MAIS 202 - Final Project

Zahur Ashrafuzzaman, Jules Barbe, Benjamin Segall

Deliverable 2

1 Problem Statement

The goal of this project is largely unchanged from what was stated in the previous Deliverable: to create an algorithm that will:

1. Given an article, classify it as legitimate news or false news
2. Redirect the user to a relevant article from a trustworthy news source if the article was classified as false.

2 Data Preprocessing

We are using tf-idf vectorizing with stop word removal included. This cleans up the data and keeps only words relevant throughout many articles.

3 Machine Learning Model

We are using an SGD classifier based on an SVM with stochastic gradient descent optimization. It is also using an optimal learning rate which calibrates itself automatically during training.

4 Preliminary Results

Our results are very good for now, with training accuracy constantly over 99 percent. As the model is very fast, we have been able to train it many different times on different splits and we always get this level of accuracy. As the hyperparamater training is done automatically through optimal learning rate, we have decided to also test the model on different splits of the test set (retraining the model from scratch everytime, obviously), where we always get an accuracy of at least 98For a test set of about 9000 samples, we always get less than a 100 false positives and negatives on every run.

5 Next Steps

1. For this model we only kept the article body of text as a feature, but it would be interesting to include the title in the model. However the pre-processing on it should be different, as we consider that stop words, caps and punctuation to be of relevant importance when it comes to the title of fake/true articles. Either we could run a seperate model on the title (which would not take long at all), or we could preprocess the titles seperately and add specific weights to the features we take out of it.

2. Even though our results are very good, we should still try to implement different models, and do so in a more automated way (like having a seperate preprocessing function and a training function that could iterate through different models from sklearn)

3. We also have to tackle article comparison which we will use in our final product. After doing research, we came to the conclusion that Latent Semantic Analysis (LSA) with Singular Value Decomposition (SVD) would be optimal. This can be done fast and on the fly to compare articles deemed fake to alternate articles we find online from trusted databases.