Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Brexit? Analyzing Opinion on the UK-EU Referendum within Twitter

Clare Llewellyn, Laura Cram

The Neuropolitics Research Lab, School of Social and Political Science, University of Edinburgh, Edinburgh, UK
C.A.Llewellyn@sms.ed.ac.uk, Laura.Cram@ed.ac.uk

Abstract

We present a demonstrator that visualizes the Twittersphere debate on whether the UK should remain in or leave the European Union. Data is collected using three strategies: hashtag search terms, extraction from the full stream and following specific users. The demonstrator can be used to show the different discussion topics identified by the different search strategies.

Introduction

We are tracking the UK debate on the EU referendum in Twitter to explore the various ways in which the public imagines the European Union. We ask how this relates to the cognitive frames that predominate in the offline public and political dialogue and explore the process through which competing cognitive frames come to predominate in political debate.

We have collected data on the EU referendum debate from the Twitter API since August 6th 2015. This data has been gathered using three search strategies, 1) tweets collected that contain a set of European Referendum specific hashtags, 2) relevant tweets extracted from the public full stream API, and 3) tweets collected from specific users, the Twitter accounts from the official campaign groups.

We examine how topics and language differ between these groups and how they influence and cross-pollinate each other. The specific hypotheses we are exploring with this demonstrator are whether 1) the different datasets contain discussion on the same topics and can be used as proxies for each other, and 2) the official campaign groups direct the discussion which we would notice through an echo chamber effect as discussion topics from the official campaign groups permeate over time to the other datasets.

This analysis is publicly accessible through our custom- built interactive demonstrator.

The UK and the EU

On the 23rd June 2016 the people of the UK will vote in a referendum on whether to remain within the European Union. The two sides of the debate are, for the UK to remain as part of the European Union (pro-Remain) or for the UK to leave the European Union (pro-Leave). Within the pro-leave campaign there are three main sub-groups: Vote Leave, Grassroots Out and LEAVE.EU. Pro-remain has a single dominant campaign group: Britain Stronger in Europe.

Gathering Data

We are using a Twitter dataset to explore the relationship between the UK and the EU and how people talk about this relationship. We aim to find out what people are saying on Twitter and to investigate how this changes leading up to a referendum on the UK's membership.

Data has been gathered as part of an ongoing process from the public Twitter API since August 6th 2015. The data is being collected in 3 sets. The first set is the 'hashtags' set, Twitter data sets on specific topics are commonly gathered by searching using hashtags, in this set data is collected from the API using UK-EU specific hashtags chosen by a panel of experts. These terms are updated periodically. This method is often criticized as the set collected is biased towards the hashtags chosen (Tufekci 2014). The second method, the 'stream' set aims reduce the bias introduced through human defined search terms.

The second set, the 'stream' set, is extracted from the full stream collected through the API. Data that is related to the EU discussion is extracted from the set using a method based on Llewellyn et al (2015). This involves using very broad non-opinion based search terms to gather a specific set from the overall dataset, for example the search terms 'euref' and 'eureferendum'. This extracted set is analyzed and the top 100 unigram, bigram and trigram terms identified. Two annotators assign each of these terms as relevant or not to UK-EU discussion and the

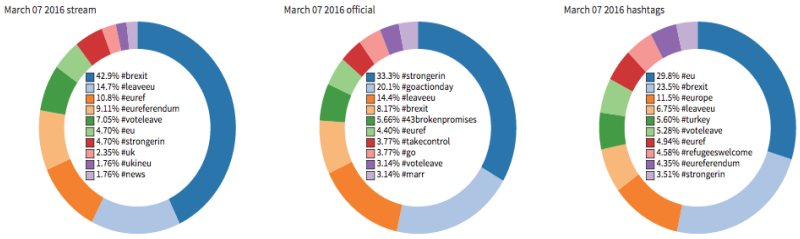

Figure1. The top 10 hashtags in each data set (March 7th 2016)

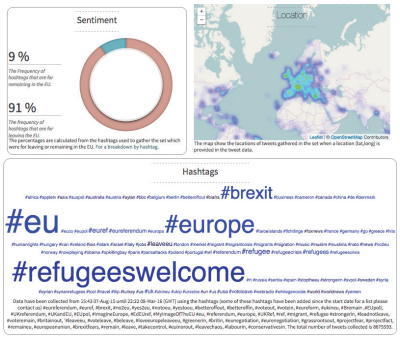

Figure 2. A visualization presenting an overview of the hashtags data set (available at http://goo.gl/fTORDX)

relevant terms are used to search the full data set and to expand the initial set. Again this process is repeated periodically to update the top terms.

The third set is the 'official' set, tweets collected from the official Twitter accounts of the campaign groups, @StrongerIn, @StrongerInPress, @LeaveEUOfficial, @Grassroots_Out and @vote_leave.

The collection of data is an ongoing process and will extend some time beyond the date of the referendum. The number of tweets and users in each set is displayed in Table 1.

	Hashtags	Stream	Official
Tweets	8,916,733	31,106	11,752
Users	1,694,412	16,926	5

Table 1. Tweets and unique users for each set (March 9th 2016)

Analysis and Visualisation

Twitter data can be visualised by the extraction of fields from the underlying data (Tran et al. 2014, Stojanovski et al. 2014). We present a visualisation of some basic information from each of the datasets (fig. 2). This includes a sentiment dial that illustrates the counts of human defined pro-Remain and pro-Leave hashtags, a map that plots the locations extracted from the set, and a wordle showing frequency of hashtags. This data is also visualised through day-by-day illustrations. We are using hashtag frequency to illustrate topics discussed. The different datasets are compared through a visualization of the top 20 hashtags both overall and day-by-day (fig. 1).

We can see from the visualizations that the different datasets do not contain the same hashtags in similar proportions and cannot therefore be used as proxies for each other. There is clearly different bias in each data set so it is important to collect all three data sets to get a better view of the ongoing discussion. The stream and hashtag sets are heavily influenced by the terms used for data collection. Those terms differ greatly when automatically extracted (the stream set) or chosen by experts (the hashtag set). The automatic method is most similar to the official set and is designed to be very specific to the topic.

The expert method is designed to follow a wider variety of terms that the experts expect will become discussion topics over the longer-term referendum debate. The day-by-day visualization shows that tweets from the official set are generally coincident with those in the stream sets of the same day, suggesting that the official campaigns are not influencing the debate. A future direction for this work is to investigate if this relationship can be seen within a smaller time frame such as hour-by-hour. In addition to this we will use this demonstrator to investigate specific terms and multiword terms to track within all three datasets to analyse how discussion is directed.

References

Llewellyn, C.A., Grover, C., Alex, B., Oberlander, J. and Tobin, R. (2015) Extracting a topic specific dataset from a Twitter archive. In TPDL, Poznan, Poland.

Tran, J., Nguyen, Q. V., & Simoff, S. (2014). IntelliViz-A Tool for Visualizing Social Networks with Hashtags. In *Advances in Visual Computing* (pp. 894-903). Springer International Publishing.

Tufekci, Z. (2014). Big Questions for social media big data: Representativeness, validity and other methodological pitfalls. In ICWSM 2014. (pp. 505-514). The AAAI Press.

Stojanovski, D., Dimitrovski, I., & Madjarov, G. (2014). TweetViz: Twitter Data Visualization. *Proceedings of the Data Mining and Data Warehouses*.