GTBT Report 23/01/2015

**Architecture**

The architecture of the project has changed since last year. The project previously comprised of three components with data passing between the three components using HTTP Requests.

This was changed so that the client application was more responsive and to reduce loading times as the number of AJAX calls needed to render a view was growing and affecting the usability of the interface.

Another reason for the change was that the client application was using the Django framework without any real requirement to do so. The CA only rendered the HTML/CSS and JS therefore the only purpose of the Django framework was to provide a webserver.

Therefore, the service component and the CA component were integrated. This integration should also make things easier to manage.

**Feed-Enrichment**

Different document features and applying different techniques – such as removing URLs – has been implemented in the NLTK implementation. The results of this have been recorded and have shown improvements and are now included in the component.

More crowdsourcing was carried out and using the CrowdFlower platform I was able to add approximately 2000 tweets to the training set. I did attempt to carry out more crowdsourcing jobs however the cost became prohibitive as the cost estimation grew and the time to complete increased.

While working on the project between the 25th of December and the 1st of January, I accidentally deleted my database. I managed to retrieve approx. 50,000 tweets back from an earlier back up but I have had to start from scratch.

I also created an implementation in Java using Weka. I started this because the resources need to build a model with NLTK exceeded the number of resources available in my MacBook. Moving to Weka managed to also reduce the time taken to build models.

I also came across a website called sentiment140 which is a service that allows users to discover the sentiment for a product or brand on twitter. This service offers the training data they used to build their model, which allowed the first foray into using sentiment within the project. The training data they provide is a collection of 1.6 million tweet texts labelled as Positive or Negative. The collection is too large to use on my MBA and currently I have not seen any sign of completion on the lab machines.

**Client Application**