Real-time Social Media Analytics through Semantic Annotation and Linked Open Data

Diana Maynard
Dept. of Computer Science
University of Sheffield
Sheffield, UK S1 4DP
d.maynard@sheffield.ac.uk

Mark A. Greenwood
Dept. of Computer Science
University of Sheffield
Sheffield, UK S1 4DP
m.a.greenwood@sheffield.ac.uk

Ian Roberts
Dept. of Computer Science
University of Sheffield
Sheffield, UK S1 4DP
i.roberts@sheffield.ac.uk

George Windsor
Policy and Research
Nesta
London, UK EC4A 1DE
george.windsor@nesta.org.uk

Kalina Bontcheva
Dept. of Computer Science
University of Sheffield
Sheffield, UK S1 4DP
k.bontcheva@sheffield.ac.uk

ABSTRACT

This paper describes an open source framework for analysing large volume social media content, which comprises semantic annotation, Linked Open Data, semantic search, dynamic result aggregation, and information visualisation. In particular, exploratory search and sense-making are supported through information visualisation interfaces, such as co-occurrence matrices, term clouds, treemaps, and choropleths. There is also an interactive semantic search interface (Prospector), where users can save, refine, and analyse the results of semantic search queries over time. These functionalities are presented in more detail in the context of analysing tweets from UK politicians and party candidates in the run up to the 2015 UK general election.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: text analysis

1. INTRODUCTION

Social media is the largest collection of information about society that we have ever had, providing an incredibly rich source of behavioural evidence. However, understanding and using it in a meaningful way is often still a major problem. Gleaning the right information can be tricky because analytics tools either do not provide the right kinds of interpretation, or are simply not accurate, aggregated, enriched or easily interpretable[1]. Our solution to these problems consists of a toolkit for social media monitoring which combines a series of generic tools inside a flexible architecture that allows each component to be easily adapted to the specific social media monitoring task and its domain. In particular, the framework includes semantic analysis, aggregation, and search tools, which allow analysts to dig deep into the data and to perform complex queries which do not just rely on surface information, plus the ability to make interesting correlations between the data.

[1]http://simplymeasured.com/blog/2015/03/09/5-problems-with-how-marketers-use-social-analytics/

WebSci '15 June 28 - July 01, 2015, Oxford, United Kingdom

DOI: http://dx.doi.org/10.1145/2786451.2786500

2. AN OPEN SOURCE FRAMEWORK FOR SOCIAL MEDIA ANALYSIS

The social media analytics toolkit is based around GATE [3], and consists of data collection, semantic annotation, indexing, search and visualisation tasks. The first stage in the process is the data collection, where a number of user accounts and hashtags are followed and their tweets collected via the Twitter streaming API. The tweet stream can also (optionally) be analysed as it comes in, in near real-time, using the "hosebird" client library to handle the connection to the API, with auto reconnection and backoff-and-retry.

GATE now has numerous tools for social media analysis, namely automatic recognition of terms via TermRaider [4], named entities via TwitIE [2], and sentiment analysis [9, 8]. Where appropriate, entities and terms are associated with relevant URIs from Linked Open Data via YODIE [6]. These tools all need adapting to the domain and task for best results, as was done in our experimental use case (Section 3. The framework also integrates Linked Open Data resources (e.g. DBpedia [1], GeoNames), which are accessed via the OWLIM (now GraphDB) knowledge repository [7]. These are used both during semantic annotation and for semantic search and visualisations.

After analysis, the social media posts are indexed using GATE Mimir [10], which enables complex semantic searches to be performed over the entire dataset. Finally, information discovery and visualisation functionalities are provided by GATE Prospector [10], a web-based user interface for searching and visualising correlations in large data sets, and thereby understanding complex content. For example, we can discover and visualise the most frequent topics associated with positive or negative sentiment, or which two topics frequently co-occur in a dynamically selected set of documents.

3. ANALYSIS OF POLITICAL TWEETS

In this section, we give an overview of a practical example of how these open source tools were used to gain insights from a large collection of political tweets, demonstrating both a general analysis of topic and sentiment distribution by different politicians and parties in different regions, and also a specific example of analysing the dataset for understanding engagement of the public with respect to the topic of climate change and the environment.

Our analysis focuses on the interaction of UK members of parliament (MPs), election candidates and members of the public on Twitter. We performed an analysis of recent tweets in the run-up to the 2015 UK elections, according to a set of 42 political themes (topics) such as immigration, climate change, Europe, etc. We first created a collection of approximately 1.8 million tweets, comprising every tweet by any MP or candidate, and every retweet and reply (by anyone) between 24 October 2014 and 13 February 2015. The data was analysed and indexed using the tools described above, customised for the political domain.

A variety of interesting findings emerged about the distribution of tweets, topics and sentiment among different political parties and in different regions. For example, we can query query and visualise a dynamically changing subset of matching tweets in Prospector, to uncover patterns in the data. Analysing the top 20 topics mentioned by MPs in a particular NUTS region returns all tweets authored by MPs representing consituencies from that region. On this dynamically selected subset, Prospector then builds frequency and co-occurrence statistics for the selected topic, and we can watch how these change over time.

Looking specifically at the question of engagement about climate change, a research question we were particularly interested in, we showed that this topic has a high level of engagement by the public, as evidenced by the number of retweets and replies, by the incidence of sentiment and optimism, by tweets containing the mention of another user and by the number of URLs contained in tweets. This research question and the analysis is described more fully in [5].

4. CONCLUSIONS

This paper presented an overview of the GATE-based open source framework for (real-time) analytics of social media, including semantic annotation, search and visualisation components. The framework is independent of the particular application domain, although domain-specific customisations can easily be incorporated through additional content analytics components. Knowledge from Linked Open Data is used to power the semantic searches, as well as as the basis for result aggregation and visualisation. For the latter, we employ both our own information discovery environment (Prospector), as well as web-based visualisations (e.g. choropleths, treemaps), which are generated using the D3 library.

In order to demonstrate the abilities of the framework, a real-life, political science application has been shown. As part of the ForgetIT project, this example scenario will also be extended to cover the House of Commons debates, which will include more information about the political roles MPs fulfil. The aim of this is to investigate the evolution of context in an organizational setting, looking at indicators such as changes to ontologies over time.

5. ACKNOWLEDGMENTS

This work was partially supported by the European Union under grant agreements No. 610829 DecarboNet and 600826 ForgetIt, and by the Nesta-funded Political Futures Tracker project.

6. REFERENCES

[1] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia – a crystallization point for the web of data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 7:154–165, 2009.

[2] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics, 2013.

[3] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854, 02 2013.

[4] E. Demidova, D. Maynard, N. Tahmasebi, Y. Stavrakas, V. Plachouras, J. Hare, D. Dupplaw, and A. Funk. Extraction and Enrichment. Deliverable D3.3, ARCOMEM, 2013.

[5] A. Dietzel and D. Maynard. Climate change: A chance for political re-engagement? 2015.

[6] G. Gorrell, J. Petrak, K. Bontcheva, G. Emerson, and T. Declerck. Multilingual resources and evaluation of knowledge modelling - v2. Technical Report D2.3.2, Trendminer Project Deliverable, 2014.

[7] A. Kiryakov. OWLIM: balancing between scalable repository and light-weight reasoner. In *Proceedings of the 15th International World Wide Web Conference (WWW2006), 23–26 May 2010*, Edinburgh, Scotland, 2006.

[8] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012*, Turkey, 2012.

[9] D. Maynard, G. Gossen, M. Fisichella, and A. Funk. Should I care about your opinion? Detection of opinion interestingness and dynamics in social media. *Journal of Future Internet*, in press.

[10] V. Tablan, K. Bontcheva, I. Roberts, and H. Cunningham. Mímir: an open-source semantic search framework for interactive information seeking and discovery. *Journal of Web Semantics*, 2014.