Name: Rohan Patel
SFSU ID: 917583698

SMS Spam Detection Using Text Mining and Machine Learning

1. Project Description

Over the years, as the use of mobile phones has increased, we have seen a huge surge in unwanted commercial advertisements sent to mobile phones using text messaging. SMS spam has become a major problem because firstly, spam messages are irritating, and secondly, because in some countries they incur a cost to the receiver as well. This, along with the lack of major mobile phone spam filtering software is the motivation behind looking into the problem of SMS spam detection.

2. Data

The dataset used is the SMS Spam Collection Data Set from the popular UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/sms+spam+collection). The dataset contains a total of 5574 messages which includes 747 spam messages and 5825 non-spam (ham) messages. The image below shows the head of the dataset after being converted into a python dataframe.

	label	message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

3. Methodology & Project Goals

The basic methodology that will be followed is as follows: first the data will be pre-processed, then after feature engineering, some text mining technique such as TF-IDF will be applied. Different machine algorithms will then be applied to classify the messages as spam and ham. The results of the machine learning techniques will be compared to evaluate the best algorithm for spam filtering of text messages. Finally, cross-validation will be used to reduce the overall error-rate of the best classifier.

The main goal of this project is to apply different machine learning algorithms to the SMS Spam detection problem, and then compare their performance to gain more knowledge that can help us filter SMS spam with higher accuracy.