Ranking election issues through the lens of social media

Stephen Wan and Cécile Paris
CSIRO
Sydney, Australia
firstname.lastname@csiro.au

Abstract

Public events are often accompanied by a social media commentary that documents the public opinion and topics of importance related to these events. In this work, we describe work in collaboration with the State Library of New South Wales (NSW) to archive the social media commentary for the Australian state election in NSW, in March 2015, as a record for social scientists and historians to study in the years to come. Here, we provide an example of how one might utilise this data set, with an analysis of the data focusing on election issues. Specifically, we describe a method to produce rankings of election issues, which we find to correlate moderately to those of official commentators. Furthermore, using our time-series data, we show how the importance of key issues stabilises approximately a month before the actual election.

1 Introduction

The archival of online content for historians and social scientists of the future to study is a challenging problem that has been tackled from various perspectives. For example, in Australia, a conglomerate of state and federal archival institutions have been archiving web content about Australia for many years through the Pandora project[1]. However, projects like Pandora, conceived before the popularisation of social media channels, have only a limited coverage of social data.

We describe work with the State Library of New South Wales (NSW) to address this problem. Specifically, we tackle the collection of social media content for the NSW state election, held in March 2015. Collecting social media content pertinent to major NSW events is part of the libary's operations, complementing data archived through projects like Pandora. As part of this mandate, the library collected physical and ephemeral materials associated with the election, such as electronic version of election campaign materials as well as public discussions on social media.[2] To collect the latter, the library employed our social media monitoring tool, Vizie (Wan and Paris, 2014), to archive public discussions on Twitter[3], a predominant social media platform, that were authored by either the community or the election candidates.

In this paper, we explore the utility of such a data resource, which is intended to support the scholarly investigations of future researchers, such as social scientists and journalists. One could ask, how accurate would a picture of the election based on this data be? To address this, we present an analysis focusing on one aspect of the election, that of election issues.

We hypothesise that social media data can shed light on which issues were the most prevalent in the lead up to the elections. Specifically, for some given election issues, we explore the use of the data to produce a ranking of the issues. Our preliminary investigation focuses on obtaining these rankings based on news content shared as embedded links on Twitter. Our results show that our data-derived rankings have a moderate correlation to those eventually published in official election commentaries. In addition, utilising the time-series nature of our data, we highlight how the rankings of these issues stabilises in time, indeed weeks before the official commentary is released.

In the remainder of the paper, in Section 2, we outline related work. We describe the data collection process in 3. Section 4 describes our col-

[1] http://pandora.nla.gov.au

[2] This effort is described in: http://www.abc.net.au/news/2015-03-12/election-tweets-added-to-nsw-library-election-collection/6306490.

[3] www.twitter.com

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 48–52,
Beijing, China, July 30, 2015.

lection of ground truth data. We describe our approach for ranking election issues in Section 5. In Section 6, we present our analyses on election issues, which we then discuss in Section 7. Finally, we summarise our findings in Section 8.

2 Related Work

There has been much work in using Twitter to predict the outcome of an election e.g., (O'Connor et al., 2010), as well as critiques of such approaches (Gayo-Avello et al., 2011) and explorations of sentiment for prediction (Tumasjan et al., 2010).

Our work focuses on different types of media, specifically news and Twitter data. There are several investigations of media which take into account the diversity of platforms and data types. For example, some have examined the effect of different information sources on public discussion, e.g., (Scharl and Weichselbraun, 2006) and (Ahmad et al., 2011). (Declerck, 2013) mentions that it would be interesting to characterise the public discussion topics for an election. In this work, we assume that these topics are provided a priori and show how a ranking of election topics is possible.

Further afield from election-focused research, (Liu et al., 2011) also utilise embedded links in Twitter but for the purposes of generating summaries of events (see also (Nichols et al., 2012) and (van Oorschot et al., 2012)). Here, we examine how our ranking of issues based on embedded links compares with that of an official commentary, rather than generating event summaries.

3 Data Collection

Data is collected using our Vizie tool which provides an interface for configuring queries to be used with a number of social media platforms including Twitter and Facebook[4], amongst others (Wan and Paris, 2014). In this paper, we focus on Twitter content, which we collect via the free Twitter API[5]. Adherence to rate limits are observed, but for most queries we do not lose any data as a result of quota limitations.

The queries about candidates and parties were prepared by the library staff in advance of the election, using a query curation framework. (For an example of their social media collection framework for all public events in 2014, see (Barwick et al., 2014).) Some candidates, such as incumbents running for election again, were known ahead of time. Other candidates were added to the query list when the official candidate list was released by the Australian Electoral Commission, approximately two weeks before the election. This was the last date to register as an election candidate.

The full set of queries included candidate names specified as multi-word phrases, along with contextual query terms such as the party name or electorate. For example, for the candidate "Luke Foley", a "Labor" party candidate running for the seat of "Auburn", we had three queries, consisting of the different possible pairings of these three elements. Each query was sent to the Twitter API. Query terms also included known election issues, electorate names and party names. Library staff were able to use the tool to set up geographical filters based on time zones to exclude non-Australian content if the query was general enough to collect content from other parts of the world. Finally, Twitter accounts for candidates and parties were subscribed to, where these existed.

For all Twitter content collected, each tweet was automatically checked for an embedded URL. If one was found, the destination web content was retrieved and archived, along with a link to the tweet that referenced it.

4 Ground Truth Data

To obtain election issues, we use a number of different online commentaries about the election. These sources were: (1) news articles from prominent news companies[6] [7]; (2) issues extracted from a Vote Compass[8] questionnaire by the Vox Populi company; and (3) Wikipedia[9].

For our ground truth on a ranking of these issues, we used a ranking published in a news article which reported the results of the Vote Compass questionnaire.[10] This ranking is reproduced in Table 1. Interestingly, not all sources had the same set of issues. We used the Vote Compass issues as the canonical set as this was the largest set with a considerable overlap with commentaries by other news agencies.

[4] www.facebook.com

[5] dev.twitter.com

[6] http://www.abc.net.au/news/2015-03-07/seven-key-things-to-watch-during-the-nsw-election-campaign/6283582

[7] http://www.smh.com.au/nsw/nsw-state-election-2015

[8] http://www.abc.net.au/votecompass/

[9] page: New_South_Wales_state_election,_2015

[10] http://www.abc.net.au/news/2015-03-05/nsw-election-2015-vote-compass-issues-economy-asset-sales/6280030

#	Issue	#	Issue
1	Economy	13	Poverty
2	Asset sales	14	Housing
3	Cost of living	15	Taxation
4	Education	16	Defence
5	Environment	17	Population
6	Healthcare	18	Racism
7	Corruption	19	Petrol pr ices
8	Public transport	20	Drug abuse
9	Unemployment	21	Indigenous issues
10	Roads	22	Personal debt
11	Immigration	23	Drought relief
12	Crime		

Table 1: Ranked issues from Vote Compass.

Rank	Issue	#articles
1	Environment	137
2	Corruption	69
3	Leadership	60
4	Asset sales	56
5	Healthcare	42
6	Roads	34
7	Social Services, Education, Domestic violence	30
8	Prime Minister	22
9	Public transport	20
10	Crime	19
11	Balance of Power	8
12	Swing back	5
13	Indigenous issues	4
14	Defence, Drought relief, Personal debt, Poverty, Unemployment	1

Table 2: Ranked issues. Ranking is based on the number of news articles associated with that issue.

5 Generating a ranking of election issues

Our aim was to see what news articles shared on social media can reveal about the relative importance of different election issues. As such, we associated each article with an issue, using the simplifying assumption of one issue per article. This then allowed us to generate a ranking of election issues based on the number of shared news articles tagged with that issue.

To begin with, we retrieved the shared news articles from our database with publication dates falling between 12 Dec. 2014 to 27 Mar. 2015, a day before the election date. Due to limited computing resources, we limited our analysis to the top 1000 articles, ranked by the number of times it was shared in a tweet using an embedded URL.

Each of the 1000 articles was associated with an election issue using standard vector space methods —for an overview, see (Salton and McGill, 1983). Each issue was represented as a vector of word frequencies, and the closest matching issue to an article was determined using cosine similarity.

To derive our issue vectors, we used text describing each issue from our the sources listed in Section 4. Although each source used slightly different names for elections issues, these were trivially reconciled with the issues provided by Vote Compass. As an example, the gloss for the issue *"asset sales"* included text such as, *"Asset sales. New South Wales should lease its electricity transmission network to the private sector. To cover infrastructure costs the government should privatise public assets rather than raise taxes."*[11] Glosses from different sources were then merged to form a single gloss for each issue.

To avoid spurious associations between articles and issues, we processed the glosses to ensure that they represented the core elements of an issue. This was done by removing words from one of three categories of words lists: i) stopwords, ii) words belonging to multiple issues, and iii) words refering to elections in general.

For (ii), we removed words occurring in more than one gloss. For example, *"taxes"* in the gloss for "asset sales" also occurs in the gloss for "taxation" and is thus not deemed to be indicative of any one particular issue.

For (iii), we determined words to do with the general topic of elections in Australia by mining specific Wikipedia pages. Words were obtained from the first paragraph of the "NSW 2015 election" Wikipedia page, and from the first two sections ("Federal Parliament" and "Voting") of the Wikipedia page on "Elections in Australia". The intuition is that by removing words about elections in general, the inferred link between an article and an issue will be more acccurate.

We use the remaining words in the glosses to produce the vector space representations of each election issue. We normalised words to be in lowercase, and all non-alphabetic characters, aside from whitespace, were removed. The vector was weighted using term frequency.

Each of these news articles was then compared to each ground truth issue based on a comparison between a vector for the news article and a vector for the election issue. For this, words in the article's title were processed in a similar manner to the glosses. We then counted the number of articles associated with each issue and then ranked issues by this count. Table 2 shows the resulting ranked issues. We note that not all ground truth

[11] Text from http://www.abc.net.au/votecompass/

election issues are represented in our data set.

6 From Social Data to Election Insights

6.1 Comparing rankings of election issues

We compare the common elements of the ranked list in Table 2 with that of Table 1 using Kendall tau Rank Correlation (Kendall, 1938).[12] We find a tau value of 0.55 (2-sided p $= 0.047$), which is statistically significant at $\alpha = 0.05$. For this test, we omitted the items at rank 14 as these were found only once in the data and may be spurious matches. Including them would inflate tau and make the result significant at $\alpha = 0.01$.

We find this moderate correlation encouraging. However, we note that the simplicity of our method for labelling election issues may be one reason that we do not find a stronger correlation. In future work, we will explore whether supervised machine learning methods for assigning labels can help improve our correlation.

6.2 Ranking stabilisation across time

A key feature of our data set is that it is time-series data and, in some future application, one could conceivably show rankings of issues before any official commentary emerges. For such a system, we would assume that it has a generic election issue detector (perhaps based on a text classification method such as labelled LDA methods (Ramage et al., 2010)). To explore this further, we repeat the study in Section 6.1 so that the end date is set at weekly intervals starting in January 2015, using our ground truth election issues.

Figure 1 shows the probability of the null hypothesis; that there is no correlation when comparing the ground truth Vote Compass ranking to the data-derived rankings each interval (Kendall's tau$= 0$). For each probability, the tau value is shown in the upper curve. We see that the p-value drops below $\alpha = 0.05$ around Feb. 23rd. This accords with our intuition: there is more uncertainty about the election issues early in the election period, and so the data-driven rankings fluctuate more. We note that our gold standard article with the ranked issues was published on Mar. 5th.

7 Discussion

With statistically significant correlation between the rankings, we conclude that Twitter shared news content about an election can provide insights on the importance of election issues. As an added advantage, our approach can also rank issues that were not mentioned by Vote Compass but which were described in our other sources.[13] These issues concern politics and government, whereas the Vote Compass issues are societal.

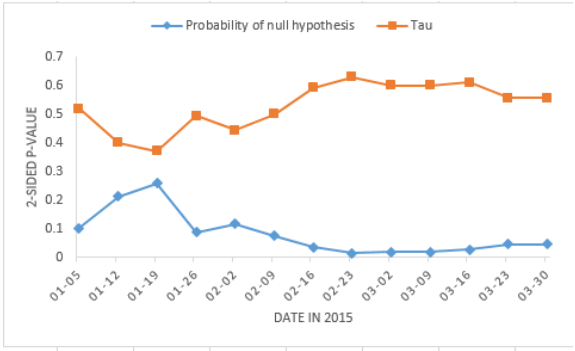

Figure 1: Kendall's tau (and the associated 2-sided p-value for significance testing) for ranked issues at weekly intervals.

We note that the analysis described here is susceptible to campaigning and lobbying activity. We are unable to tell from this analysis whether the prevalence of an issue is due to intensive lobbying or a reflection of widespread concern.

8 Conclusion

In this work, we presented an analysis which provided a ranking of election issues based on shared news articles found in Twitter content about the 2015 NSW state election. With respect to the issues that found a voice on Twitter, we observed a moderate correlation with official commentaries. Furthermore, utilising the time-series nature of our data set, we show when the ranking of the election issues seems to stabilise during the election period, suggesting the potential for this analysis to provide some monitoring functionality.

Acknowledgments

We thank Brendan Somes and Kathryn Barwick, who curated queries and oversaw the data collection process at the State Library of NSW; Brian Jin and James McHugh from the CSIRO for their software engineering expertise; and the anonymous reviewers for their insights on improving the readability of this paper.

[12] Kendall's tau was calculated with the online tool: http://www.wessa.net/rwasp_kendall.wasp (Wessa, 2012)

[13] These issues were Leadership, Prime Minister, Swing back, and Balance of Power issues.

References

Khurshid Ahmad, Nicholas Daly, and Vanessa Liston. 2011. What is new? news media, general elections, sentiment, and named entities. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 80–88, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Kathryn Barwick, Mylee Joseph, Cécile Paris, and Stephen Wan. 2014. Hunters and collectors: seeking social media content for cultural heritage collections. In *VALA 2014: Streaming With Possibilities*.

Thierry Declerck. 2013. Integration of the thesaurus for the social sciences (thesoz) in an information extraction system. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 90–95, Sofia, Bulgaria, August. Association for Computational Linguistics.

Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. 2011. Limits of electoral predictions using twitter. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*.

Fei Liu, Yang Liu, and Fuliang Weng. 2011. Why is "sxsw" trending? exploring multiple text sources for twitter topic summarization. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 66–75, Portland, Oregon, June. Association for Computational Linguistics.

Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pages 189–198, New York, NY, USA. ACM.

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 William W. Cohen and Samuel Gosling, editors, *ICWSM*. The AAAI Press.

Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. In *ICWSM*.

G. Salton and M. J. McGill. 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York.

Arno Scharl and Albert Weichselbraun. 2006. Web coverage of the 2004 us presidential election. In *Proceedings of the 2Nd International Workshop on Web As Corpus*, WAC '06, pages 35–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.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.

Guido van Oorschot, Marieke van Erp, and Chris Dijkshoorn. 2012. Automatic extraction of soccer game events from twitter. In Marieke van Erp, Laura Hollink, Willem Robert van Hage, Raphal Troncy, and David A. Shamma, editors, *Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, volume 902, pages 21–30, Boston, USA, 11. CEUR.

Stephen Wan and Cécile Paris. 2014. Improving government services with social media feedback. In *IUI'14 19th International Conference on Intelligent User Interfaces, IUI'14, Haifa, Israel, February 24-27, 2014*, pages 27–36.

Wessa. 2012. Kendall tau rank correlation (v1.0.11) in free statistics software (v1.1.23-r7).