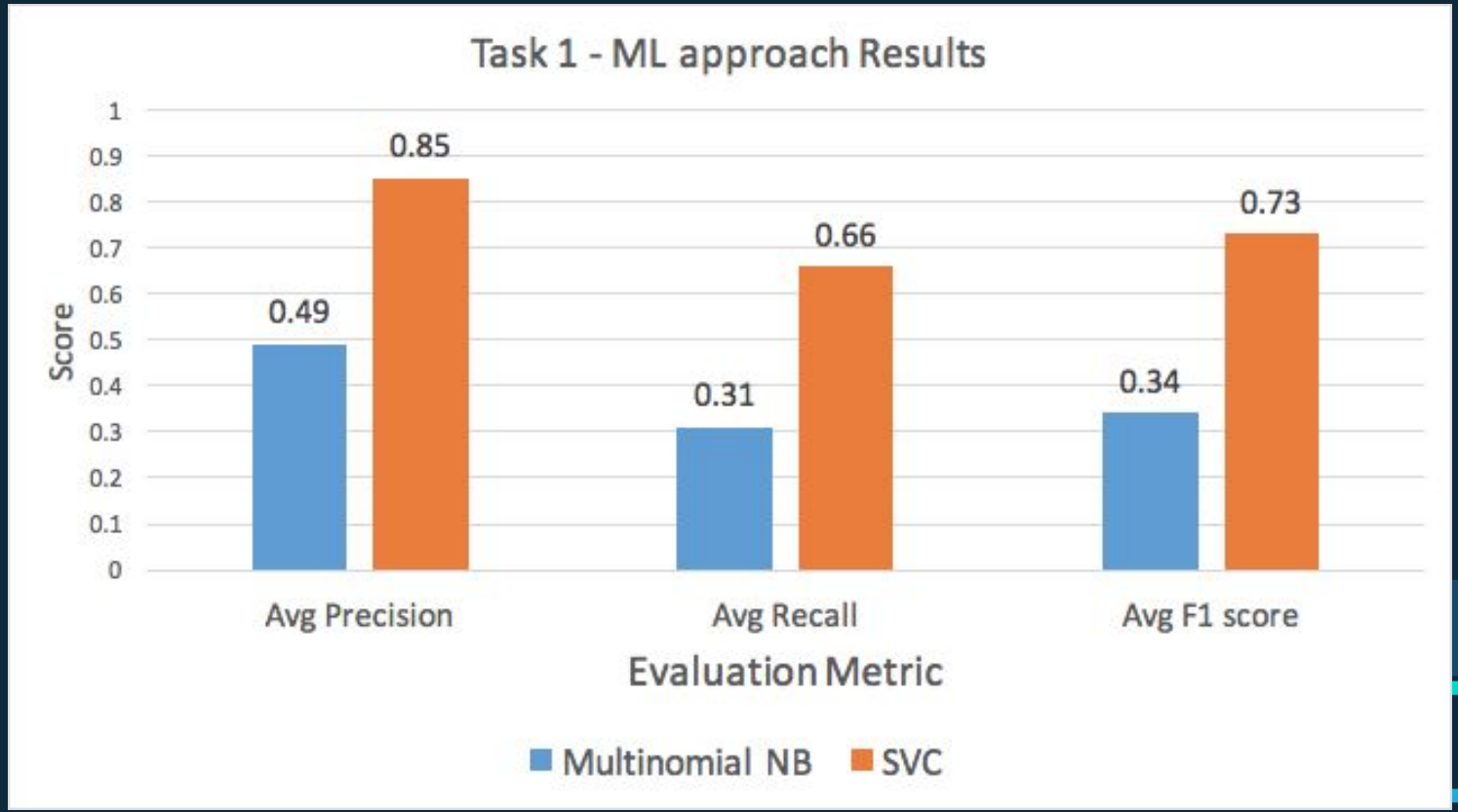
- **Entire** data Split used 80:20 and 60:40 using Stratified Sampling, both gave similar results

- One vs Rest Classification approach to handle multiple labels.

1

Results

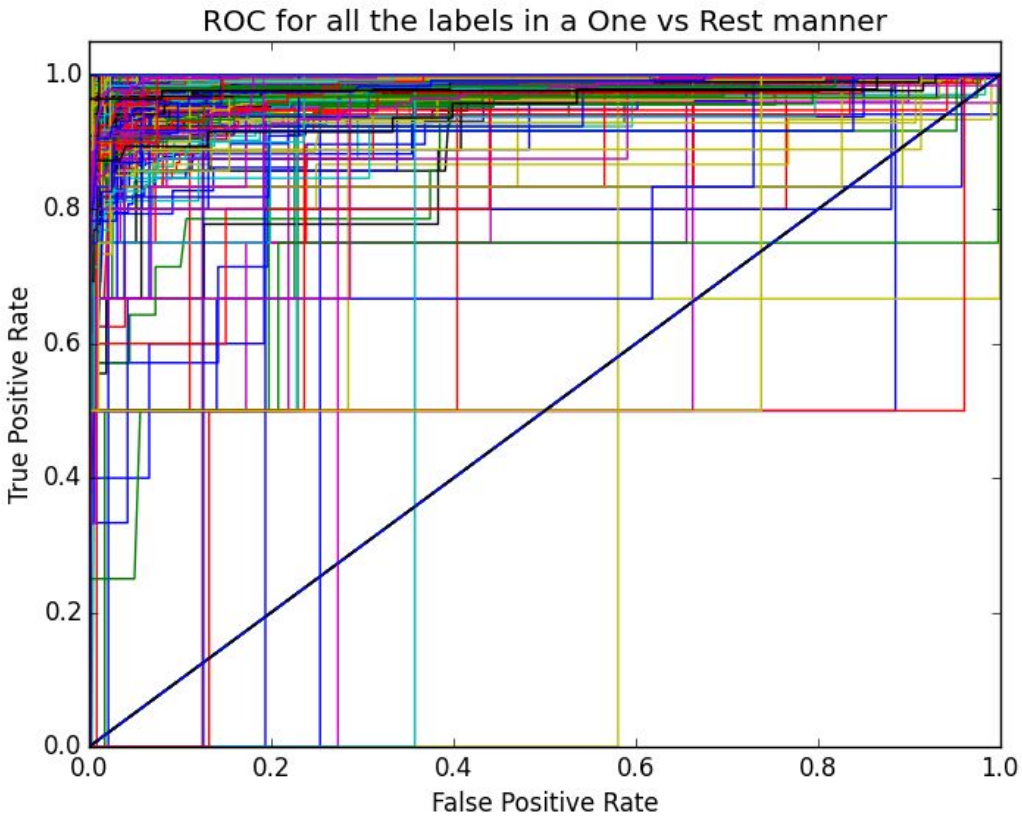
A. Machine Learning Approach



1

Results

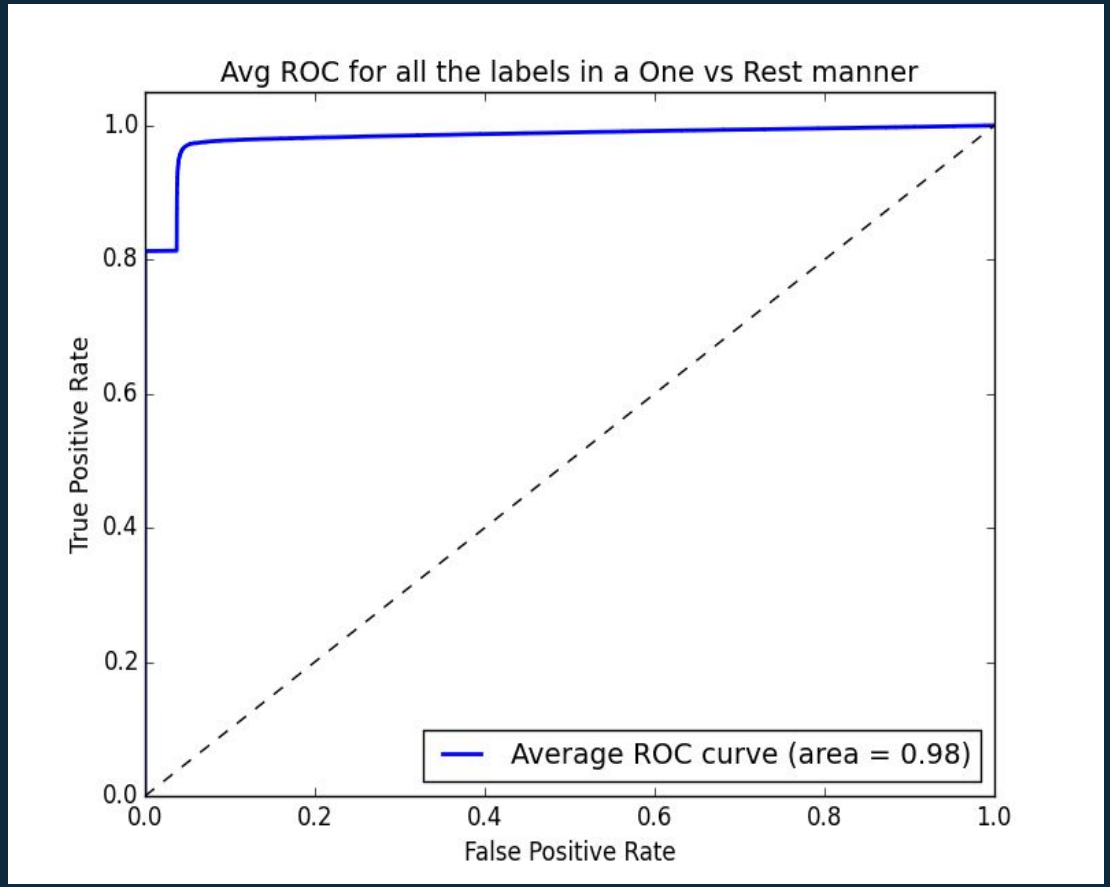
A. Machine Learning Approach

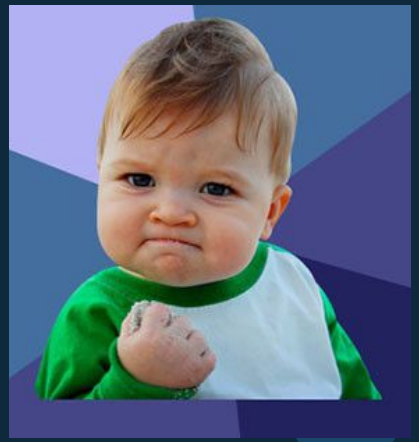


1

Results

A. Machine Learning Approach





B

Information Retrieval Approach

1

B. Information Retrieval Approach

Data Preprocessing

1. Build indexes for business, review and tip collections to speed up the search process.

2. Build the (MongoDB)collection “test_set”, which includes the business information, all the reviews and tips for each business_id.

1

Ground Truth File

B. Information Retrieval Approach

3. For each business_id, it already has some categories assigned to it. We will use this as ground truth file.

eg. if business_id1 has category [category2]

Business_id1	Category1	0
Business_id1	Category2	1

1

B. Information Retrieval Approach

4. For each query, search the fields “review” and “tip”, and find the top 10 related business_id.

Algorithms:

Vector Space Model, BM25, Language Model (Dirichlet Smoothing), Language Model (Jelinek Mercer Smoothing)

review

	BM25	LM(D)	LM(J)	VSM
P	0.4361	**0.4396**	0.3637	0.3637
R	0.1281	0.1230	0.1133	0.1133
F1	0.1244	0.1228	0.1071	0.1071

tip

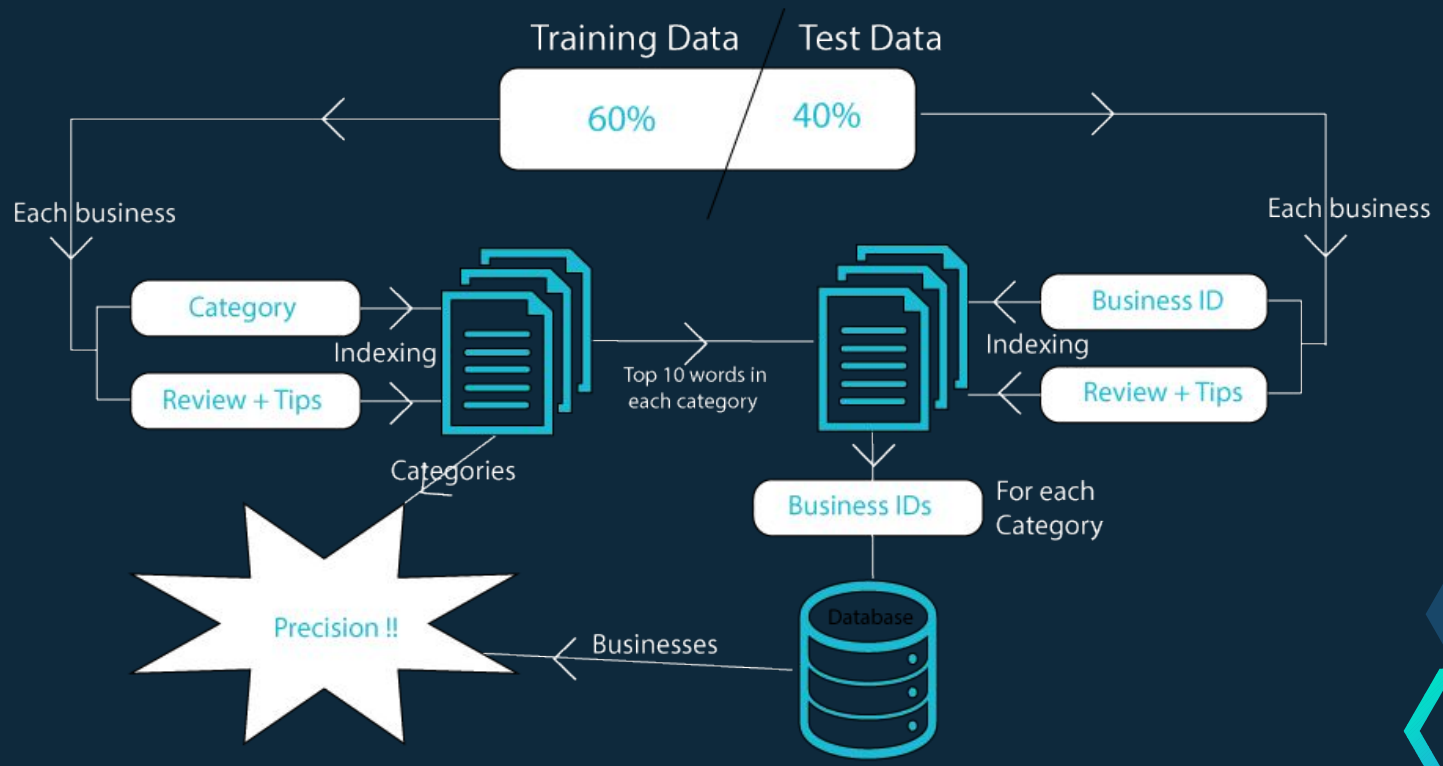
	BM25	LM(D)	LM(J)	VSM
P	0.2367	0.2575	0.2192	0.2192
R	0.3067	0.0366	0.0343	0.0343
F1	0.0462	0.0472	0.0430	0.0430

LM(D) Language Model with Dirichlet Smoothing
LM(J) Language Model with Jelinek Mercer Smoothing

1

Approach 2

B. Information Retrieval Approach



P = 0.42

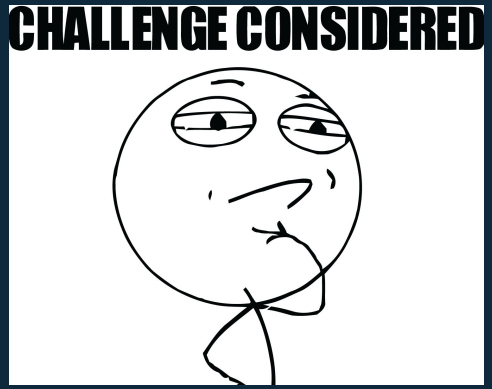
R = 0.24

F1 = 0.28

Task 2

A. Predict Rating of a Review from its text

B. Predict Helpfulness of Review from its text



2

A. Predict Rating of a Review from its text

Problem Statement

Given a review text, predict the rating on a scale of 1 to 5

Data

All Reviews from the dataset, irrespective of the business

Data Preprocessing

Steps done in task 1, Lemmatization using WordNet, Append not_ to negated words eg: not good will be not_good, and Remove Punctuation

Features

Same as Task 1 (836818 number of unigram features)

2

Algorithmic Approaches

A. Predict Rating of a Review from its text

Pure Classification Task

- Linear SVC
- Multinomial NB

with Recursive Feature Elimination

Classification & Regression Task

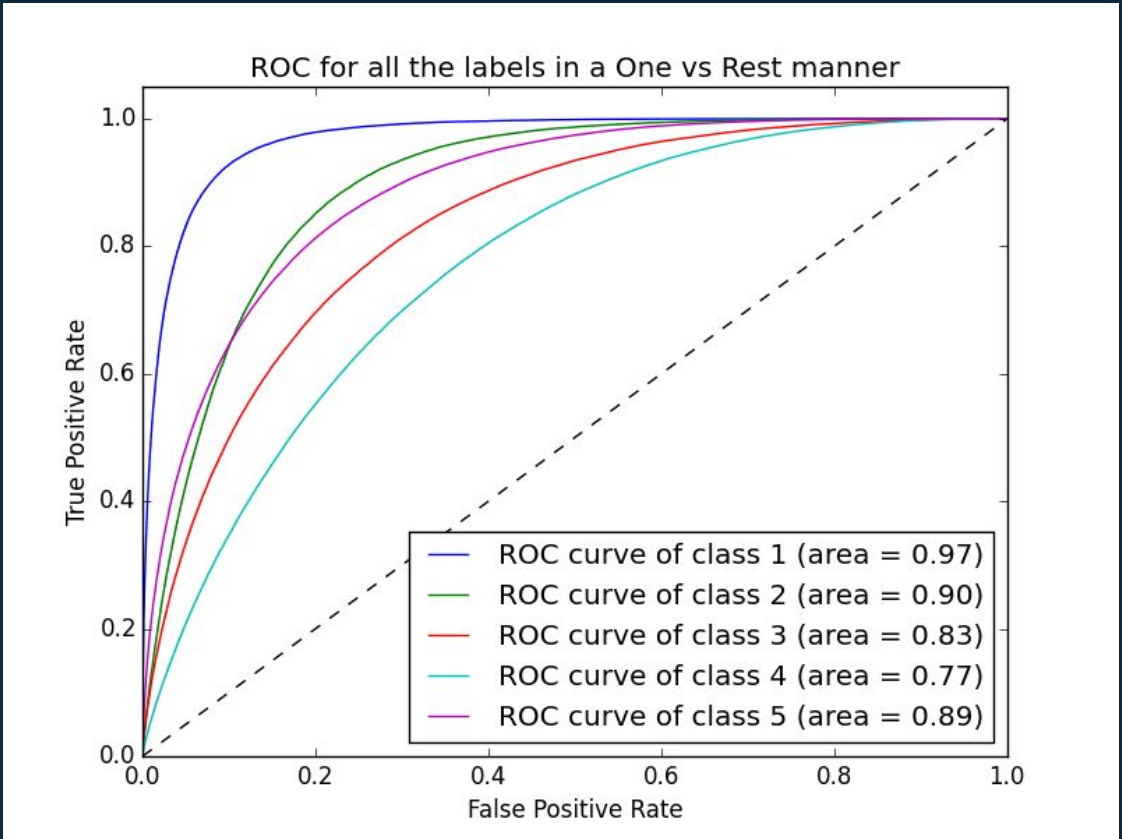
- Classification & Regression Trees (CART)

(Data Split 60:40 using Stratified Sampling)

2

Detailed Results for Classification task

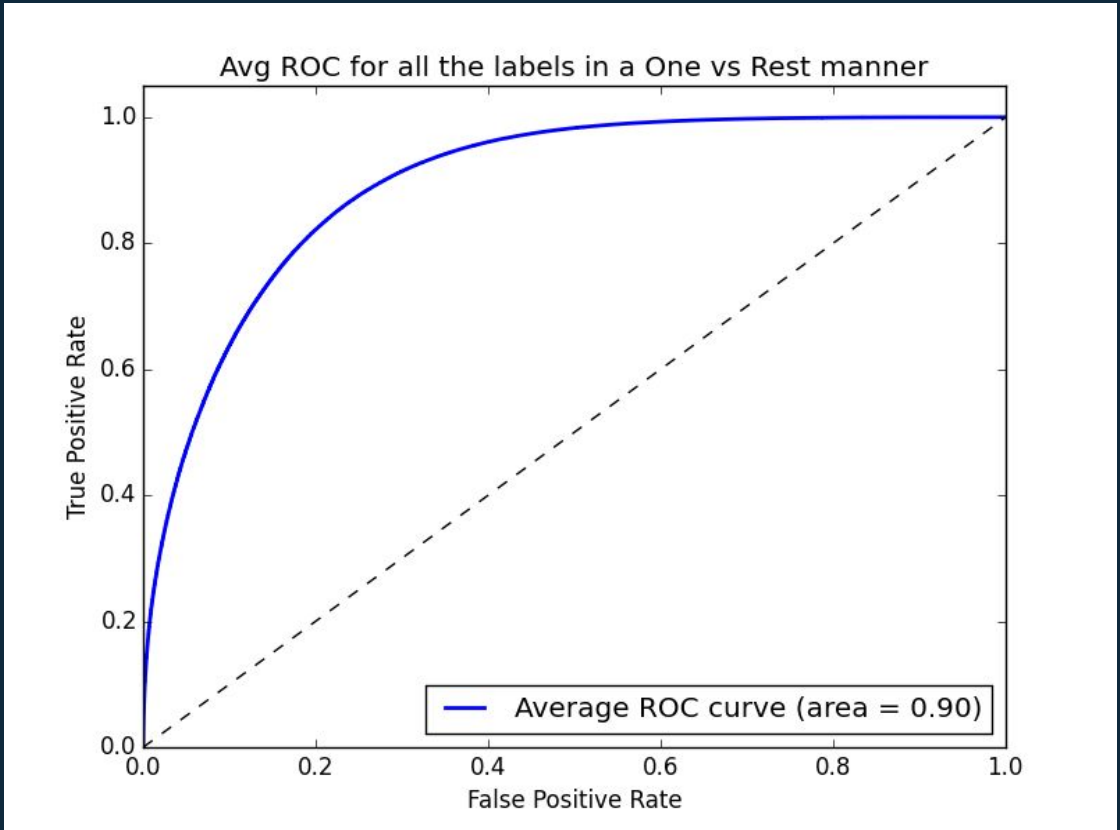
A. Predict Rating of a Review from its text



2

Detailed Results for Classification task (contd.)

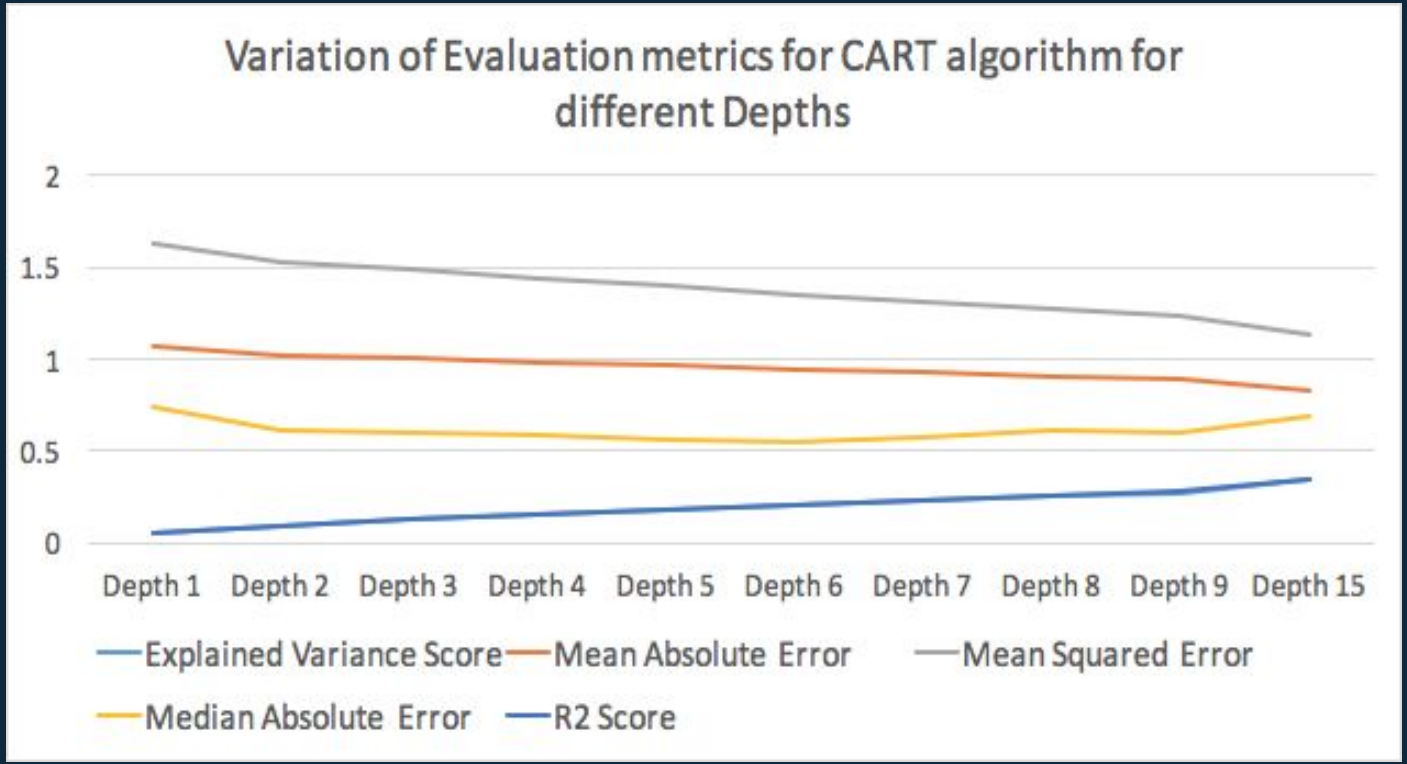
A. Predict Rating of a Review from its text



2

CART Results

A. Predict Rating of a Review from its text



2

CART Results

A. Predict Rating of a Review from its text

Explained Variance Score	0.343
Mean Absolute Error	0.832
Mean Squared Error	1.13
Median Absolute Error	0.6957
R2 Score	0.343

2

B. Predict Helpfulness of a review from its text

Problem Statement

Given a review text, predict whether the review will be helpful or not

Data

All Reviews from the dataset, irrespective of the business

Data Preprocessing

Same as Task 2a, extract helpfulness and aggregate all helpfulness > 1 to helpful = 1

Features

Same as Task 1

2

Algorithmic Approaches

B. Predict Helpfulness of a review from its text

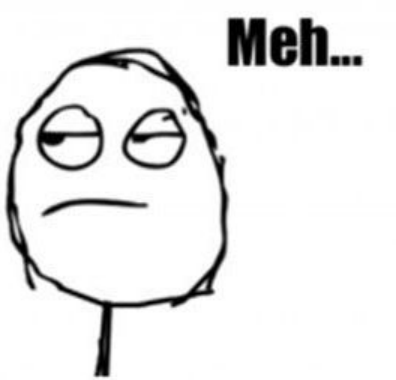
Linear SGD

Multinomial NB

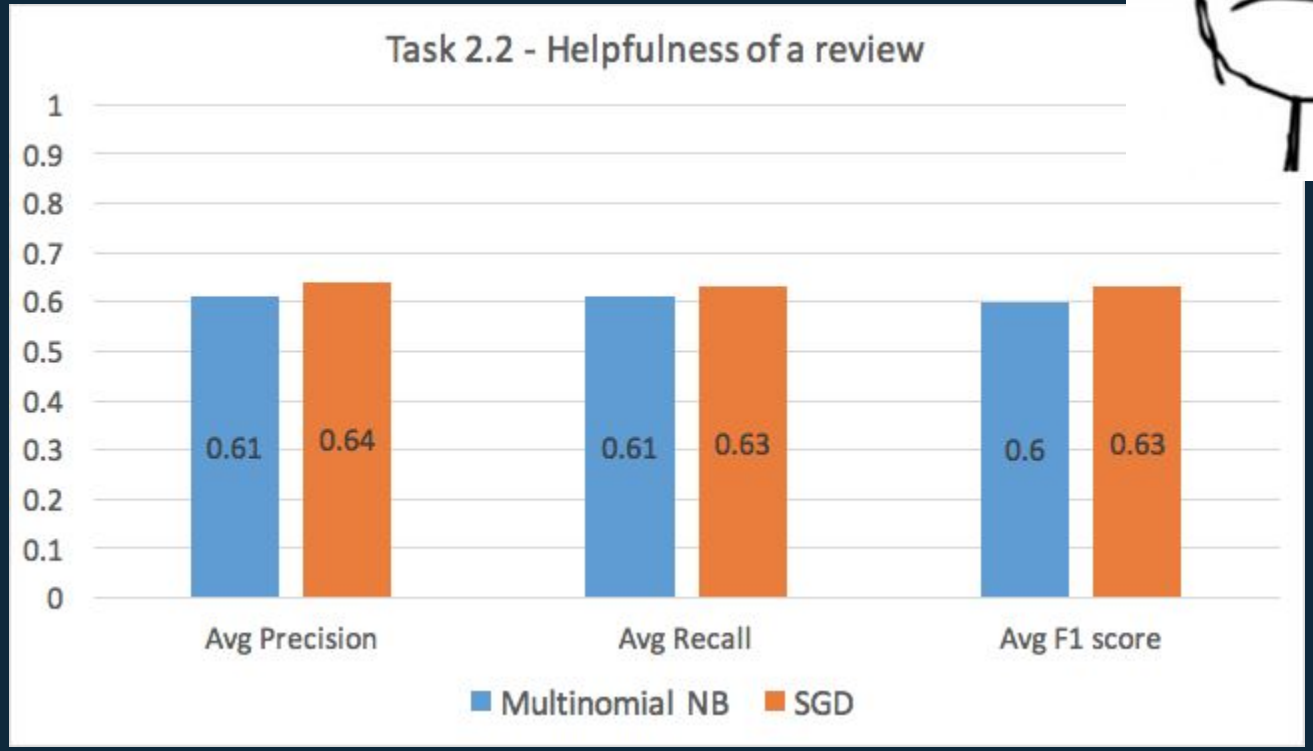
(Data Split 80:20 and 60:40)

2

Results



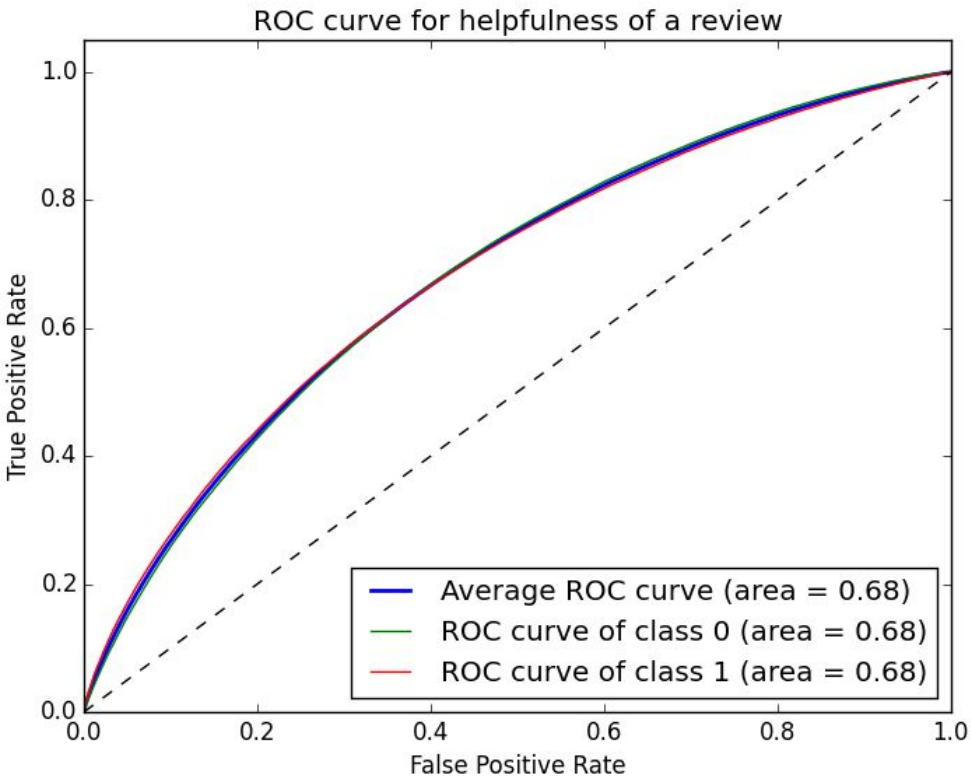
B. Predict Helpfulness of a review from its text

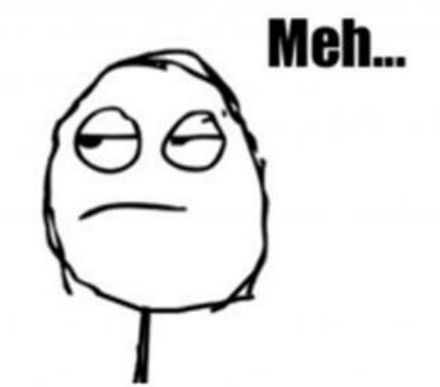


2

Results

B. Predict Helpfulness of a review from its text





Insights for Task 1

1. Class imbalance problem
2. Bag of Words Model
3. Simple linear classifier
4. Precision values - 0?

Insights for Task 2

1. Classification / Regression problem?
2. Class imbalance good or bad?
3. CART = Intractable for depth > 15

Overall experience of the project,


THANK YOU!
ANY QUESTIONS?
memegenerator.net