# Project Summary: Document Classifier

Arjun Patel[1], Harshil Prajapati[1], and Bowen Song[1]

[1]Department of Electrical and Computer Engineering, Boston University

March 24, 2018

## 1 Problem Definition

This project explores machine learning algorithms for document classification. The specific application of the project is sensitive to datasets. The contribution of the project is to compare and contrast algorithm performances based on results and computation efficiency with respect to different types of datasets.

## 2 Literature Review

## 3 Proposed Work

The project concerns algorithms including:

- Bag-of-words model with Naive Bayes assumption
- Multi-class logistic regression (also known as maximum entropy classifier)
- Sensing-aware kernel SVM [1]
- Unigram model
- Markov model.

The bag-of-words model with Naive Bayes assumption is considered the baseline for prediction accuracy. This project also includes stop-words filtering based on stop-words vocabulary list and term frequencyinverse document frequency (tfidf) in attempt to improve performances for each model. The project learns from different datasets; the overview of the datasets is included in later chapters. In addition to discussing algorithm performances, the project also concerns the importance of preserving word ordering within a document with respect to classification accuracy. The project attempts to explore correlations between word vectors generated from Google word2vec algorithm and its occurrences.

### 3.1 Model and Algorithms

### 3.2 Code and Dataset

### 3.3 Minimum Achieviable Plan

## 4 Conclusion

## Division of Labor

## References

[1] W. Ding, P. Ishwar, V. Saligrama, and W. C. Karl, "Sensing-aware kernel SVM," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2947–2951.