

# Project Summary: Document Classifier

Arjun Patel<sup>1</sup>, Harshil Prajapati<sup>1</sup>, and Bowen Song<sup>1</sup>

<sup>1</sup>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

Classification of documents is a traditional problem where usually a bag-of-words approach is to extract features and are used for supervised classification using Naive Bayes or Support Vector Machine (SVM) algorithms [1]. In classical papers, the order of words is ignored but recent study shows use of Convolutional Neural Network (CNN) [2] to get improved results.

Due to the size of documents, different techniques like topological spaces for developing Information Retrieval System (IRS) [3], Ontology-Based Feature Vector Generation [4] have been developed which can effectively reduce the computation while extracting features that aids in classification. Kernel methods (KMs) are an effective alternative to explicit feature extraction. Another way to efficiently extract features is based on finding conditional probability using n-gram models [5].

More over the traditional supervised learning techniques, research has been done to improve prediction in k-NN by using pyramidal decomposition [6], skip-gram and paragraph vectors-distributed bag of words (PV-DBOW) with multiple discriminant analysis [7], expansion method and Powerset-label mechanism for the short hierarchical classification using the Support Vector Machine (SVM) classifier [8], Bayes-optimum decision with the maximum margin principle yields kernels for SVMs [9].

## 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 [9].

- 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

We have multiple datasets in which we want to implement these algorithms. Initially we want to start by using the news group data set, due to previous use of the data set. This dataset has 20 class, 11,314 training documents and 7,532 test documents. The dataset is already preprocessed in a form which can be used. It is broken down into document ID, word ID and word count. Once we test the algorithms on the news group data set, we will test it on another dataset, Reuters21578 but will be using the ModApt version, which is a smaller overall dataset from the Reuters21578 dataset [10]. In this dataset there are 5,945 training documents and 2,347 testing documents. This dataset has also been preprocessed. Finally we want to run these algorithms on the dataset obtained from a Kaggle project, Personalized Medicine: Redefining Cancer Treatment [11]. The data has not been preprocessed so we have the raw files. The data we have is currently, a text file of each paper. There 3,320 documents for training and 5,667 documents for testing. There are more test documents due to Kaggle adding machine generated documents, but these machines generated documents can be removed and the test dataset will be down to 987 documents. There is a total of 9 classes in this dataset. But the biggest job with this dataset is all the preprocessing has to be done by us and made into a usable form, like the other datasets.

### 3.3 Minimum Achievable Plan

## 4 Conclusion

## Division of Labor

## References

- [1] D. S. Sachan, M. Zaheer, and R. Salakhutdinov, "Investigating the working of text classifiers," *arXiv preprint arXiv:1801.06261*, 2018.
- [2] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.
- [3] B. Parlak and A. K. Uysal, "On feature weighting and selection for medical document classification," in *Developments and Advances in Intelligent Systems and Applications*. Springer, 2018, pp. 269–282.

- [4] M. K. Elhadad, K. M. Badran, and G. I. Salama, “A novel approach for ontology-based feature vector generation for web text document classification,” *International Journal of Software Innovation (IJSI)*, vol. 6, no. 1, pp. 1–10, 2018.
- [5] J. Fürnkranz, “A study using n-gram features for text categorization,” *Austrian Research Institute for Artificial Intelligence*, vol. 3, no. 1998, pp. 1–10, 1998.
- [6] P. Heroux, S. Diana, A. Ribert, and E. Trupin, “Classification method study for automatic form class identification,” in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol. 1. IEEE, 1998, pp. 926–928.
- [7] P. Lauren, G. Qu, F. Zhang, and A. Lendasse, “Discriminant document embeddings with an extreme learning machine for classifying clinical narratives,” *Neurocomputing*, vol. 277, pp. 129–138, 2018.
- [8] Z. F. Salih and S. Tiun, “Term expansion and powerlabel set for multi-label hierarchical on short document classification,” *International Journal of Applied Engineering Research*, vol. 13, no. 1, pp. 539–544, 2018.
- [9] W. Ding, P. Ishwar, V. Saligrama, and W. C. Karl, “Sensing-aware kernel svm,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2947–2951.
- [10] “Collection of datasets,” <http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html>.
- [11] “Kaggle data: Personalized medicine: Redefining cancer treatment,” <https://www.kaggle.com/c/msk-redefining-cancer-treatment>.