# **Steps of Project**

## *Dataset Used:*

* Yelp NYC dataset @ (1)

## *Pre-processing:*

* Removing punctuations
* Convert to Lower Case
* Tokenization
* Removing Stop Words
* Lemmatization
* Removing empty strings
* Class Imbalance
* ***Bow:***
* N-Grams
* Hashing TF
* TF-IDF Model
* K-Means
* LDA
* **Train-test split(70-30 split)**
* ***Model Building: (with Cross validation 3 folds, Grid Search Hyperparameter optimization)***
* Naïve Bayes
* Logistic Regression
* Random Forest
* GBT
* SVM
* ***Evaluation Metrics***
* Accuracy
* Area under ROC
* Area under Precision-Recall Curve
* Precision
* Recall

References Used:

1. @inproceedings{DBLP:conf/sigkdd/Akoglu15,

author = {Shebuti Rayana and Leman Akoglu},

title = {Collective Opinion Spam Detection: Bridging Review Networks and metadata},

booktitle = {Proceeding of the 21st ACM international conference

on Knowledge discovery and data mining, {KDDí15}},

year = {2015},

}