NLP Project

SPAM DETECTION IN TWEETS ENGLISH

K. HEMANT

K. CHETAN

Problem Description

Twitter has been one of the fastest adopted social network in recent history. Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, **500 million** tweets per day and around 200 billion tweets per year.

With such popularity, it is no surprise to see the high occurrence of spam tweets, which never cease to irritate and waste our time. We aim to build a classifier which given a tweet, can output whether the tweet is spam or not. We use the available SMS data to train our classifier.

Methodology

We basically use a Naïve Bayesian classifier, which classifies a given tweet based on probability values obtained from training data.

Initially, we make dictionaries of each word's occurrences both for spam sentences and ham sentences.

Then we use the Bayes Formula to find the probability whether a given sentence is ham and spam, treating each word to be independent in the sentence. This probability measure is what we use to classify.

The "frequent item" idea was also used, where we remove the words with high occurrence(stop words), which increases our accuracy overall.

Laplacian smoothing was also used for smoothing the probability values.

The given data was split into two groups of 80 percent and 20 percent, with the former being used to train and the latter being used to test. It was ensured that both have a similar ratio of spam to ham sentences.

Problems Faced

As the given data was SMS data, there was a high occurrence of abbreviations like ‘btw’ for ‘by the way’, ‘brb’ for ‘be right back’, etc. So, this skewed our results a bit.

We had to use trial and error to find a correct vocabulary size, too high or too low values were giving high discrepancy in results.

Our results were improved when we used the “frequent item” method. With just Naïve Bayes Classification, there was a higher error rate.

Also, many words were followed by ‘.’ or’..’, which we had to remove.

Results/Observations

Our classifier had an accuracy of 98%.

Our spam recall value was around 91% and the ham recall value was around 99 percent.

Thus, the Naïve Bayes Classifier with the “frequent item” method thus gives a high accuracy than just the Naïve Bayes Classifier itself.(for this given SMS data)

The high accuracy percentage is also attributed to the properties of the data, which is small and has only around 800 spam SMSs in around 5000 good SMSs.

The low spam recall value is attributed to the data given, where we can see quite a number of one time occurrences like ‘wining’ instead of ‘winning’, random numbers with words, etc. This skews the probability value, and even results in some of the wrong classifications.

Our spam recall value tells us that our classifier is a safe classifier, insofar as it almost always does not classify a ham as spam, but vice versa has more chance of happening, as seen from the recall values.

Naïve was mainly used as it was recommended by our mentor to use for this SMS data and we also found it to have high accuracy.

THANK YOU