Language Technology for Agile Social Media Science

Simon Wibberley
Department of Informatics
University of Sussex
sw206@susx.ac.uk

Jeremy Reffin
Department of Informatics
University of Sussex
j.p.reffin@susx.ac.uk

David Weir
Department of Informatics
University of Sussex
davidw@susx.ac.uk

Abstract

We present an extension of the DUALIST tool that enables social scientists to engage directly with large Twitter datasets. Our approach supports collaborative construction of classifiers and associated gold standard data sets. The tool can be used to build classifier cascades that decomposes tweet streams, and provide analysis of targeted conversations. A central concern is to provide an environment in which social science researchers can rapidly develop an informed sense of what the datasets look like. The intent is that they develop, not only an informed view as to how the data could be fruitfully analysed, but also how feasible it is to analyse it in that way.

1 Introduction

In recent years, automatic social media analysis (SMA) has emerged, not only as a major focus of attention within the academic NLP community, but as an area that is of increasing interest to a variety of business and public sectors organisations. Among the many social media platforms in use today, the one that has received the most attention is Twitter, the second most popular social media network in the world with over 400 million tweets sent each day. The popularity of Twitter as a target of SMA derives from both the public nature of tweets, and the availability of the Twitter API which provides a variety of flexible methods for scraping tweets from the live Twitter stream.

A plethora of social media monitoring platforms now exist, that are mostly concerned with providing product marketing oriented services[1]. For example, brand monitoring services seek to provide companies with an understanding of what is being said about their brands and products, with language processing technology being used to capture relevant comments or conversations and apply some form of sentiment analysis (SA), in order to derive insights into what is being said. This paper forms part of a growing body of work that is attempting to broaden the scope of SMA beyond the realm of product marketing, and into areas of concern to social scientists (Carvalho et al., 2011; Diakopoulos and Shamma, 2010; Gonzalez-Bailon et al., 2010; Marchetti-Bowick and Chambers, 2012; O'Connor et al., 2010; Tumasjan et al., 2011; Tumasjan et al., 2010).

Social media presents an enormous opportunity for the social science research community, constituting a window into what large numbers of people are talking. There are, however, significant obstacles facing social scientists interested in making use of big social media datasets, and it is important for the NLP research community to gain a better understanding as to how language technology can support such explorations.

A key requirement, and the focus of this paper, is agility: the social scientist needs to be able to engage with the data in a way that supports an iterative process, homing in on a way of analysing the data that is likely to produce valuable insight. Given what is typically a rather broad topic as a starting point, there is a need to see what issues related to that topic are being discussed and to what extent. It can be important to get a feeling for the kind of language being used in these discussions, and there is a need to rapidly assess the accuracy of the automated decision making. There is little value in developing an analysis of the data on an approach that relies on the technology making decisions that are so nuanced that the method being used is highly unreliable. As the answers to these questions are being exposed, insights emerge from the data, and it becomes possible for the social scientist to progressively refine the topics that are be-

[1] http://wiki.kenburbary.com/social-media-monitoring-wiki lists 230 Social Media Monitoring Solutions

Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 36–42,
Sofia, Bulgaria, August 8 2013.

ing targetted, and ultimately create a way of automatically analysing the data that is likely to be insightful.

Supporting this agile methodology presents severe challenges from an NLP perspective, where the predominant approaches use classifiers that involve supervised machine learning. The need for substantial quantities of training data, and the detrimental impact on performance that results when applying them to "out-of-domain" data mean that exisiting approaches cannot support the agility that is so important when social scientists engage with big social media datasets.

We describe a tool being developed in collaboration with a team of social scientists to support this agile methodology. We have built a framework based on DUALIST, an active learning tool for building classifiers (Settles, 2011; Settles and Zhu, 2012). This framework provides a way for a group of social scientists to collaboratively engage with a stream of tweets, with a goal of constructing a chain (or cascade) of automatic document classification layers that isolate and analyse targeted conversions on Twitter. Section 4 discusses ways in which the design of our framework is intended to support the agile methodology mentioned above, with particular emphasis on the value of DUALIST's active learning approach, and the crucial role of the collaborative gold standard and model building activities. Section 4.3 discusses additional data processing step that have been introduced to increase the frameworks usefulness, and section 5 introduces some projects to which the framework is being applied.

2 Related Work

Work that focuses on addressing sociological questions with SMA broadly fall into one of three categories.

- Approaches that employ automatic data analysis without tailoring the analysis to the specifics of the situation e.g. (Tumasjan et al., 2010; Tumasjan et al., 2011; O'Connor et al., 2010; Gonzalez-Bailon et al., 2010; Sang and Bos, 2012; Bollen et al., 2011). This body of research involves little or no manual inspection of the data. An analytical technique is selected *a-priori*, applied to the SM stream, and the results from that analysis are then aligned with a real-world phenomenon in order to draw predictive or correlative conclusions about social media. A typical approach is to predict election outcomes by counting mentions of political parties and/or politicians as 'votes' in various ways. Further content analysis is then overlaid, such as sentiment or mood anlysis, in an attempt to improve performance. However the generic language-analysis techniques that are applied lead to little or no gain, often causing adjustments to target question to something with less strict assessment criteria, such as poll trend instead of election outcome (Tumasjan et al., 2010; Tumasjan et al., 2011; O'Connor et al., 2010; Sang and Bos, 2012). This research has been criticised for applying out-of-domain techniques in a 'black box' fashion, and questions have been raised as to how sensitive the results are to parameters chosen (Gayo-Avello, 2012; Jungherr et al., 2012).

- Approaches that employ manual analysis of the data by researchers with a tailored analytical approach (Bermingham and Smeaton, 2011; Castillo et al., 2011).This approach reflects traditional research methods in the social sciences. Through manual annotation effort, researchers engage closely with the data in a manual but interactive fashion, and this effort enables them to uncover patterns in the data and make inferences as to how SM was being used in the context of the sociocultural phenomena under investigation. This research suffers form either being restricted to fairly small datasets.

- Approaches that employ tailored automatic data analysis, using a supervised machine-learning approach(Carvalho et al., 2011; Papacharissi and de Fatima Oliveira, 2012; Meraz and Papacharissi, 2013; Hopkins and King, 2010). This research infers properties of the SM data using statistics from their bespoke machine learning analysis. Mannual annotation effort is required to train the classifiers and is typically applied in a batch process at the commencement of the investigation.

Our work aims to expand this last category, improving the quality of research by capturing more of the insight-provoking engagement with the data seen in more traditional research.

3 DUALIST

Our approach is built around DUALIST (Settles, 2011; Settles and Zhu, 2012), an open-source project designed to enable non-technical analysts to build machine-learning classifiers by annotating documents with just a few minutes of effort.

In Section 4, we discuss various ways in which we have extended DUALIST, including functionality allowing multiple annotators to work in parallel; incorporating functionality to create 'gold-standard' test sets and measure inter-annotator agreement; and supporting on-going performance evaluation against the gold standard during the process of building a classifier. DUALIST provides a graphical interface with which an annotator is able to build a Naïve Bayes' classifier given a collection of unlabelled documents. During the process of building a classifier, the annotator is presented with a selection of documents (in our case tweets) that he/she has an opportunity to label (with one of the class labels), and, for each class, a selection of features (tokens) that the annotator has an opportunity to mark as being strong features for that class.

Active learning is used to select both the documents and the features being presented for annotation. Documents are selected on the basis of those that the current model is most uncertain about (as measured by posterior class entropy), and features are selected for a given class on the basis of those with highest information gain occurring frequently with that class. After a batch of documents and features have been annotated, a revised model is built using both the labelled data and the current model's predictions for the remaining unlabelled data, through the use of the Expectation-Maximization algorithm. This new model is then used as the basis for selecting the set of documents and features that will be presented to the annotator for the next iteration of the model building process. Full details can be found in Settles (2011).

The upshot of this is two-fold: not only can a reasonable model be rapidly created, but the researcher is exposed to an interesting non-uniform sample of the training data. Examples that are relatively easy for the model to classify, i.e. those with low entropy, are ranked lower in the list of unlabelled data awaiting annotation. The effect of this is that the training process facilitates a form of data exploration that exposes the user to the hardest border cases.

4 Extending DUALIST for Social Media Science Research

This section describes ways in which we have extended DUALIST to provide an integrated data exploration tool for social scientists. As outlined in the introduction, our vision is that a team of social scientists will be able to use this tool to collaboratively work towards the construction of a cascade of automatic document classification layers that carve up an incoming Twitter data stream in order to pick out one or more targeted 'conversations', and provide an analysis of what is being discussed in each of these 'conversations'. In what follows, we refer to the social scientists as the researchers and the activity during which the researchers are working towards delivering a useful classifier cascade as data engagement.

4.1 Facilitating data engagement

When embarking on the process of building one of the classifiers in the cascade, researchers bring preconceptions as to the basis for the classification. It is only when engaging with the data that it becomes possible to develop an adequate classification policy. For example, when looking for tweets that express some attitude about a targeted issue, one needs a policy as to how a tweet that shares a link to an opinion piece on that topic without any further comment should be classified. There are a number of ways in which we support the classification policy development process.

- One of the impacts of the active learning approach adopted in DUALIST is that by presenting tweets that the current model is most unsure of, DUALIST will very rapidly expose issues around how to make decisions on boundary cases.

- We have extended DUALIST to allow multiple researchers to build a classifier concurrently. In addition to reducing the time it takes to build classifiers, this fosters a collaborative approach to classification policy development.

- We have added functionality that allows for the collaborative construction of gold standard data sets. Not only does this provide feedback during the model building process as to when performance begins to plateau, but, as a gold standard is being built, researchers are shown the current inter-annotator agreement score, and are shown examples of tweets where there is disagreement among annotators. This constitutes yet another way in which researchers are confronted with the most problematic examples.

4.2 Building classifier cascades

Having considered issues that relate to the construction of an individual classifier, we end this

section by briefly considering issues relating to the classifier cascade. The Twitter API provides basic boolean search functionality that is used to scrape the Twitter stream, producing the input to the cascade. A typical strategy is to select query terms for the boolean search with a view to achieving a reasonably high recall of relevant tweets[2]. An effective choice of query terms that actually achieves this is one of the things that is not well understood in advance, but which we expect to emerge during the data engagement phase. Capturing an input stream that contains a sufficiently large proportion of interesting (relevant) tweets is usually achieved at the expense of precision (the proportion of tweets in the stream being scraped that are relevant). As a result, the first task that is typically undertaken during the data engagement phase involves building a relevancy classifier, to be deployed at the top of the classifier cascade, that is designed to filter out irrelevant tweets from the stream of tweets being scraped.

When building the relevancy classifier, the researchers begin to see how well their preconceptions match the reality of the data stream. It is only through the process of building this classifier that the researchers begin to get a feel for the composition of the relevant data stream. This drives the researcher's conception as to how best to divide up the stream into useful sub-streams, and, as a result, provides the first insights into an appropriate cascade architecture. Our experience is that in many cases, classifiers at upper levels of the cascade are involved in decomposing data streams in useful ways, and classifiers that are lower down in the cascade are designed to measure some facet (e.g. sentiment polarity) of the material on some particular sub-stream.

4.3 Tools for Data Analysis

As social scientists are starting to engage with real-world data using this framework, it has emerged that certain patterns of downstream data analysis are of particular use.

Time series analysis. For many social phenomena, the timing and sequence of social media messages are of critical importance, particularly for a platform such as Twitter. Our framework supports tweet volume analysis across any time frame, allowing researchers to review changes over time in any classifier's input or output tweet flows (classes). This extends the common approach of sentiment tracking over time to tracking over time any attitudinal (or other) response whose essential features can be captured by a classifier of this kind. These class-volume-by-time-interval plots can provide insight into how and when the stream changes in response to external events.

Link analysis. It is becoming apparent that link sharing (attaching a URL to a tweet, typically pointing to a media story) is an important aspect of how information propagates through social media, particularly on Twitter. For example, the meaning of a tweet can sometimes only be discerned by inspecting the link to which it points. We are introducing to the framework automatic expansion of shortened URLs and the ability to inspect link URL contents, allowing researchers to interpret tweets more rapidly and accurately. A combination of link analysis with time series analysis is also providing researchers with insights into how mainstream media stories propagate through society and shape opinion in the social media age.

Language use analysis. Once a classifier has been initially established, the framework analyses the language employed in the input tweets using an information gain (IG) measure. High IG features are those that have occurrence distributions that closely align the document classification distributions; essentially they are highly indicative of the class. This information is proving useful to social science researchers for three purposes. First, it helps identify the words and phrases people employ to convey a particular attitude or opinion in the domain of interest. Second, it can provide information on how the language employed shifts over time, for example as new topics are introduced or external events occur. Third, it can be used to select candidate keywords with which to augment the stream's boolean scraper query. In this last case, however, we need to augment the analysis; many high IG terms make poor scraper terms because they are poorly selective in the more general case (i.e. outside of the context of the existing query-selected sample). We take a sample using the candidate term alone with the search API and estimate the relevancy precision of the scraped tweet sample by passing the tweets through the first-level relevancy classifier. The precision of the

[2] In many cases it is very hard to estimate recall since there is no way to estimate accurately the volume of relevant tweets in the full Twitter stream.

new candidate term can be compared to the precision of existing terms and a decision made.

5 Applications and Extensions

The framework's flexibility enables it to be applied to any task that can be broken down into a series of classification decisions, or indeed where this approach materially assists the social scientist in addressing the issue at hand. In order to explore its application, our framework is being applied to a variety of tasks:

Identifying patterns of usage. People use the same language for different purposes; the framework is proving to be a valuable tool for elucidating these usage patterns and for isolating data sets that illustrate these patterns. As an example, the authors (in collaboration with a team of social scientists) are studying the differing ways in which people employ ethnically and racially sensitive language in conversations on-line. The framework has helped to reveal and isolate a number of distinct patterns of usage.

Tracking changes in opinion over time. Sentiment classifiers trained in one domain perform poorly when applied to another domain, even when the domains are apparently closely related (Pang and Lee, 2008). Traditionally, this has forced a choice between building bespoke classifiers (at significant cost), or using generic sentiment classifiers (which sacrifice performance). The ability to rapidly construct sentiment classifiers that are specifically tuned to the precise domain can significantly increase classifier performance without imposing major additional costs. Moving beyond sentiment, with these bespoke classifiers it is in principle possible to track over time any form of opinion that is reflected in language. In a second study, the authors are (in collaboration with a team of social scientists) building cascades of bespoke classifiers to investigate shifts in citizens' attitudes over time (as expressed in social media) to a range of political and social issues arising across the European Union.

Entity disambiguation. References to individuals are often ambiguous. In the general case, word sense disambiguation is most successfully performed by supervised-learning classifiers (Màrquez et al., 2006), and the low cost of producing classifiers using this framework makes this approach practical for situations where we require repeated high recall, high precision searches of large data sets for a specific entity. As an example, this approach is being employed in the EU attitudinal survey study.

Repeated complex search. In situations where a fixed but complex search needs to be performed repeatedly over a relatively long period of time, then a supervised-learning classifier can be expected both to produce the best results and to be cost-effective in terms of the effort required to train it. The authors have employed this approach in a commercial environment (Lyra et al., 2012), and the ability to train classifiers more quickly with this framework reduces the cost still further and makes this a practical approach in a wider range of circumstances.

With regard to extension of the framework, we have identified a number of avenues for expansion and improvement that will significantly increase its usefulness and applicability to real-world scenarios, and we have recently commenced an 18-month research programme to formalise and extend the framework and its associated methodology for use in social science research[3].

Conclusions and Future Work

We describe an agile analysis framework built around the DUALIST tool designed to support effective exploration of large twitter data sets by social scientists. The functionality of DUALIST has been extended to allow the scraping of tweets through access to the Twitter API, collaborative construction of both gold standard data sets and Naïve Bayes' classifiers, an Information Gain-based method for automatic discovery of new search terms, and support for the construction of classifier cascades. Further extensions currently under development include grouping tweets into threads conversations, and automatic clustering of relevant tweets in order to discover subtopics under discussion.

Acknowledgments

We are grateful to our collaborators at the Centre for the Analysis of social media, Jamie Bartlett and Carl Miller for valuable contributions to this work. We thank the anonymous reviewers for their helpful comments. This work was partially supported by the Open Society Foundation.

[3]Towards a Social Media Science, funded by the UK ESRC National Centre for Research Methods.

References

[Bermingham and Smeaton2011] Adam Bermingham and Alan F Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011*, pages 2–10.

[Bollen et al.2011] Johan Bollen, Alberto Pepe, and Huina Mao. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453.

[Carvalho et al.2011] Paula Carvalho, Luís Sarmento, Jorge Teixeira, and Mário J. Silva. 2011. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, pages 564–568, Stroudsburg, PA, USA.

[Castillo et al.2011] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World wide web*, pages 675–684.

[Diakopoulos and Shamma2010] Nicholas A Diakopoulos and David A Shamma. 2010. Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1195–1198.

[Gayo-Avello2012] Daniel Gayo-Avello. 2012. I wanted to predict elections with twitter and all i got was this lousy paper a balanced survey on election prediction using Twitter data. *arXiv preprint arXiv:1204.6441*.

[Gonzalez-Bailon et al.2010] Sandra Gonzalez-Bailon, Rafael E Banchs, and Andreas Kaltenbrunner. 2010. Emotional reactions and the pulse of public opinion: Measuring the impact of political events on the sentiment of online discussions. *arXiv preprint arXiv:1009.4019*.

[Hopkins and King2010] Daniel J. Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

[Jungherr et al.2012] Andreas Jungherr, Pascal Jürgens, and Harald Schoen. 2012. Why the Pirate Party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, Sprenger, Sander, & Welpe. *Social Science Computer Review*, 30(2):229–234.

[Lyra et al.2012] Matti Lyra, Daoud Clarke, Hamish Morgan, Jeremy Reffin, and David Weir. 2012. Challenges in applying machine learning to media monitoring. In *Proceedings of Thirty-second SGAI International Conference on Artificial Intelligence (AI-2012)*.

[Marchetti-Bowick and Chambers2012] Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: political forecasting with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612.

[Màrquez et al.2006] Lluís Màrquez, Gerard Escudero, David Martínez, and German Rigau. 2006. Supervised corpus-based methods for wsd. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation*, volume 33 of *Text, Speech and Language Technology*, pages 167–216. Springer Netherlands.

[Meraz and Papacharissi2013] Sharon Meraz and Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on #egypt. *The International Journal of Press/Politics*, 18(2):138–166.

[O'Connor et al.2010] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129.

[Pang and Lee2008] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in Information Retrieval*, 2(1-2):1–135.

[Papacharissi and de Fatima Oliveira2012] Zizi Papacharissi and Maria de Fatima Oliveira. 2012. Affective news and networked publics: the rhythms of news storytelling on #egypt. *Journal of Communication*, 62(2):266–282.

[Sang and Bos2012] Erik Tjong Kim Sang and Johan Bos. 2012. Predicting the 2011 dutch senate election results with Twitter. *Proceedings of the European Chapter of the Association for Computational Linguistics 2012*, page 53.

[Settles and Zhu2012] Burr Settles and Xiaojin Zhu. 2012. Behavioral factors in interactive training of text classifiers. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 563–567.

[Settles2011] Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478.

[Tumasjan et al.2010] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fourth international AAAI conference on weblogs and social media*, pages 178–185.

[Tumasjan et al.2011] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2011. Election forecasts with Twitter how 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4):402–418.