# COSC 522 - Machine Learning: Milestone 3

Hector D. Ortiz-Melendez and Kevin De Angeli
*University of Tennessee - Fall 2019*

## Preliminary Results

### A. Pre-Processing

The dataset selected for this project consists of three individual datasets containing people's reviews of movies, products, and restaurants from IMDB, Amazon, and Yelp, respectively. Each of these individual datasets consists of 1000 entries, where 500 are positive reviews, and 500 are negative reviews. So far, we have worked with these datasets individually, training the algorithms using one of these three datasets at the time. However, we plan to experiment by eventually merging the datasets into one single dataset and observe the accuracy of the models. The outcomes of this experiment will be interesting because the type of vocabulary used to review a movie, for example, can be significantly different than the type of words you may find on a product review; these results will be part of the final report. For our preliminary results, we have applied two common feature extraction techniques in Natural Language Processing (NLP): bag-of-words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Table 1 shows the dimension of the training datasets after obtaining the features. The number of columns represent the size of the dictionary of each of the datasets minus words such as articles that do not carry much information.

**Table 1 Dataset dimensions after features were extracted.**

| Method | Amazon | IMDB | Yelp |
|---|---|---|---|
| **BoW & TF-IDF** | [1000 x 1848] | [1000 x 3047] | [1000 x 2035] |

### B. Accuracy

A Pythom script was written to train a Support Vector Machine (SPM), Decision Trees (DT), and Backpropagation Neural Network (BPNN) with two hidden layers. Table 2 displays the accuracies for both feature extraction methods chosen: BoW and TF-IDF. For all three classifiers, it can be concluded that using TF-IDF features, SVM is the most accruate model (up to 82.73% accuracy), and with BoW features, BPNN performs better than the rest (up to 79.39% accuracy). The maximum accuracy was obtained with SVM on the Yelp data set. BPNN with BoW features in the Yelp dataset was the worst performing method. However, BPNN proved to be the best method for the Amazon dataset. Finally, in comparison with the other two classifiers, DT was consistently mediocre across the board.

**Table 2 Accuracy results for: Support Vector Machine (SVM), Decision Tree (DT), and Backpropgation Neural Network (BPNN)**

| Method | Feature | IMDB (%) | Amazon (%) | Yelp (%) |
|---|---|---|---|---|
| **SVM** | BoW | 66.36 | 76.67 | 75.45 |
| **BPNN** | BoW | 73.64 | 79.39 | 47.58 |
| **DT** | BoW | 64.55 | 72.12 | 71.82 |
| **SVM** | TF-IDF | 79.09 | 81.21 | 82.73 |
| **BPNN** | TF-IDF | 78.18 | 82.12 | 47.58 |
| **DT** | TF-IDF | 62.73 | 70.61 | 70.91 |

### C. Discussion

Future works includes applying three Gaussian classifiers from scratch, and the K-nearest neighbor algorithm. We will apply classifier fusion to see if a greater accuracy can be obtained when multiple algorithms' predictions are

considered. We will attempt to include a section on a most complex feature extraction technique such as word2vec. The problem with these methods is that vectors representing individual reviews will vary on length, and this can present a challenge for ML algorithms. We plan to analyze what aspect of the Yelp dataset makes BPNN fail so spectacularly, and what makes SVM perform so well in the IMDB dataset. We will also research dimensionality reduction techniques for NLP. We will apply n-fold cross validation (right now the dataset is split as 70-30%), and analyze the performance using metrics such as the ROC curve.