Machine Learning Engineer Nanodegree
Capstone Proposal
Bhavya Garg
March 28, 2017

Domain Background

Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken. NLP is a component of artificial intelligence (AI) which is used to analyze text, allowing machines to understand how human's speak. This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more. NLP is commonly used for text mining, machine translation, and automated question answering. Natural language processing technology is designed to derive meaningful and actionable data from freely written text. This is particularly useful because it allows to record information in a natural manner and when executed well, it enables a more natural transition between person and database. Here being inspired by spam email classification I've chosen to work on text messaging data to classifying spam/ham messages. Spam is a waste of time and takes unnecessary storage space. There are many other areas where NLP are being used, for example YELP is using NLP to understand people comments and provide ratings based on that.

Problem Statement

Natural language processing (NLP) principles and machine-learning algorithms move beyond traditional basic keyword and regular expression search and word count statistics that are commonly used in mobile device forensic analysis. The ability to detect linguistic patterns is an invaluable tool when applied to text messaging data.

The goal of this project is to classify SMS phone messages to Ham/Spam using a collection of more than 5 thousand SMS phone messages labeled ham and spam examples. I'll first train a machine learning model to learn to discriminate between ham/spam automatically. Then, with a trained model, I'll be able to classify arbitrary unlabeled messages as ham or spam.

Datasets and Inputs

This project will use SMS Spam Collection(CORPUS)which is a set of SMS tagged messages that have been collected for SMS Spam research)
(https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection).
The dataset of SMS messages has 5574 messages in total out of which 4827 SMS are ham and 747 are spam.

The SMS Spam Collection file contain one message per line. Each line is composed by two columns: one with label (ham or spam) and other with the raw text
The files contain one message per line. Each line is composed by two columns: one with label (ham or spam) and other with the raw text. Below are some examples:

ham What you doing?how are you?
ham Ok lar... Joking wif u oni...
ham dun say so early hor... U c already then say...
ham Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.
spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop
spam Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B
spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Solution Statement

This project will attempt to Classify the correct class label for a given input. There would be some exploratory data analysis and text preprocessing steps involved in this project and once patterns are identified and a data set is constructed from the text-messaging corpus utilizing the Python-based Natural Language Toolkit (NLTK), machine learning algorithms can be applied to a training set. I'll be using Naive Bayes classification algorithm because it really works well in classifying texts since it apply Bayes' theorem on the context classification of each messages, with a strong assumption that the words included in the messages are independent to each other and also they typically use bag of words features to identify spam. The model will then be used to predict spam vs ham classification.

Benchmark Model

Although the main aim of this project is to classify the SMS messages to ham/spam but the measure of success is how accurately my model is making prediction in classifying ham vs spam messages. I am choosing the benchmark model with the accuracy of 87% considering all messages as ham since my data sets are strongly biased towards ham and I can get a minimum accuracy of 87% by just saying that every single email is ham, which is obviously something that indicates a non-ideal classifier. So, my aim here is to focus on getting over 90% accuracy in Area Under ROC curve (which considers all the threshold values and not discriminative) instead of just the overall accuracy.

Evaluation Metrics

Model evaluation is the phase where we want to determine how well our model will do overall on the entire dataset. There are quite a few possible metrics for evaluating model performance. Which one is the most important depends on the task and the business effects of decisions based off the model. I will use SciKit Learn's built-in classification report, which returns Accuracy, precision, recall, f1-score. Since my dataset is imbalanced where one class dominates over the other, Accuracy is not only the metric to prefer because that is misleading and in my case, there are 4827 ham SMS and 747 spam meaning the ham messages are dominating the spam and if I calculate the accuracy it will give me 87% accuracy considering all messages as ham. So I prefer calculating precision which is the ratio of correct positive predictions made out of the total positive predictions made, recall which is the ratio of correct positive predictions made out of the actual total that were positive(correct positive prediction plus incorrect false negative prediction)and F1 score which can be interpreted as a weighted average of the precision and recall, as well along with the Accuracy.

Project Design

I'll follow the below steps:

- ***Download Dataset:*** The first step of the project will be to download dataset from the UCI datasets.
- ***Exploratory Data Analysis***: Once I have the dataset ready, will do some basic Exploratory data analysis on the dataset to get familiar with it.
- ***Pre-Processing:*** Once done with EDA, I'll do some text preprocessing steps like split a message into its individual words, remove common words, ('the', 'a', 'can', 'they' etc.) using NLTK library.
- ***Normalization***: This includes stemming words or distinguishing by part of speech
- ***Vectorization***: Converting each of the messages into a vector that machine learning models can understand. We'll do that in three steps using the bag-of-words model
- ***Training a model:*** Once the messages are represented as vectors, we can finally train our spam/ham classifier. We can use any kind of classifier to build a model but I prefer Naive Bayes classification algorithm since it serves well for classifying texts.
- ***Model evaluation***: Now once we have the model we want to determine how well our model will do overall on the entire dataset. The best way to evaluate our model is to split the data into a training/test set, where the model is built based on the training dataset and we check the performance of the model on the test dataset.