Ayleah K. Hill

Prof Ananth Jillepalli

CPTS 315

November 2, 2020

**Project Proposal**

1. Data Mining Task:
   Can I build a prediction model that can accurately determine whether text messages are legitimate or spam messages?
   Spam messages have become a common nuisance that everyone deals with. With the development of Covid-19 over these last few months, I have noticed a spike in spam messages that take on a variety of formats including texts, emails or phone calls. With people being stuck at home due to the pandemic, some people may have more free time to pay attention to messages that they receive, including spam messages. As spam messages continue to plague us with no end in sight, it would help to be able to build machine learning algorithms and utilize data mining skills to be able to identify whether texts are spam messages based on their content. Spam messages not only have the ability to waste our time but also have the ability to obtain sensitive information from us that can put us in dangerous situations.
2. Dataset:
   The dataset I intend to use is provided by the UCI Machine Learning Group which posted the dataset on Kaggle:
   https://www.kaggle.com/uciml/sms-spam-collection-dataset
   That dataset contains 5,574 text messages total, consisting of a mix between legitimate texts and spam texts (identified as either ham or spam accordingly) where 87% of the messages are “ham” and the other 17% are “spam”. The two columns in the data set are for labelling the texts as “ham” or “spam” and for holding the content of the texts.
3. Methodology:
   Using something like Weka to develop classifications and clusters based on text content thus create visualizations to show what words are most commonly used in texts that are spam messages, potentially using semi-supervised learning techniques. I also was to use the Apriori algorithm on the data set to see if I can find trends of particular groups of words (not just individual words) that can be found together in spam or legitimate texts and see if the sets of words in both legitimate texts and spam messages are unique.
4. Final Product:
   The final product will consist of visualizations to show what words are most commonly used in spam texts (in the form of charts). My accuracy will be measured in the accuracy of the learning model being able to identify in the test set of texts (with hopefully 100% accuracy) what messages should be identified as spam.
   This project will teach me how to be able to mine data, use machine learning techniques to find trends in data and how to make precise and accurate predictions based on a small amount of labelled data.