We process all the tweets as follows:

1. Replace all the URL's with blank.

2. Replace all hashtags (#) and targets (For ex: @James) with blank.

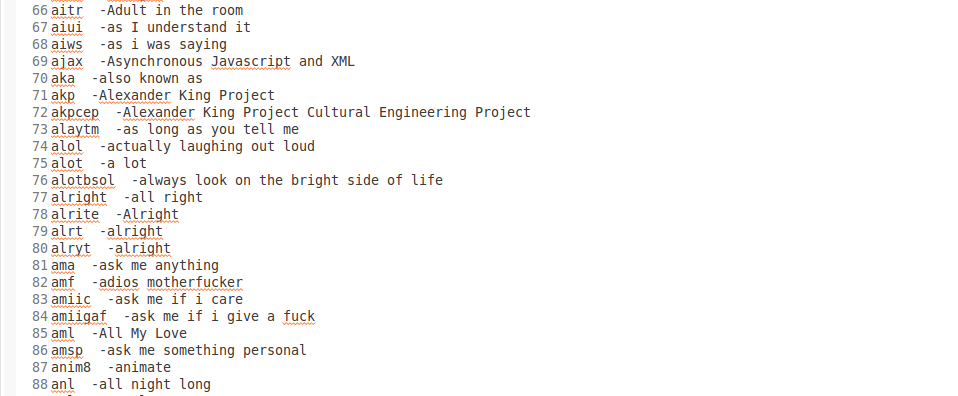
3. Replace acronyms and emoticons with their assigned meanings.

4. Replace negations with word 'NOT'.

5. Tagged the tweets.

6. Stemmed the words to its root word.

For replacing acronyms we have prepared an acronym dictionary from all the slangs listed in www.noslang.com/dictionary/. The dictionary has translations for 5300 acronyms. Our program compares replaces the acronyms with their translated meanings in the tweet. Below image is the snapshot of our acronyms dictionary.



For handling the presence of emoticons we have compiled a list of 160 emoticons listed in Wikipedia (http://en.wikipedia.org/wiki/List\_of\_emoticons) and have assigned them polarity according to their usage. For ex: “:)” this smile symbol has been assigned a positive polarity as it is mostly used for expressing happiness.

We have also replaced negations with 'NOT' and for selecting negations we have referred to the negative words provided by the Grammarly handbook ([www.grammarly.com/handbook/sentences/negatives/](http://www.grammarly.com/handbook/sentences/negatives/)). Our algorithm converts can't to can NOT (Since, n't is the negation here) and cannot to NOT.

We have used Standford's POS tagger to tag the tweets and referred to Apache Lucence Stemming algorithm to stem the words. In any text adjectives are good indicators of subjective, evaluative sentences. However an adjective might give us an understanding of subjectivity but an adjective alone could not shed much light on a sentence semantic orientation. For ex: the adjective “unpredictable” might have a negative orientation in automotive context (unpredictable steering) but it might have a positive context in a movie review(unpredictable plot). Thus, we have implemented an algorithm which extracts the two consecutive words in a tweet if and only if the combination of the two extracted words satisfies any of the below mentioned conditions:

|  |  |  |
| --- | --- | --- |
| **First Word** | **Second Word** | **Third Word(not selected)** |
| JJ | NN or NNS | anything |
| RB, RBR, RBS | JJ | Not NN nor NNS |
| JJ | JJ | Not NN nor NNS |
| NN or NNS | JJ | Not NN nor NNS |
| RB, RBR or RBS | VB, VBD, VBN or VBG | anything |

For ex: if we take the third case then the combination JJ , JJ will only be selected if the third consecutive word is not a category of noun(NN or NNS).

After cleaning the tweets by applying all the above mentioned steps we apply the input to train our models.

References:

List of Slangs: [www.noslang.com/dictionary/](http://www.noslang.com/dictionary/)

List of emoticons: <http://en.wikipedia.org/wiki/List_of_emoticons>

Negations: [www.grammarly.com/handbook/sentences/negatives/](http://www.grammarly.com/handbook/sentences/negatives/)

Study of Semantic Orientation by National Research Council Canada

Research on Sentiment Analysis of Twitter data by Columbia University