**NPL 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.