Data Mining Capstone Task 6

April 18, 2020

Asad Bin Imtiaz CS598
Data Mining Capstone
MCS-DS UIUC
Email: aimtiaz2@illinois.edu

1 Introduction:

In this task, reviews from Yelp Reviews from Yelp Academic Challenge Dataset were analyzed to predict if the restaurant had failed or passed its health inspection. The analysis and findings are part of Task-6 for Data Mining Capstone course, wherein the Yelp restaurant reviews were explored to predict hygiene of the restaurant and their inspection status.

In this task, important features from reviews texts for a specific business and their additional facility related properties were mined and classified to predict whether these restaurants will pass the public health inspection tests. The goal of this task was to come up with an appropriate approach to identify important features for classification which can be used for this prediction.

In the following, the implementation of these tasks and their sub tasks are explained.

1.1 Tools and Libraries:

The exploration and analysis of Yelp restaurant dataset was performed with **Python 3** using Jupyter notebook. Several python libraries such as **Gensim**, **scipy, numpy** and **Scikit-learn** were used to analyze and cleanse the review texts. For visualization, **python Matplotlib** was used to generate bar and line charts.

1.2 Dataset:

For this task, the dataset for 13299 restaurants with their reviews and some additional properties was already provided. Among these, inspection labels for 546 restaurants were provided for training of classification models. Below is a description provided files:

- **hygiene.dat**: Each line contains the concatenated text reviews of one restaurant.
- **hygiene.dat.labels**: Each line contains binary labels (0 or 1) for the first 546 restaurants where a 0 indicates that it has passed for latest public health inspection test, while a 1 indicates that it failed the test.
- **hygiene.dat.additional**: Each line in this file is a comma-separated list containing the cuisines offered, zip code, number of reviews and average rating.

2 Approach

The processing steps of this task was divided into several sub-tasks as below:

1. Data pre-processing and cleansing
2. Text based prediction

3. Exploring additional features for prediction
4. Analyzing additional features collected
5. Prediction based on additional features
6. Results analysis and interpretation

These subs tasks are discussed as separate chapters in the following.

2.1 Data Preprocessing and cleansing

As always, the first step when processing text data is to clean it and remove unnecessary elements such as stop-words from the data. The text data was taken and following cleansing was performed:

- URLs from the data were removed
- Extra whitespaces and newlines were removed
- Word abbreviations were expanded and punctuations removed
- Stop-words were removed

Additionally, from text data, bi-grams and tri-grams were extracted. This included:

- All tokens were converted into base lemma form.
- Merging frequently occurring words into bigrams and trigrams
- Merging Noun-Phrases by creating bi-grams by merging adjectives and nouns, or negations (such as not, barely etc.) and adjectives or adverbs. This gave phrases such as not-good, great-place etc. Please note that not is a stop-word, but is retained in phrases e.g. in not-bad.
- Only words representing Proper-Nouns, Nouns, Verbs, Adjectives, Adverbs, Determiners, interjections, Conjunctions and ad-positions were kept.
- All stop-words were removed.

After application of above pre-processing steps, processed review text for each restaurant was retrieved and was used in subsequent steps. The processed text was divided into training and testing splits based on the restaurants for which labels were already provided for training.

2.2 Text based prediction

From the raw text data, the classifiers were built to predict inspection labels. The review text for the restaurants for which inspection labels were provided were used to train the classifiers. Please note that these classifiers were only using text features from the raw text and no additional features were used. Multiple classifiers were built and their accuracy was assessed for the training data. In particular, the following classifiers were trained and assessed:

1. Multinomial Naive Bayes Classifier
2. Random Forest Classifier
3. Decision Tree Classifier
4. Support Vector Machine Classifier
5. K-nearest Neighbors Classifier
6. Multilayer Perceptron Classifier

All these classifiers were built from python **sklearn** library using default settings. The classification reports were generated for each of the classifiers using a train-test split of 80% to 20%. Below are the classification reports for each of the classifiers trained

Random Forest:
Classification Report:

Label	Precision	Recall	F1
0	0.76	0.56	0.65
1	0.63	0.81	0.71
Accuracy			0.68
macro avg	0.70	0.69	0.68
weighted avg	0.70	0.69	0.68

Score: 68.18

Multinomial Naive Bayes:
Classification Report:

Label	Precision	Recall	F1
0	0.60	0.58	0.59
1	0.56	0.58	0.57
Accuracy			0.58
macro avg	0.58	0.58	0.58
weighted avg	0.58	0.58	0.58

Score: 58.18

Decision Tree:
Classification Report:

Label	Precision	Recall	F1
0	0.55	0.67	0.60
1	0.54	0.42	0.47
Accuracy			0.55
macro avg	0.54	0.54	0.54
weighted avg	0.54	0.55	0.54

Score: 54.55

Support Vector Machine:
Classification Report:

Label	Precision	Recall	F1
0	0.57	0.74	0.64
1	0.58	0.40	0.47
Accuracy			0.57
macro avg	0.58	0.57	0.56
weighted avg	0.57	0.58	0.56

Score: 57.27

K-Nearest Neighbors:
Classification Report:

Label	Precision	Recall	F1
0	0.57	0.75	0.65
1	0.60	0.40	0.48
Accuracy			0.58
macro avg	0.59	0.58	0.56
weighted avg	0.59	0.58	0.57

Score: 58.18

MLP Classifier:
Classification Report:

Label	Precision	Recall	F1
0	0.67	0.82	0.74
1	0.75	0.57	0.65
Accuracy			0.70
macro avg	0.71	0.70	0.69
weighted avg	0.71	0.70	0.79

Score: 70.55

As can be see, the Multi-Layer perceptron Classifier gave the best overall score of 70 on the training data. Therefore, this classifier was used on the processed test review text to predict labels for inspections. The result gave an overall score of 68% for F1-Score on the leaderboard.

Overall Score		hygiene (1.0)			
Current	**Previous**	**Precision (0.6)**	**Previous**	**Recall (0.4)**	**Previous**
0.6868	*-inf*	0.701	*-inf*	0.6656	*-inf*

2.3 Exploring additional features

In this subtask additional features were analyzed to increase the overall accuracy of the classification. Four different types of features were analyzed. These include:

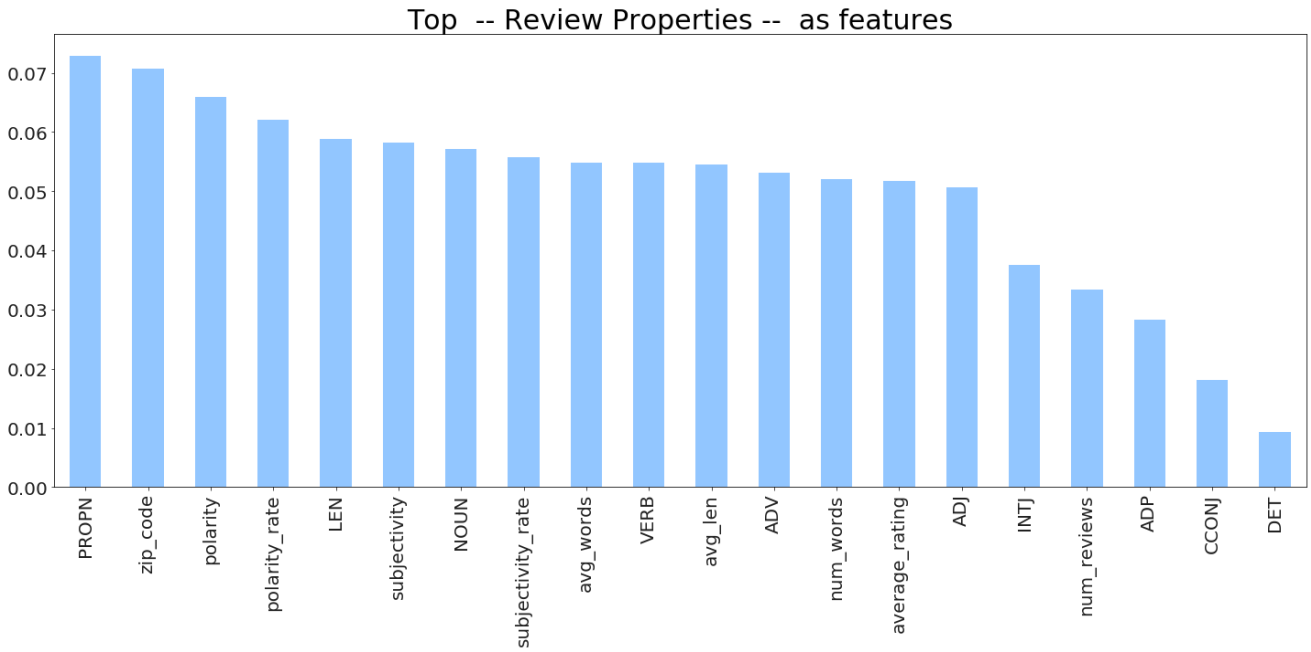
1. Features from Review Data and additional Restaurant features
2. Features from Restaurant Category
3. Features from Top Words
4. Features from Top Phrases

2.3.1 Review text and Restaurant features

These features included:

- Number of POS tags in the review for POS tags of types ['PROPN', 'NOUN', 'ADJ', 'VERB', 'ADV', 'DET', 'ADP', 'CCONJ', 'INTJ']
- Length of the review text
- Polarity and Subjectivity of the review text (calculated from python **textblob** library) and their respective rates.
- Number of Words in the text
- Number of Characters in the text
- Additional features such as Zip code, Number of reviews & Average rating

With a random forest classifier on these features and the training labels, most important features were identified as shown in the figure below:

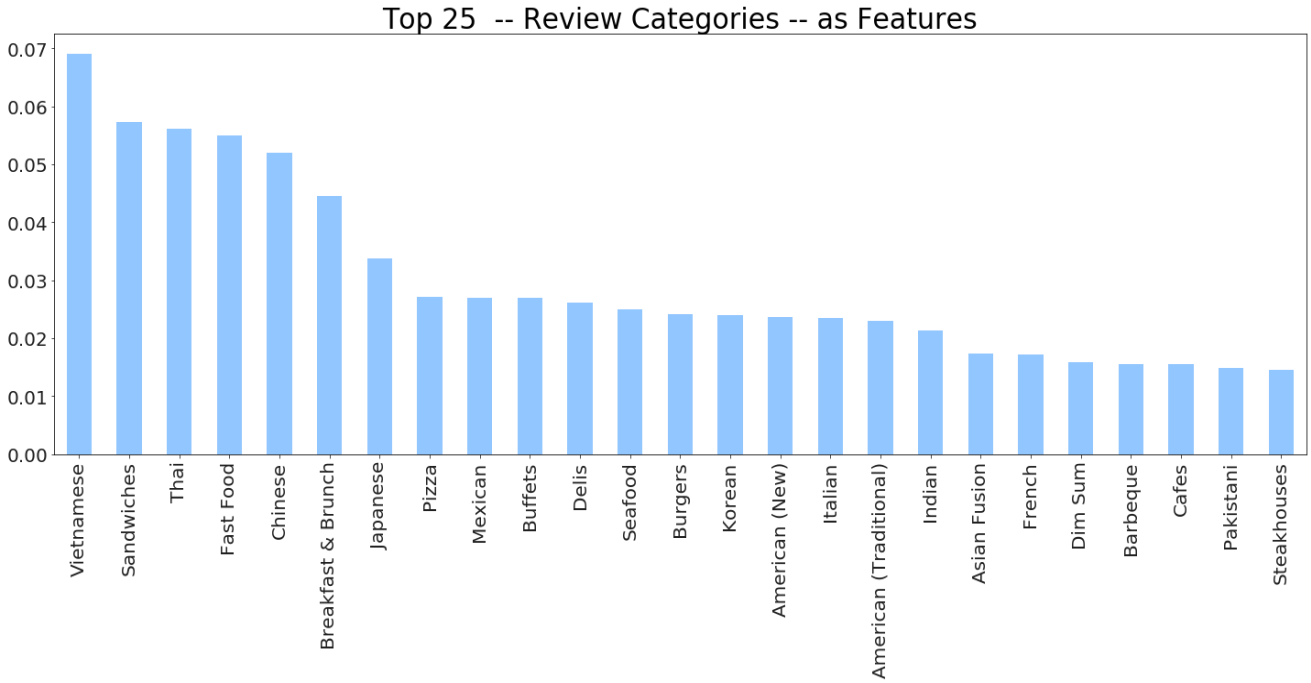


As can be seen, features such as Zip-Code and Proper Nouns (most probably restaurant names) were among the most distinguishing features for Inspection label prediction. This suggests that there is a correlation between geographical locations and failed inspections.

2.3.2 Features from Restaurant Category

These features included the food cuisine and category the restaurant is offering. A data-frame was created for each category, and for each restaurant, 1s were added where restaurant offered this category and 0s otherwise.

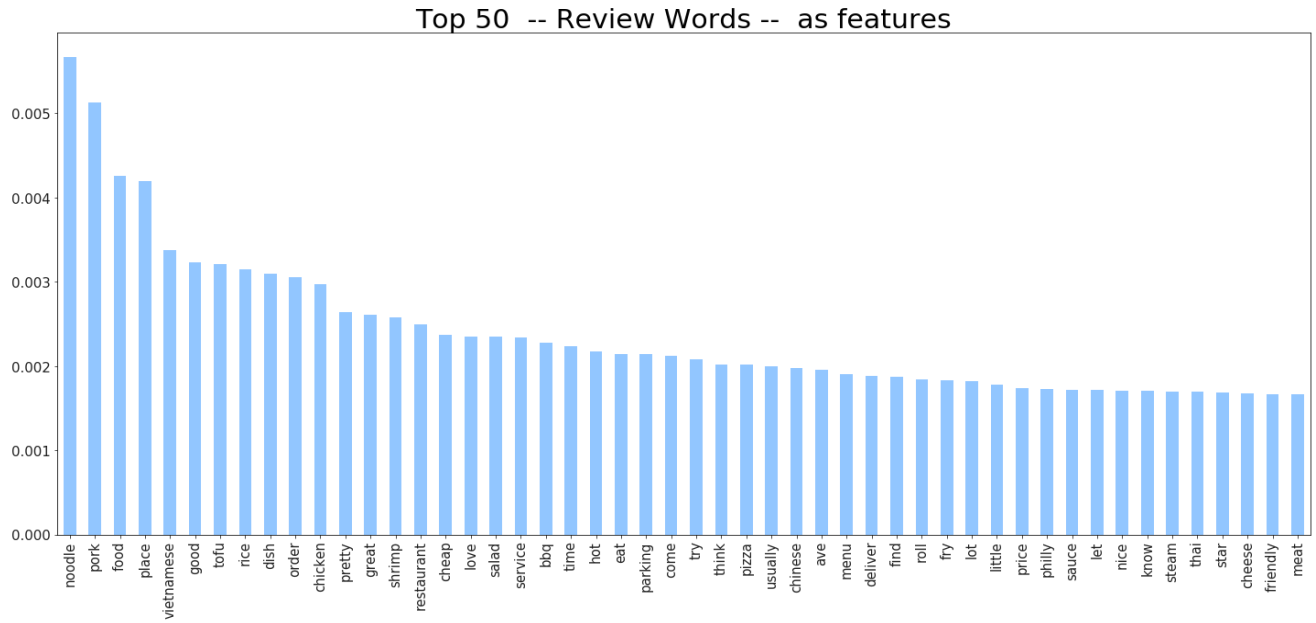
With a random forest classifier on these features and the training labels, most important features were identified as shown in the figure below:



As can be seen, mostly Asian restaurants such as Vietnamese, Thai, Chinese, Japanese etc appear in top words. In these restaurants, fresh meat products are served making them more susceptible to hygiene inspection failure.

2.3.3 Features from top most used Word/Unigrams

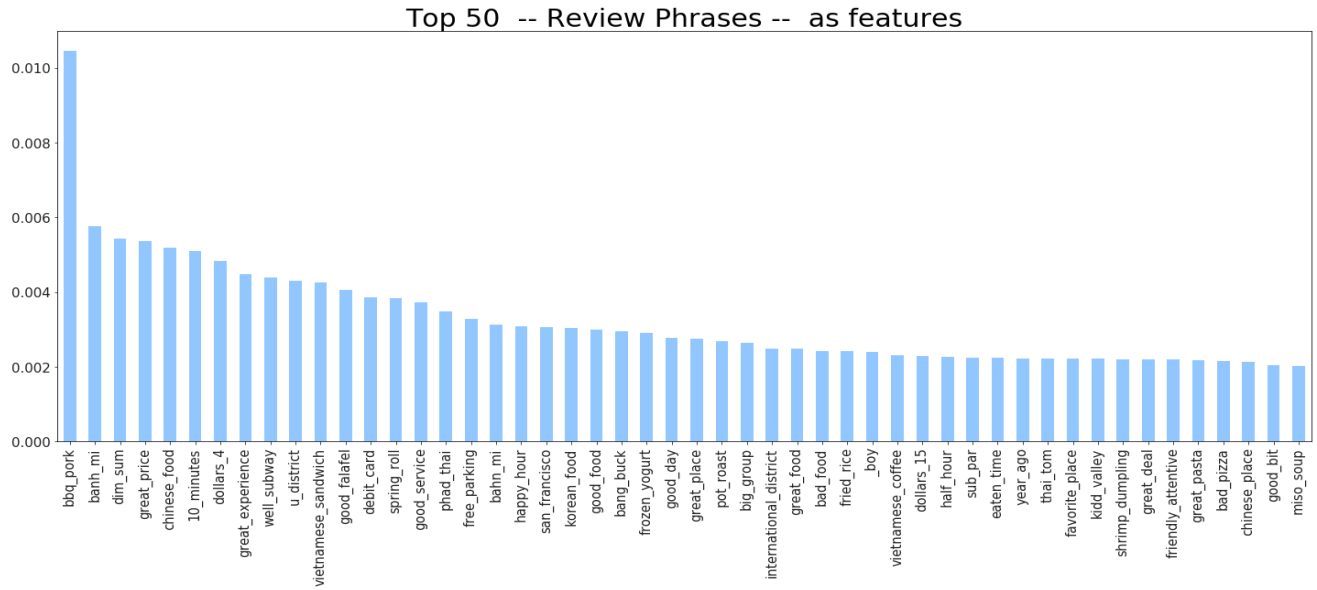
These features included most discriminating unigrams. These are words most commonly associated with a specific inspection label. Top 50 of such words are shown in figure below:



One can clearly see words such as noodle, Vietnamese, pork etc. in the top words, suggesting meat serving Asian Restaurants. This again corresponds to top food offered or related words for the categories identified.

2.3.4 Features from top most used Phrases

These features included most discriminating phrases (bi-grams, tri-grams). Top 50 of such phrases are shown in figure below:



Again, mostly phrases related to Asian cuisine related or to meat related words came on the top.

2.4 Analyzing additional features

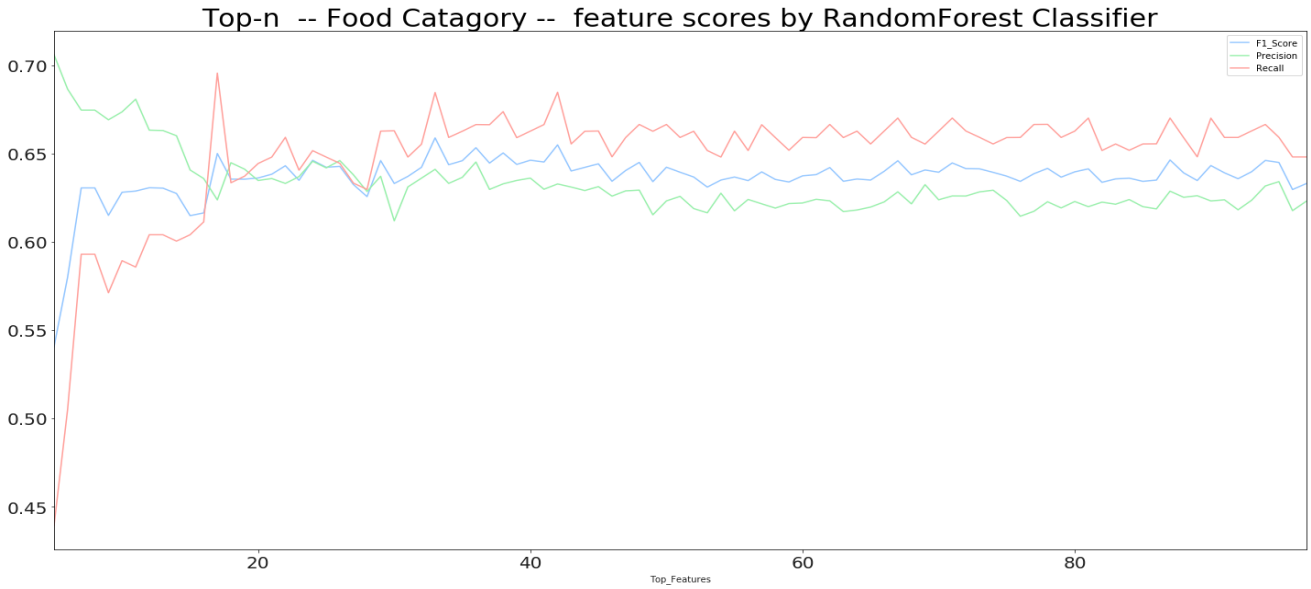
In this sub task, the four different set of additional features were analyzed for prediction accuracy of following two classifiers with cross-validation factor=5:

- Random Forest Classifier
- Multilayer Perceptron Classifier

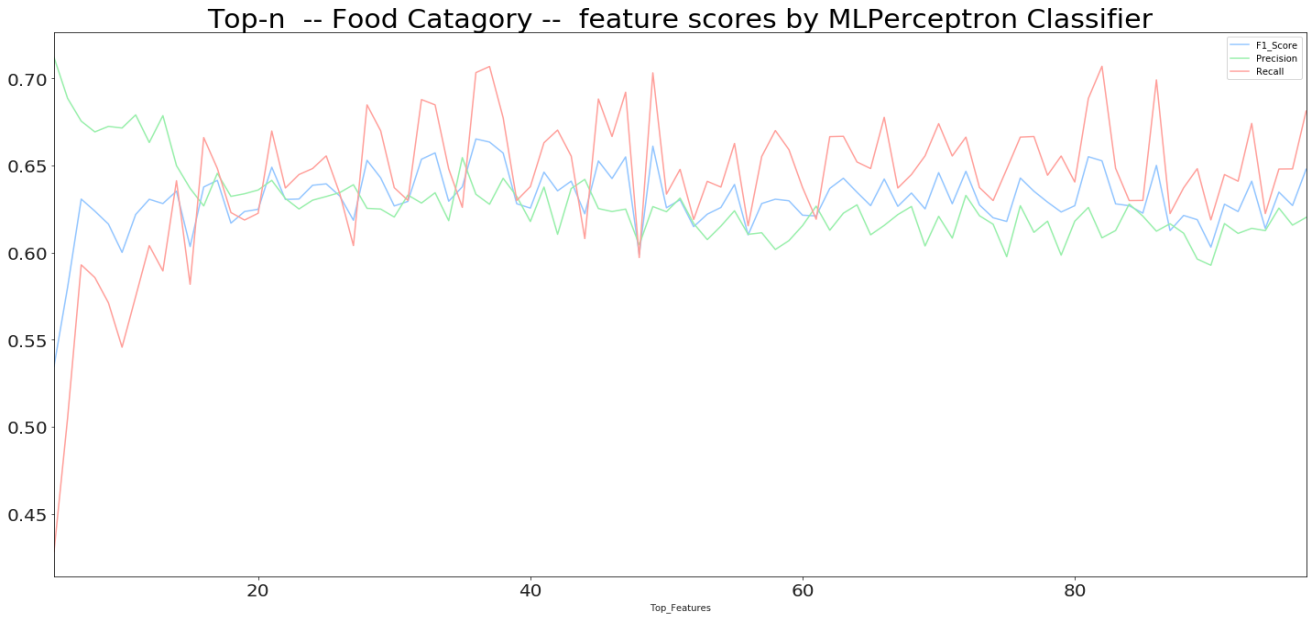
The top-N most discriminating 'Restaurant category' feature combinations were identified by calculating F1, Precision and Recall scores:

2.4.1 Restaurant Category features

For **Random Forest Classifier**, it gave top 33 features as most discriminating with highest average F1 scores:

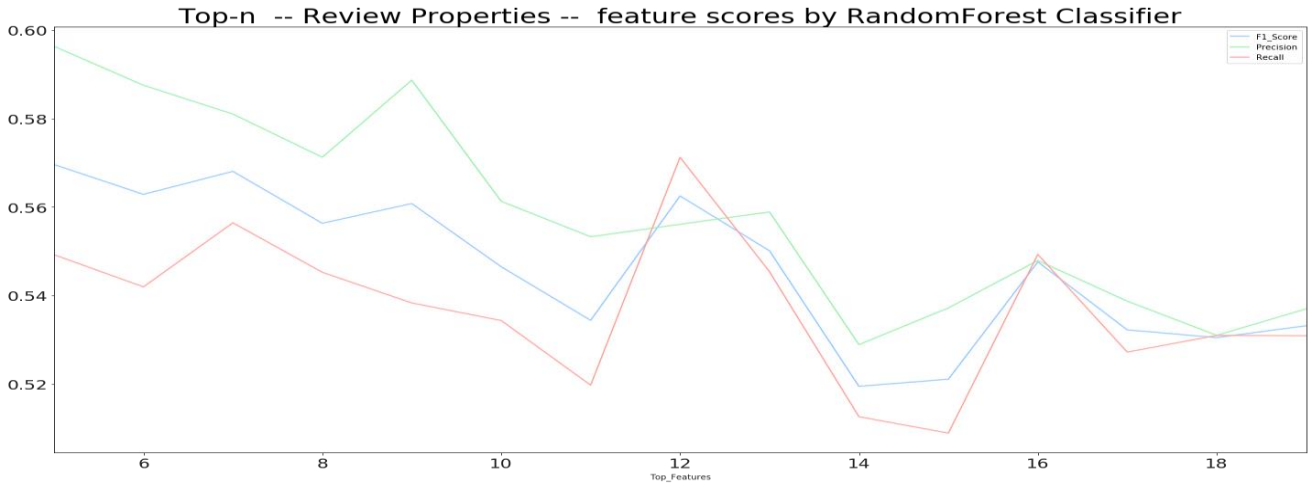


For **MLPClassifier**, it gave top 37 features as most discriminating with highest average F1 scores:

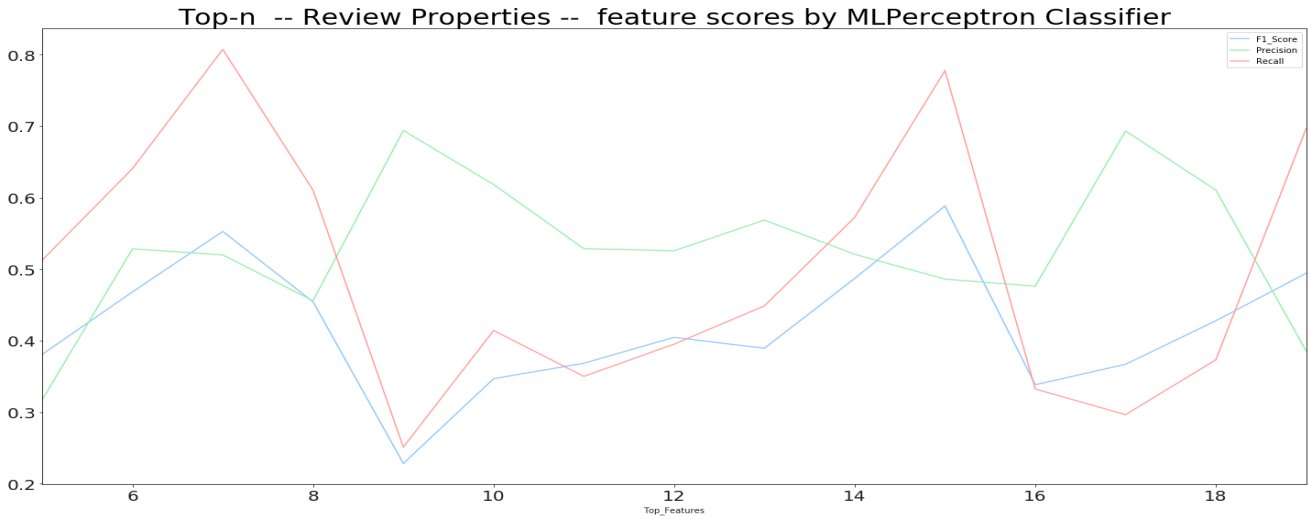


2.4.2 Restaurant Additional features

For **Random Forest Classifier**, it gave top 5 features as most discriminating:

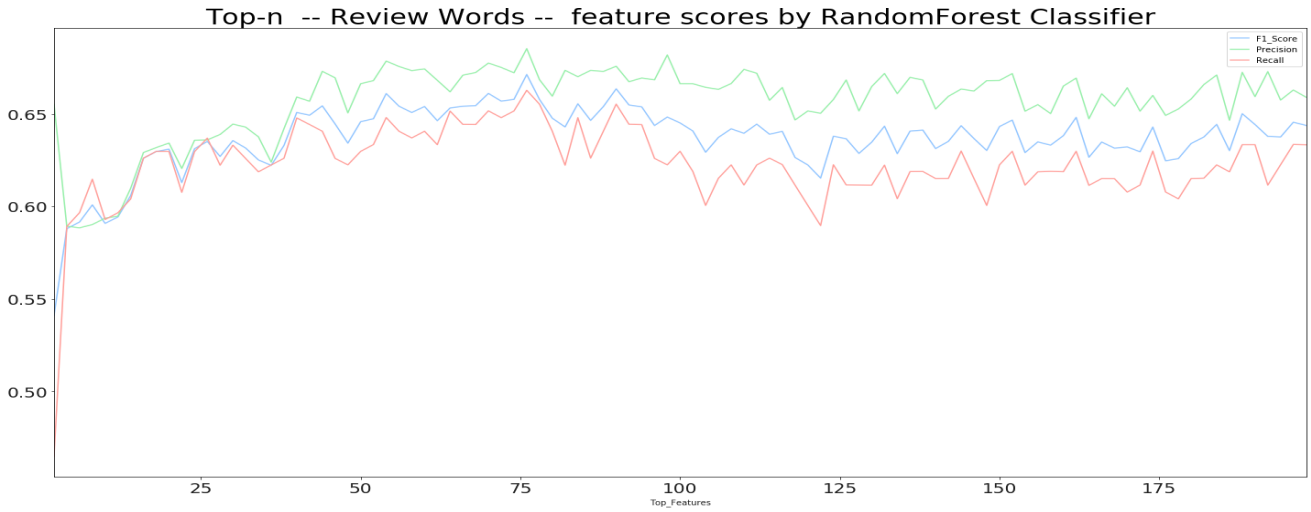


For **MLPClassifier**, it gave top 15 features as most discriminating with highest average F1 scores:

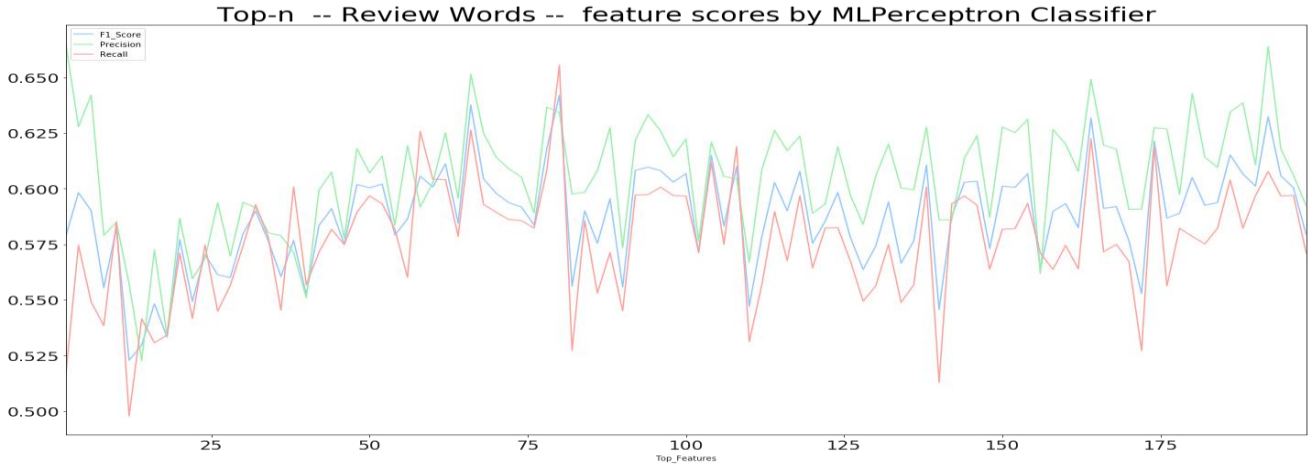


2.4.3 Restaurant Top Token features

For **Random Forest Classifier**, it gave top 76 words as most discriminating features:

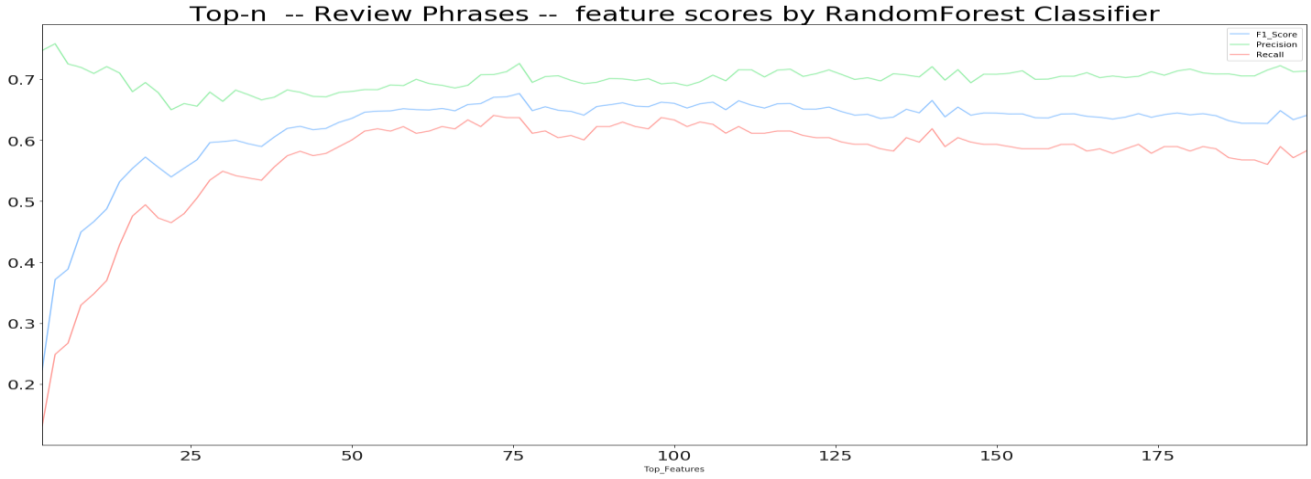


For **MLPClassifier**, it gave top 80 words as most discriminating features with highest average F1 scores:

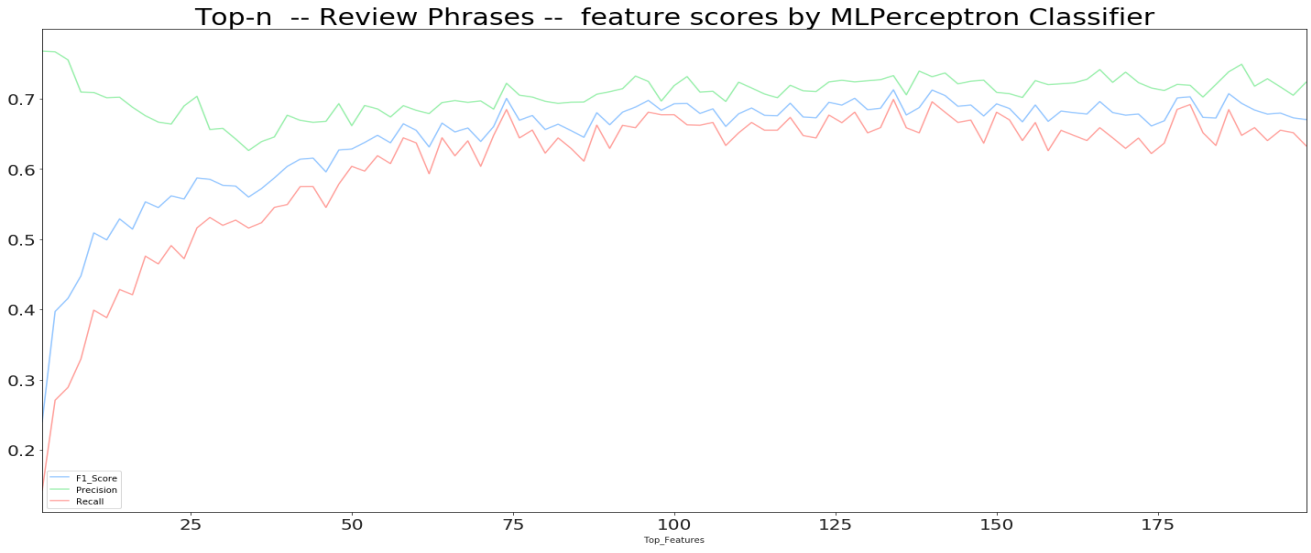


2.4.4 Restaurant Top Pheas features

For **Random Forest Classifier**, it gave top 76 phrases features as most discriminating:

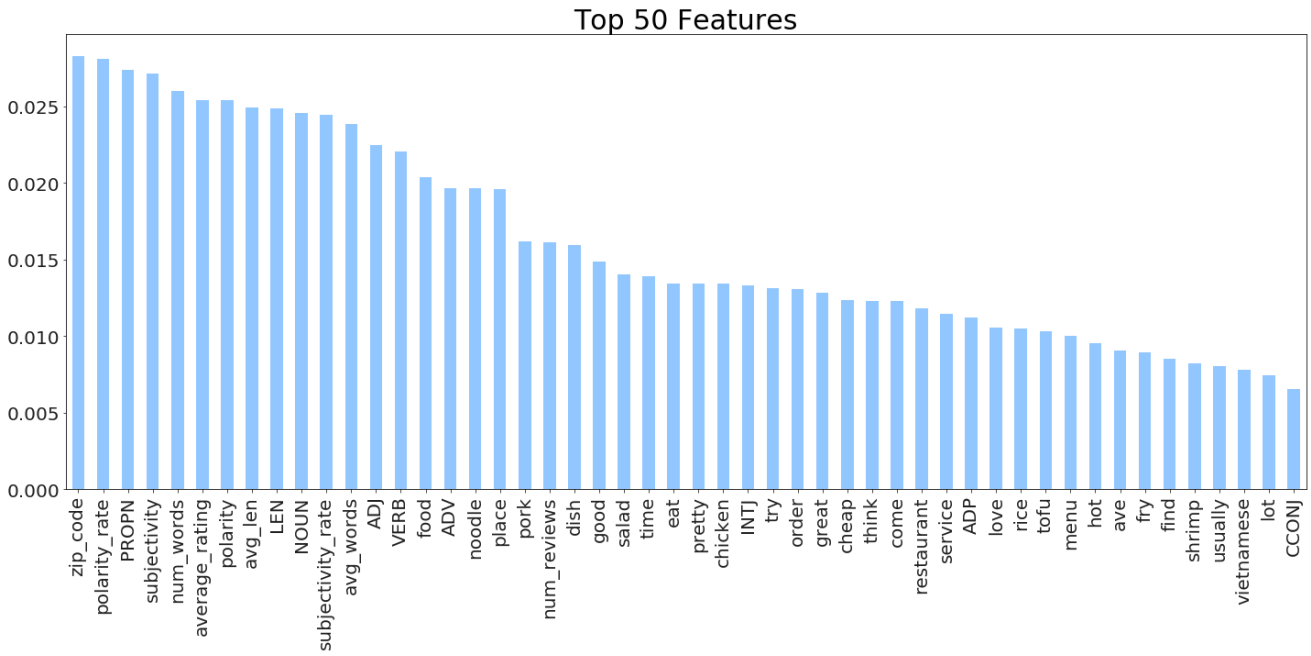


For **MLPClassifier**, it gave top 80 phrases as most discriminating features:



2.5 Analyzing additional features

In this subtask, the top discriminating features from all four feature categories were selected, merged and sorted based on their importance. The top 50 features from all combined additional features are shown in the figure below:



These selected features used instead of the entire text and classified again using an MLP Classifier with 5000 iterations. The reduction result from leaderboard showed and increase in accuracy with an overall score of **72** for F1-Score.

Overall Score		hygiene (1.0)			
Current	Previous	Precision (0.6)	Previous	Recall (0.4)	Previous
0.7203	*0.6622*	0.7325	*0.6643*	0.7021	*0.6592*

These was done using an MLP Classifier with alpha=4.2 iterations. Again, there was an increase in accuracy with an overall score of **75** for F1-Score.

Overall Score		hygiene (1.0)			
Current	Previous	Precision (0.6)	Previous	Recall (0.4)	Previous
0.7494	*0.6622*	0.7543	*0.6643*	*0.7422*	*0.6592*

2.6 Results analysis and interpretation

As can be seen, just review text was not sufficient to predict the inspection status correctly. There where several other attributes about the restaurant facility which directly correlated the failed inspection status.

It was observed that with only a very few of words/phrases the prediction of health inspection failure can be done more reliably. Most important of these features were:

- Location
- Cuisine or category
- Specific dish and restaurant names

Counterintuitively it was also observed that there was no significant correlation between the restaurant rating and the inspection status. Polarity and subjectivity of working in the review was more useful than the average rating for inspection label prediction.

3 Conclusions

It can be concluded that:

1. Preprocessing the review text and generating phrases and n-grams as features increases the accuracy of the classified
2. Only a few words and phrases were needed for accurate prediction.
3. The prediction was also related on non-textual features such as Location, Cuisine etc.
4. The was no significant relation between the business score or restaurant sentiment with inspection status.
5. Most of the failed inspections were in the restaurants handling meat products.