Distinguishing the Wood from the Trees: Contrasting Collection Methods to Understand Bias in a Longitudinal Brexit Twitter dataset

Clare Llewellyn and Laura Cram
School of Social and Political Science
University Of Edinburgh, Edinburgh, UK

Abstract

Various methods can be used for searching or streaming Twitter data to gather a sample on a specific topic. All of these methods introduce a bias into the resulting datasets. Here we examine, and try to define, the bias that the different strategies introduce. Understanding the bias means that we can extrapolate wider meaning from the data in a more precise manner.

We use datasets collected on topics from the UK-EU Brexit referendum conducted in 2016. Each dataset discussed draws data from Twitter over a twelve-month period, from 1st September 2015 until 31st August 2016. Three data collection strategies are considered: collecting on human defined topic specific hashtags; collecting using a semi-automated technique to identify topic terms which are then used to collect tweets; and collecting from predefined users known to be tweeting on the topic. To investigate bias in the data we look at, and find wide variation in: group level metadata attributes such as size of the dataset; number of users in each set; average numbers of friends and followers; likely re-tweet status; and levels of inclusion of various add-ons such as hashtags, URLs and media. We also find that relevance to the topic differs between the sets; being far higher in the known users set. We investigate how readability of tweets within each set varies, particularly between known users and topic term sets. We also find that there is a surprising lack of overlap in the data obtained using different collection methods.

Introduction

The huge strength of social media datasets, such as those generated from Twitter, is the sheer volume of spontaneous conversation that can be used to gain insights into social science research questions (Boyd and Crawford 2012). The use of datasets created from social media sources is increasingly being challenged as it is thought there is a lack of population representation and coverage. The robust sampling techniques generally used in this domain are not in effect (Jungherr, Jürgens, and Schoen 2012; Gayo Avello, Metaxas, and Mustafaraj 2011) and bias is introduced.

Generally, within the social sciences, social media research has focused on responses to particular events, limiting analysis to short time frames, very specific search terms (see (Tufekci 2014) for examples and a wider discussion of the issue), specific search terms expanded using lexicons (Olteanu et al. 2014) or specific users (Barberá and Rivero 2014). The methods used to collect data, such as using hashtags as search terms, are thought to strongly influence the results of any subsequent analysis. A common problem has been identified; 'selecting on a dependent variable' (Tufekci 2014), where hashtags or specific search terms are used for collection and those terms are subsequently measured in the analysis. The search terms used are thought to introduce a biased sampling technique. Here we examine, and try to define, the bias that the different strategies introduce. Understanding the bias means that we can extrapolate wider meaning from the data in a more precise manner.

In this paper we compare three commonly used methods for gathering data from Twitter in a longitudinal study on the UK-EU referendum (Brexit). We compare and contrast the resulting data using various extracted metadata components, conduct analyses into the relevance of the data collected to the topic of study, and investigate whether there are differences in the quality of the language in the different sets.

The full Twitter data stream can be purchased but researchers wanting to avoid this cost, can stream a sample for free. There are two API methods that can be used to gather data, a streaming and a search method. These methods are carried out either as the data is produced (the streaming API) or from data which has been previously published (the search API). A random sample can be collected, or, these methods can be used in combination with search terms and other facets to retrieve specific data. If a search query generates a volume of data that is greater than 1% of the total Twitter stream, then the data issued is capped at 1% (Morstatter, Pfeffer, and Liu 2014). There are biases in this sample but the exact nature of these are not known. Morstatter et al (2014) found that, in general, the streaming API does give access to a representative sample of Twitter activity.

A commonly used technique for collecting data from the Twitter API is to query using hashtags (Llewellyn, Cram, and Favero 2016). Each tweet collected has been annotated by the author (or possibly a previous author if the tweet is a re-tweet) using a keyword or compound phrase that suggests a topic label or context. Collecting data in this way can introduce bias into the data sampling procedure. Relevant data will be missed if it does not contain a hashtag. It is likely that there are differences between the type of users who in-

clude hashtags and those that do not. The sample is more likely to contain tweets from those who use hashtags known to the community discussing the issues. Those that use hashtags are thought to be more motivated, actively wanting their content to be seen beyond their immediate network and to become part of the wider conversation. For example, Haung et al (2010) found that using a hashtag increased the likelihood of it being used by Twitter to illustrate a trending topic.

We contrast three methods for collecting data from Twitter: using hashtags chosen by an expert panel as search queries (An et al. 2016); collecting a random sample without specified search terms and extracting appropriate data (Llewellyn et al. 2015); and collecting from specific users that are known to be contributing to the debate (O'Callaghan et al. 2014). We hypothesize that these differing collection strategies result in systematically different datasets and that the users in these datasets use Twitter in different ways. We hypothesize that users of hashtags may be more sophisticated users of Twitter in general, they may use more of Twitter's other facilities such as adding photos and mentioning other users, they may have more followers in general and they may use language in a more sophisticated way.

Data

The datasets described here were collected to enable the study of public conversation on the UK-EU referendum, which took place on June 23rd 2016. In this referendum the citizens of the UK voted on whether the UK should remain within the EU. The result of this vote was in favour of leaving the EU (51.9%). As part of this study, we investigated which topics were related to the debate and how they were discussed. The datasets have been collected in an ongoing manner since August 2015, but this paper focuses on data from a twelve-month period from 1st September 2015 until 31st August 2016. This encompasses discussion leading up to the referendum and subsequent reaction.

We use three methods to collect relevant data. The Hashtag Dataset is gathered from the Twitter Streaming API using UK-EU specific hashtags chosen by a panel of experts. These terms are updated monthly, new hashtags are added in addition to any hashtags previously used. This includes referendum specific terms such as #brexit, #strongerin, #voteleave and those reflecting topics which were likely to be debated, such as #migrants and #refugees, and more general terms such as #eu and #europe.

The second method, the Stream Dataset, aims to reduce the bias introduced through human defined search terms. This set is extracted from the random stream, limited to Tweets in English, and collected through the Twitter API. Data is extracted using broad search terms to gather a custom set of EU referendum related tweets from the overall dataset, a method based on (Llewellyn et al. 2015). The top 100 unigram, bigram and trigram terms are identified, two annotators assign each of the terms as relevant or not to UK-EU discussion, a third annotator mediates this, and the relevant terms are used to search the full dataset and expand the initial set. This process is repeated monthly to update the top terms. Other methods are available for identifying appropriate search terms, for example, using an automatic method for adaptive filtering of tweets using machine learning (Magdy and Elsayed 2016). This would have resulted in a larger dataset that would include tweets that do not contain appropriate hashtags. It is unknown at this point if these methods would have given relevant results at the level required for this study, future work would be required to confirm this. It is hoped that these methods could be used in the future to generate similar, larger, datasets.

The third, Official Dataset, is derived by searching the Twitter API for those users who are known to be contributing to the debate in an approach inspired by O'Callaghan et al (2014). Tweets are collected from the Twitter accounts of the campaign groups, @StrongerIn, @StrongerInPress, @LeaveEUOfficial, @Grassroots_Out and @vote_leave. This data is collected once a day using the Twitter search API. Twitter is commonly used for political campaigning and these accounts provide a reflection of the way in which each campaign sought to frame the debate. It was decided to gather data from a small number of users as the nature of these accounts ensured that most tweets gathered would be relevant to the desired topic.

Analysis

The overall aim of the analysis is to see if there are differences in the datasets gathered using the three different collection methods. We initially look at the data generated by the gathering strategies. We will look at the amount of data generated, the number of users in each dataset and the relevance of the data to the topic studied. We will then look at the overlap between the datasets, how much of the same data is gathered and whether the approaches give rise to systematically different datasets. We hypothesized that these sets would be gathered from different types of users, who use Twitter in significantly different ways. We will seek to measure these differences in tweet strategy by looking at the use of Twitter features such as the ability to add photos, URLs, hashtags, mentions of other users and use of the re-tweeting facility. We will also identify different user types by looking at the number of friends and followers each user in the set has. Finally, we will investigate the quality of the text within the tweets by looking at reading ease and grade level scores.

Size

The Hashtag set is the largest set (over 34 million tweets), followed by the Stream set (over 400K tweets) and the Official set (over 30K tweets) (Tab. 1) Using Twitter's search facility, rather than post filtering, allows us to gather a much larger dataset. It is thought that adaptive filtering techniques may allow us, in the future, to gather a set similar in content to the Stream set, but at a larger scale. Limiting collection to a small number of users also clearly limits the volume of data that is collected. Within the Hashtag and Stream sets most users only appear once (median of 1 in both cases). We can see that the mean value for the Hashtag set (7.27) is higher than the Stream set (1.79). In the Hashtag set we are catching more tweets from some users. When investigated further we found that we are catching many more tweets from the most frequent users in both sets, but more so in the

Table 1: Amount of Tweets and Users in each dataset

Dataset	Tweets	Users	Tweets per User	
			Mean	Median
Official	37,461	5	7492.20	6286
Stream	417,143	233,885	1.79	1
Hashtag	34,405,327	4,748,427	7.27	1

Table 2: Relevance of tweets to the topic

Dataset	Task 1				Task 2			
	A1	A2	A3	Ave	A1	A2	A3	Ave
Official	91	72	85	82.67	94	80	95	89.67
Stream	95	58	83	78.67	96	79	92	89
Hashtag	18	8	24	16.67	49	38	68	51.67

Table 3: Overlap between datasets

Dataset	Tweets	Users
Hashtag and Stream	163,912	194,073
Hashtag and Official	15,396	5
Official and Stream	194	5
Hashtag, Official and Stream	130	5

Table 4: Percentages of tweets that contain additional content

Dataset	Hashtags	URL	Photos	Mentions	Retweets
Official	59.94	42.56	35.65	64.60	47.80
Stream	58.42	49.11	30.35	71.36	59.59
Hashtag	93.16	51.45	34.06	71.68	61.40

Hashtag set. This set is biased to reflect a greater number of tweets from those who tweet frequently, thereby reflecting the views of a very small sub-set of the most highly motivated Twitter users.

Relevance to Topic

Three annotators (A1, A2, A3) evaluated relevance independently over two tasks. We randomly selected 100 tweets from each dataset. In the first task the annotators were asked to decide if each tweet was directly relevant to a debate on the UK-EU referendum. The second task asked if each tweet was relevant to the referendum or about a topic that would likely influence voter opinion. We found that the Official set and the Stream set are more relevant to both the referendum, and to topics relating to the referendum, than the Hashtag set (Tab. 2). The Hashtag set has a low score for 'directly relevant to the referendum debate' but this rises significantly when topics that will influence the debate are considered. The results indicate that, although the Hashtag set contains non-relevant information, it also covers the topics likely to influence voters that are not identified in the other sets. This result was due to the broad range of hashtags given by our expert panel; some experts selected hashtags that were on topics they thought would be relevant to the debate such as #migrant and #refugee. Even given the higher score in task 2, both the Stream and the Official approaches gave data that was more relevant to the topic of study. As hashtags can also be misused, adopting this collection strategy can lead to the introduction of a large amount of irrelevant data.

Overlap

There was a surprisingly small total overlap of all tweets collected; only 130 tweets were common to all sets (Tab. 3). The different collection strategies clearly gave rise to different data. As most users tweet more than once there was a higher overlap in users between sets. All 5 official users were found in all datasets and there were a higher number of common users than common tweets between the Hashtag and Stream sets. We found that 39% of the stream data, and 41% of the official data was found in the Hashtag set whereas 83% of users found in the Stream set, and 100% of users found in the Official set were found in the Hashtag set. This tells us that the large majority of users do use hashtags but not always. There is not a subset of users who always use hashtags and those that do not. Most users include hashtags some of the time or re-tweet tweets that contain hashtags.

Use of Additional Media and Twitter facilities

We investigated whether tweets contained additional content beyond traditional text. In the definition of additional content we included hashtags, URLs, photos or images, mentions, and if they included another tweet (a re-tweet) (Tab. 4). We found, surprisingly, not all of the tweets in the Hashtag set contained a hashtag (93.16%). This was because the API search facility also searches content from a re-tweeted tweet if included. While the tweet we gathered does not contain a hashtag it does contain a re-tweet that does.

We found that tweets from the Hashtag set are more likely to include URLs, mentions and re-tweets than data from the other sets. Collecting using hashtags gives a bias towards richer additional content. Users of hashtags are also more likely to include photos or images than users in the Stream set, but, less likely than the Official set. There is a clear bias towards including tweets with photos or images using the official collection strategy; the official campaign groups use more imagery than other users.

Friends and Followers

Followers are the number of users that follow a particular user and friends are those that a particular user follows. The Official set contains data from official campaign groups that are well known. As they are known they have many more followers than friends (Tab. 5), likely echoed in other user based collection strategies. This collection strategy gives data with a lower number of friends than other approaches. This is likely because the accounts are used primarily to disseminate information from the official groups. As previous work has highlighted, there is a large variation in mean and median scores for each strategy, a small amount of users have many followers. In contrasting the results from the stream and hashtag approaches we can see that the stream approach has a lower median and higher standard deviation for both followers and friends, suggesting there is some variation in the users in the different sets. The result indicates a wider diversity in the number of followers of users found in the Stream set.

Table 5: Followers of users in each dataset

	Followers			Friends		
Dataset	Mean	Median	SD	Mean	Median	SD
Official	32680	17701	30367	963	1034	614
Stream	7052	462	176754	1434	495	5913
Hashtag	5758	491	121638	1461	521	5858

Table 6: Reading ease of tweet text

Dataset	Mean	Median	SD
Official	59.58	21.73	61.67
Stream	60.30	62.34	23.48
Hashtag	57.61	53.83	34.88

Language Use

To study the language use in tweets we measured the Flesch Reading Ease Score, this is scored from 0 (very confusing) to 100 (very easy to read). This is thought to be a rough measure of how easy text is to read and gives an indication of the complexity of the language as measured using the total number of syllables, words and sentences. Mean reading scores were very similar for all sets, but we found the median, and therefore standard deviation, varied widely amongst sets (Tab. 6). We found the lowest median and highest variance of scores in the Official set, this was in general the hardest set to read and contained more complex sentence structures more often. The stream data was the least complex and easiest to read most often with the lowest median and lowest standard deviation. We found that for the hashtag data the language used was more complex than the stream data but less than the official data.

Conclusion

Social media data gathering methods introduce a bias into the resulting data. Here we have tried to define the bias that the different strategies introduce. We tested several hypotheses whether: data collected using different collection strategies results in different datasets and if users in these datasets vary and use the facilities provided by Twitter differently.

We found that the data was different with a surprising lack of overlap between each set. The hashtag approach gives a much larger but less topic relevant dataset and contains more tweets from the same, highly motivated, users. Although there was a low overlap between tweets within the datasets there was a higher overlap between users. Then same users were found even though different tweets were included, this disproves the notion that the sets would contain different types of users and suggests more that it contains tweets generated by the same users using different strategies depending on the context. However, we did find that there was some diversity of users and this gave rise to a difference in friend and follower numbers.

We found a difference in the use of additional content in the different sets, in particular a higher use of photos or images in the Official set and a higher use of additional content in general in the Hashtag set. We found that in collecting from a specific set of users, the Official set, we introduced a bias due to a specific tweeting strategy used by all accounts (using images). This bias may change if different users are collected from, but, highlights the need for caution.

We also found that the users in the Official set had more followers and less friends than the other sets. These users were well known to be tweeting about the topic, which was why we included them, but, also why others followed them. The users in the official group were 'broadcast accounts'; those that have many followers and relatively less friends.

Collecting from specific users also influenced the language used in the tweets; measured through reading ease, they were more complex than those collected from the other sets. We also found that the language in the Hashtag set was more complex than in the Stream set.

References

An, J.; Kwak, H.; Mejova, Y.; Oger, D.; Saenz, S. A.; and Fortes, B. G. 2016. Are you Charlie or Ahmed? Cultural pluralism in Charlie Hebdo response on Twitter. *arXiv preprint arXiv:1603.00646*.

Barberá, P., and Rivero, G. 2014. Understanding the political representativeness of Twitter users. *Social Science Computer Review* 0894439314558836.

Boyd, D., and Crawford, K. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15(5):662–679.

Gayo Avello, D.; Metaxas, P. T.; and Mustafaraj, E. 2011. Limits of electoral predictions using Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence.

Huang, J.; Thornton, K. M.; and Efthimiadis, E. N. 2010. Conversational tagging in Twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, 173–178. ACM.

Jungherr, A.; Jürgens, P.; and Schoen, H. 2012. Why the Pirate Party won the German election of 2009 or the trouble with predictions: A response to predicting elections with Twitter: What 140 characters reveal about political sentiment? *Social science computer review* 30(2):229–234.

Llewellyn, C.; Grover, C.; Alex, B.; Oberlander, J.; and Tobin, R. 2015. Extracting a topic specific dataset from a Twitter archive. In *Research and Advanced Technology for Digital Libraries*. Springer. 364–367.

Llewellyn, C.; Cram, L.; and Favero, A. 2016. Avoiding the drunkard's search: Investigating collection strategies for building a Twitter dataset. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 205–206. ACM.

Magdy, W., and Elsayed, T. 2016. Unsupervised adaptive microblog filtering for broad dynamic topics. *Information Processing & Management* 52(4):513–528.

Morstatter, F.; Pfeffer, J.; and Liu, H. 2014. When is it biased?: assessing the representativeness of Twitter's streaming api. In *Proceedings of the 23rd International Conference on World Wide Web*, 555–556. ACM.

O'Callaghan, D.; Prucha, N.; Greene, D.; Conway, M.; Carthy, J.; and Cunningham, P. 2014. Online social media in the Syria conflict: Encompassing the extremes and the in-betweens. In *ASONAM, 2014 IEEE/ACM*, 409–416. IEEE.

Olteanu, A.; Castillo, C.; Diaz, F.; and Vieweg, S. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*.