Extracting a Topic Specific Dataset from a Twitter Archive

Citation for published version:
Llewellyn, C, Grover, C, Alex, B, Oberlander, J & Tobin, R 2015, Extracting a Topic Specific Dataset from a Twitter Archive. in Research and Advanced Technology for Digital Libraries: 9th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9316, Springer International Publishing, pp. 364-367. https://doi.org/10.1007/978-3-319-24592-8_36

Digital Object Identifier (DOI):
10.1007/978-3-319-24592-8_36

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Research and Advanced Technology for Digital Libraries

Extracting a Topic Specific Dataset from a Twitter Archive

Clare Llewellyn, Claire Grover, Beatrice Alex, Jon Oberlander, and Richard Tobin

School of Informatics, University of Edinburgh, United Kingdom
C.A.Llewellyn@sms.ed.ac.uk

Abstract. Datasets extracted from the microblogging service Twitter are often generated using specific query terms or hashtags. We describe how a dataset produced using the query term 'syria' can be increased in size to include tweets on the topic of Syria that do not contain that query term. We compare three methods for this task, using the top hashtags from the set as search terms, using a hand selected set of hashtags as search terms and using LDA topic modelling to cluster tweets and selecting appropriate clusters. We describe an evaluation method for accessing the relevance and accuracy of the tweets returned.

Keywords: Social Media, Topic Modelling, Data Selection

1 Introduction

This work compares three methods for extracting tweets to form a topic-specific dataset from a Twitter archive. An evaluation method for assessing the relevance of the set produced (to the topic specified) is described and results provided.

Many Twitter datasets are gathered for a particular need. The Twitter streaming API allows the collection of Twitter data at the time it is produced. This means that you need to know what you are looking for ahead of time. As this is not always possible we describe methods for extracting data from a previously harvested Twitter dataset collected from the streaming API. We query this data with specific search terms and then augment the extracted dataset depending on the terms included in the original set. We describe and evaluate three different methods for enriching this dataset, based upon commonly used hashtags, hand selected hashtags and topic relevant tweets as identified by topic modelling. Enriching a dataset brings about a requirement for testing the relevance of this data to the original search parameters. We describe an evaluation technique that can be used to determine the relevance of the data.

2 Background

Twitter provides a streaming API giving access to up to one percent of the data as it is produced. Users can either take a random sample or query using search words, phrases, hashtags, location bounding boxes or user IDs [4]. Previously it was possible to freely share sets of tweets between researchers, but changes to Twitter's terms of service mean that this is now not possible [5]. Instead it is possible to share the user id, tweet id and software for gathering those tweets directly from Twitter. Sharing Twitter data sets allows collaborative research, reproducibility of results and the use of Twitter as a research tool by non-technical researchers. One of the specific aims of the TREC microblog task is to encourage the re-use of Twitter data sets [4].

3 Methods

We did not know which search terms to query the Twitter API with ahead of time. We therefore investigated the best way to extract a topic-specific data set from a previously gathered Twitter archive provided by the ReDites research group [3]. This study investigates the best methods for extracting data relating to the conflict in Syria. Two events are studied and a week's worth of data has been selected associated with each of those events. In the first week, 1-8 March 2012 (2012 set), the UK embassy in Damascus was closed. In the second week, 29 August - 4 September 2013 (2013 set), the UK Parliament voted not to authorise military action over chemical weapons use in Syria. Data was taken from the one percent stream limited to English tweets. We looked to select tweets that discussed Syria and Syria specific events. Initially we gathered all tweets that contained the term 'syria' in either upper or lower case, and we used this as a base set from which we could expand. Methods for increasing the size of the dataset are discussed below.

Method 1: Top Hashtags From the tweets that contained the term 'syria' we extracted all of the hashtags. The top 40 hashtags from both time periods were selected and normalised to give 34. The hashtag terms (hashtags with the # removed) were used as search terms to gather more data. Not all tweets in the set were about the Syrian conflict, for example, the hashtag #UK collected tweets about various activities that were happening in the UK in the selected weeks.

Method 2: Hand Selected Hashtags In order to make the dataset more focused on the Syrian conflict the hashtags about Syria were hand selected. The amount of content on the conflict varies between the datasets with the 2013 dataset being larger. Therefore, all hashtags that had a frequency over 10 from the 2012 set (this gave 60 in total) and all hashtags that occurred with a frequency above 20 in the 2013 set (giving 148 in total) were selected. Each hashtag was annotated by two human coders as either directly relating to the Syrian conflict or not. This included all locations, people and institutions from Syria or formed to deal with Syria or anything with any of those items incorporated into a compound term, for example 'norway4syria'. The human

Dataset	2012			2013				
	Size	1	2	Kappa	Size	1	2	Kappa
Full Set	9988193				11272991			
Top Hashtags	25753	9	8	0.936	231724	14	17	0.886
Hand Selected Hashtags	2555	95	91	0.695	23838	100	100	1.00
Topic Modelling	2292	92	89	0.826	60613	61	57	0.876

Table 1. Total number of tweets returned as relevant per set (size) and the percentages of tweets that were annotated as relevant per set per annotator (1 and 2) and inter-annotator agreement (Kappa)

coders were in perfect agreement (Kappa 1.0) on which tags were related giving 32 which were used as search terms to gather more data.

Method 3: Topic modelling The clustering method used in this work is Latent Dirichlet Allocation (LDA) topic modelling [1]. It is used to identify patterns in text and thereby derive topics. A topic is formed from words that often co-occur, the words that co-occur more frequently across multiple documents are most likely to belong to the same topic. LDA provides a score for each document for each topic. We assign the document to the topic for which it has the highest score. This approach was implemented using the Mallet tool-kit [2]. The system provides a list of top words in each topic. The topics that were classed as relevant for this task were those which have ‘syria’ as one of these most frequent terms. Any tweets that were allocated one of these topics were classed as relevant. In this case the number of clusters generated was set to 15.

4 Results and Discussion

The results are presented in terms of percentage of those that are relevant to the topic, and the F-score. The percentage that are relevant give an overview of the likely pollution of the dataset and the F-score gives an indication of accuracy for each method.

Relevance The percentage of tweets that are relevant was calculated through a manual evaluation. There were 6 datasets: one for each of the 3 methods for each time period. For each set 100 tweets were randomly selected for manual examination. Each tweet was coded as relevant or irrelevant to the conflict in Syrian by two annotators. As can be seen in Table 1, inter-annotator agreement scores for this task show that in general there was high agreement. We can see in Table 2 the relevance of tweets extracted to the topic. The top hashtags approach gives very low relevance results. Therefore, while this approach gave a large data set it was not relevant to the topic. The hand selected hashtags method gives high relevance scores. Therefore, while the sets are fairly small in comparison

2012	Precision	Recall	F-Score
Hand Selected Hashtags	0.92	0.98	0.95
Topic Modelling	0.76	0.80	0.78
2013			
Hand Selected Hashtags	0.95	0.66	0.78
Topic Modelling	0.60	0.90	0.72

Table 2. Accuracy as shown by Precision, Recall and F-Scores per set

with the other approaches they are relevant to the topic. The topic modelling approach provides a high level of relevance for the smaller 2012 set but lower scores and less relevant results for the larger 2013 set.

Accuracy An F-score was calculated by comparing the automatically generated results against a gold standard set. The tweets used to create the gold standard were extracted from the top hashtag set. This gave a set with a higher number of relevant tweets and, therefore, made the accuracy evaluation task difficult and the results more robust. We randomly chose 1000 tweets from each time period which were then annotated as relevant or not. The highest F-score was for the hand selected approach for 2012 set as seen in Table 2. Both approaches showed a drop in F-score for the larger 2013 set. This drop in accuracy was smaller for the topic modelling than the hand selected hashtag approach. This was because while the precision score of the hand selected approach increased for the 2013 set the recall score decreased. The hand selected approach did select appropriate tweets but it also missed many, providing a relevant but small set. The opposite happens for the topic modelling approach. Overall, the F-scores for both datasets are lower but there was a lower drop in accuracy between the two sets. As a drop precision is balanced by a rise in the recall. Therefore while the topic modelling approach selected a larger set it was less relevant.

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
2. A. K. McCallum. *mallet : Amachinelearningforlanguagetoolkit*. 2002.
3. M. Osborne, S. Moran, R. McCreadie, A. Von Lunen, M. D. Sykora, E. Cano, N. Ireson, C. Macdonald, I. Ounis, Y. He, et al. Real-time detection, tracking, and monitoring of automatically discovered events in social media. 2014.
4. I. Soboroff, D. McCullough, J. Lin, C. Macdonald, I. Ounis, and R. McCreadie. Evaluating real-time search over tweets. *Proc. of ICWSM*, 2012.
5. J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011.