# Evaluation of Machine Learning based Text Classifiers

- Gautham Giridharan (A20359074)
- Goutham Kannan (A20361163)

# Problem Statement

Evaluating the different methods of Text Classification.

The question that are answered  in this project are,

- How much time it took to classify given set of documents ?
- How  accurately the documents got classified ?

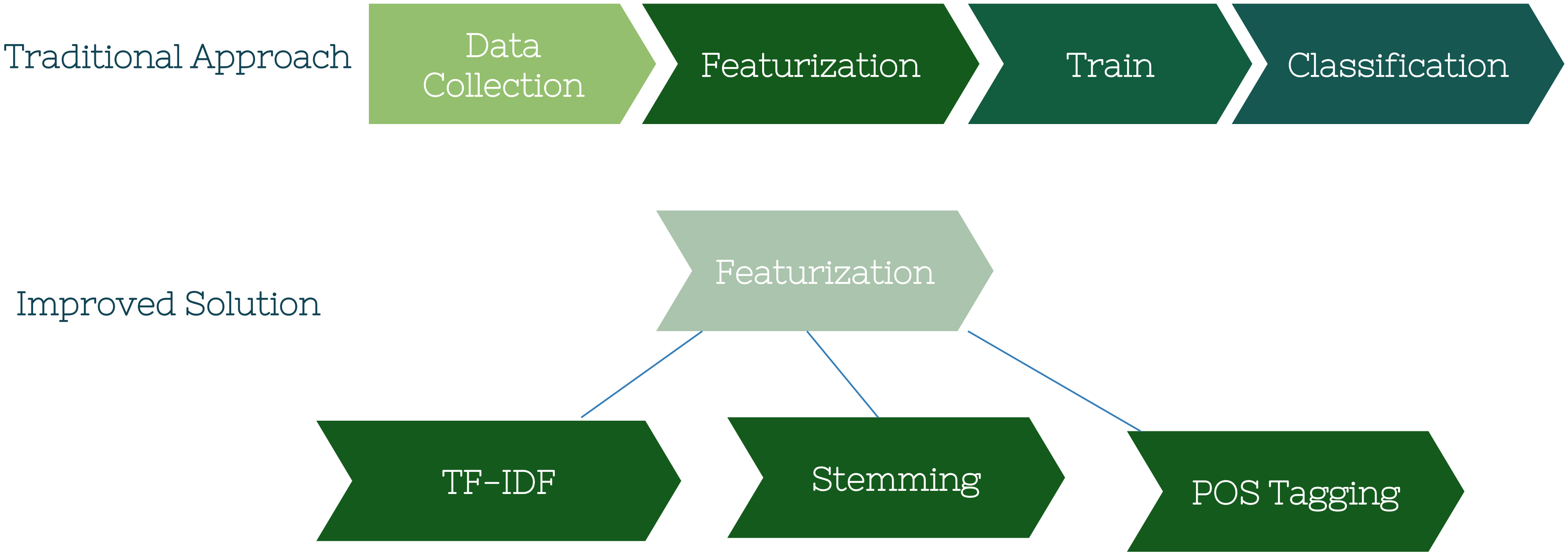
# Proposed Solution !

The idea here is to classify different text datasets using multiple classfiers. SVM , GNB , LR Classifiers are used to classify the text document.

Experiment with different parameters of the classifier.

Better Feature Extraction Techinques. Like Stemming and Part of Speech tagging.

# Implementation



Traditional Approach

Data Collection

Featurization

Train

Classification

Improved Solution

Featurization

TF-IDF

Stemming

POS Tagging

**TF-IDF**

Term Frequecny – Inverse Document Frequency, Feature matrix is build by measuring , the number of time a term occurred in all documents divided by the number of documents the term occurred in.

This results in huge and parse feature matrix. [samples X total number of words]

**Stemming**

Each word is mapped to back to its stem word. This reduces the number of features but doen't lose any information.
Eg. Amazing and Amazed are mapped to Amaze .

**POS**

Part of Speech tagging. Feature matrix contains maximum of words that doesn't affect the classification (Viz.) the,that,and,but,who etc.

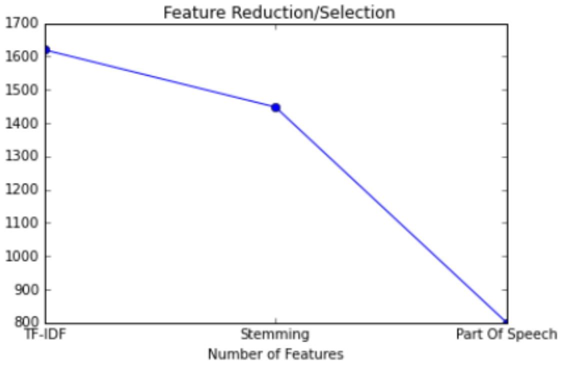
The part of speech of every feature is tagged and words which are adjective,modal auxilary , adverb , pre-determiner ,noun,verbs are only selected.
This reduces the size of feature matrix while retaining the useful informations.

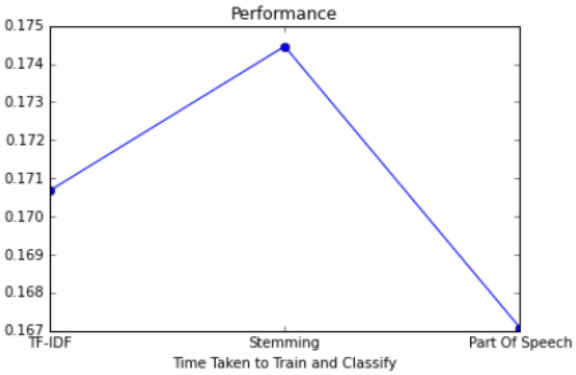
# And tables to compare data

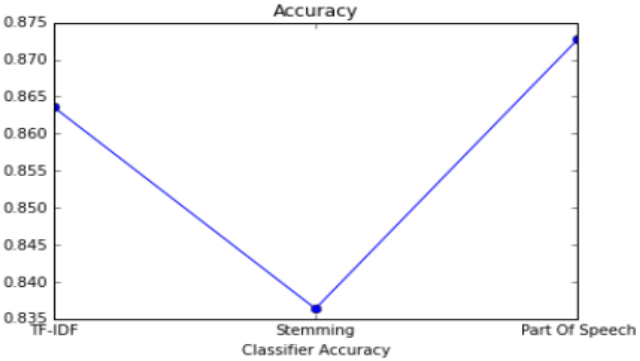
| Feature Size | Yelp Data | IMDB Data |
|---|---|---|
| TF-IDF | 1620 | 4524 |
| Stemming | 1448 | 3835 |
| POS | 801 | 2224 |

| Yelp Data (SVM) | Accuracy | Time Taken(in s) |
|---|---|---|
| TF-IDF | 0.863 | 0.174 |
| Stemming | 0.836 | 0.171 |
| POS | 0.872 | 0.167 |

# And Graphs to show results







# Top Coefficients

| Positive Coefficients |
|---|
| u'great' |
| u'excellent' |
| u'superb' |
| u'worth' |
| u'liked' |
| u'loved' |
| u'enjoyed' |

| Negative Coefficients |
|---|
| u'boring' |
| u'bad' |
| u'terrible' |
| u'poor' |
| u'awful' |
| u'waste' |
| u'almost |

# Conclusion

From the experiments we can see the clear effect of feature selection,using POS tagging, on compution time as well as on accuracy of the classificatiom.

We conclude that for classifying text based on the sentiment , SVM with POS tagging performed the best.

# Reference

1. Lillian Lee, Bo Pang, “Thumbs up? Sentiment Classification using Machine Learning Techniques”,
http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

2. Baharum Baharudin, Lam Hong Lee, and Khairullah Khan.A review of machine learning algorithms for text-documents classification. Journal of advances in information technology,1(1):4–20, 2010
https://www.researchgate.net/publication/43121576_A_Review_of_Machine_Learning_Algorithms_for_Text-Documents_Classification

**THANKS!**

# Any questions?