CS 5100: Foundations of Artificial Intelligence

Final Project: Milestone 1 (11/11)

Team: Amey Arya and Nikhita Singh

BUILDING A SPAM FILTER USING MACHINE LEARNING

For the first milestone, we had four major action items. They are discussed as follows:

1. **Learning Naïve Bayes, Random Forest, and Decision Tree classifiers:**

   One of the most common ways to execute a spam filter is Naïve Bayes classifier. The Naive Bayes algorithm relies on Bayes Rule. This algorithm will classify each object by looking at all its features individually. Bayes Rule below shows us how to calculate the posterior probability for just one feature. The posterior probability of the object is calculated for each feature and then these probabilities are multiplied together to get a final probability. This probability is calculated for the other class as well. Whichever has the greater probability that ultimately determines what class the object is in? We read various papers and articles to gain a better understanding of the same. Provided are the links we referred:
   https://iopscience.iop.org/article/10.1088/1757-899X/226/1/012091/pdf
   https://pdfs.semanticscholar.org/2021/61ac4674000870a0d450c6410f140f42216c.pdf
   https://en.wikipedia.org/wiki/Naive_Bayes_classifier

   Random forest, like its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. We read various papers and articles to gain a better understanding of the same. Provided are the links we referred:
   https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1
   https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
   https://en.wikipedia.org/wiki/Random_forest
   https://pdfs.semanticscholar.org/a06f/4cee2b93ee14224a8aedfe10e653a2d7c879.pdf

   Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine

learning. We read various papers and articles to gain a better understanding of the same. Provided are the links we referred:

https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567

https://en.wikipedia.org/wiki/Decision_tree_learning

https://pdfs.semanticscholar.org/2fbc/3021ecba5eadfaea280fc08b55c5ecd24aae.pdf

https://www.researchgate.net/publication/258651642_Spam_Mail_Filtering_Technique_using_Different_Decision_Tree_Classifiers_through_Data_Mining_Approach_-_A_Comparative_Performance_Analysis

2. **Setting up the environment on the local machine as well as GitHub:**

We have set our local machines for the project and worked on the basic data cleaning. Additionally, we have set up the private ccs GitHub repository for the same.

3. **Cleaning and scraping the dataset:**

We are using the Ling-Spam Dataset, provided by Ion Androutsopoulos which is available publicly. It is unfiltered and requires cleaning. Here is a link for the same: https://aclweb.org/aclwiki/Spam_filtering_datasets. The filters we applied are:

- Lower-casing: The entire email is converted into lower case, so that capitalization is ignored (e.g., IndIcaTE is treated the same as Indicate).
- Stripping HTML: All HTML tags are removed from the emails. Many emails often come with HTML formatting; we remove all the HTML tags so that only the content remains.
- Normalizing URLs: All URLs are replaced with the text “httpaddr”.
- Normalizing Email Addresses: All email addresses are replaced with the text “emailaddr”.
- Normalizing Numbers: All numbers are replaced with the text “number”.
- Word Stemming: Words are reduced to their stemmed form. For example, “discount”, “discounts”, “discounted” and “discounting” are all replaced with “discount”. Sometimes, the Stemmer strips off additional characters from the end, so “include”, “includes”, “included”, and “including” are all replaced with “includ”.
- Removal of non-words: Non-words and punctuation have been removed. All white spaces (tabs, newlines, spaces) have all been trimmed to a single space character.

As an example, we take the 3-380msg4.txt from the above-mentioned dataset. The message looks something like this:

Subject: postings

hi , i ' m working on a phonetics project about modern irish and i ' m having a hard time finding sources . can anyone recommend books or articles in english ? i ' , specifically interested in palatal (slender) consonants , so any work on that would be helpful too . thanks ! laurel sutton (sutton @ garnet . berkeley . edu

Post cleaning the data we obtain a result as follows:

posting hi m work phonetics project modern irish m hard source anyone recommend book article english specifically interest palatal slender consonant work helpful too thank laurel sutton sutton garnet berkeley edu

In this way we have cleaned the dataset and categorized it into four broad categories, namely, non-spam train dataset, spam train dataset, non-spam test and spam test dataset.

4. **Deciding the train-to-test ratio:**

   Upon cleaning the dataset and further discussions, we decided on a train to test ratio of 9:1. This will allow us to properly train our ML model and also will leave the 10% part for proper testing of the model.