**NLP Project Report**

**Pre-processing**

First, I stripped out the speech marks from the text replacing them with spaces.

Then I expanded out standard contractions that I have compiled in a txt file. In cases of ambiguity of contractions, I selected one of the pair. This still leaves some contractions where they are less widely used and more likely to be character specific.

Then I striped apostrophise to avoid names splitting. This is to avoid my regex tokenizer splitting john’s into john s rather than johns when it is removing punctuation.

**Training.**

**Testing**

**Improvement**

To make the classifier more advances I used word2vec. Mostly because I found it interested because of the knowledges that can be represented and explore from the vector relations between words and I wanted to see those results on a more specific data set as in the project.

**Training**

I first trained a word2vector model over the whole of the training data set with the hope that it generalises the data well and will function on the testing set without having trained the model on this set. I used a minimum word frequency of 3 in the word2vector model parameters in a hope to capture as many words in the model as possible because this is a reasonably small dataset.

I created a vector for each sentence by averaging the word2vectors of each word in the given sentences. This should results in a vector that is a descriptor of the sentences semantic meaning.

If a word is not in the word2vector model ignore the word in the averaging.

**Cross validation**