Retrieving and parsing linguistic expressions of political attitudes*

Paul Nulty
paul.nulty@gmail.com

Social media is frequently used as a platform for the broadcast of political opinion and engagement, both by citizens and their representatives. In this paper we introduce methods from information retrieval and natural language processing to improve the analysis of political expression from Twitter posts. Research which makes use of information retrieved from social media typically performs a keyword search to return posts relevant to the topic of interest. Text analysis of the retrieved content often uses a stemmed bag-of-words model, whereby the linguistic expression is reduced to a count of word lemmas. These approaches ignore two fundamental aspects of linguistic communication. First, the Zipfian frequency distribution of word types and word senses means that the precision of results returned from keyword searches varies widely according to the baseline frequency, ambiguity, and vagueness of the search term used. Second, human language is richly structured, and this structure is lost when utterances are reduced to word counts.

Borrowing methods from information retrieval, we can refine the keyword search procedure to balance precision and recall, while accounting for the fact that most words, names and acronyms have multiple meanings. Once we have retrieved tweets of interest, we use syntactic dependency parsing to extract a more subtle interpretation of political statements than the word-frequency matrix approach. We demonstrate these methods by collecting and parsing tweets that indicate political engagement or opinion in the UK, both from the general public and their political representatives. To do this, we isolate posts pertaining only to UK political actors, using a confidence threshold to balance precision and recall. We then use syntactic dependency parsing to measure sentiment and behaviour by extracting the verbs and adjectives that fill the argument roles of noun phrases signifying political actors, enabling us to differentiate, for example, between *'Cameron criticized Clegg'* and *'Clegg criticised Cameron'*. Phrases such as this, which vary only in word order, are impossible to distinguish using the bag-of-words model.

*Author address: Department of Methodology, London School of Economics and Political Science, London WC2A 2AE. Abstract prepared for presentation at the Computational Social Science workshop, Lucca, Italy, 24th September 2014. This research was supported by the European Research Council grant ERC-2011-StG 283794-QUANTESS.