Analysis of Sentiment Direction Based on Two Centuries of the Hansard Debate Archive

Author:
Moreno Vardanega

Supervisor:
Prof. Leslie Smith

September 2016

Dissertation submitted in partial fulfilment for the degree of

Master of Science in Big Data

Computing Science and Mathematics

University of Stirling

Abstract

This project was conceived by Nalanda Technology, a text data search and analysis company. This study concerned the following central research question: Can Machine Learning techniques be used to analyze automatically the change in aggregate levels of sentiment polarity within a defined topic in UK parliamentary speeches?

We addressed the question by making use of methods of natural language processing (NLP). The project evaluated whether or not opinion-mining techniques can successfully be applied to the textual analysis of parliamentary debates. For data, we used the transcriptions of the speeches made at the plenary parliamentary meetings as reported in the Hansard Archive (digitised debates of the UK Parliament from 1803 to the present). A Sentiment Model was implemented to retrieve meaningful patterns based on the classification results. A substantial selection of parliamentary proceedings was compiled and a Gold Standard Corpus (GSC) created, covering both subjectivity and orientation.

The main idea of this study is that an automatic sentiment analysis model can be very useful for rapidly visualizing how the opinion of the parliament – which normally reflects the national sentiment – changes over a period of time. This enables the user to put together an overall picture of the position on a specific issue and to pick up on the underlying trends without having to go through the entire sequence of the Hansard speeches. The visualizations have to be clear and easy to understand, with all ambiguities eliminated.

The Sentiment Classifier was built on top of the Nalytics search engine, the core product of Nalanda Technology, to add a new analytical feature to this data-search and discovery platform. The classifier was the outcome of a series of experiments, including an analysis of the performance levels of various supervised learning algorithms, such as Support Vector Machines, Linear Regression, Naive Bayes, etc. These techniques and methodologies were written in Python using the NLTK and SciKit-Learn toolkits. The model achieved a classification accuracy level of 75% when trained and tested with the GSC – this was considered a promising result, in light of the complex nature of parliamentary discourse.