Selected Papers of AoIR 2016:
The 17th Annual Conference of the
Association of Internet Researchers
Berlin, Germany / 5-8 October 2016

A TOPIC ANALYSIS APPROACH TO REVEALING DISCUSSIONS ON THE AUSTRALIAN TWITTERSPHERE

Brenda Moon
Queensland University of Technology

Abstract

This paper investigates techniques to identify the topics being discussed in one week of tweets from the Australian Twittersphere. Tweets were extracted from a comprehensive dataset which captures all tweets by 2.8m Australian: the Tracking Infrastructure for Social Media Analysis (TrISMA) (Bruns, Burgess & Banks et al., 2016). Bruns & Moe (2014) suggest that most Twitter research to date has focussed on "the macro layer of Twitter communication" (p. 23-24), partly because it is methodologically difficult to move beyond this. The TrISMA dataset enables the selection of a dataset based on a date range, rather than being limited to keywords or hashtags. As a result, the extracted one-week dataset of 5.5 million tweets is not focussed on a particular topic, and contains tweets from all three layers of Twitter communication defined by Bruns & Moe (2014), not just predominately from the macro level of hashtag conversations. This study seeks to identify the themes present in this dataset using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003).
The results of the topic analysis are triangulated with the themes found by the different types of analysis as part of a wider methodological study determining other metrics for the same week. The ability to identify the themes present in a dataset has many applications, including identifying changes in themes over time, extracting subsets of the corpus for further study, and understanding the diversity of themes present.

Keywords: Twitter, social media, big data, topic analysis, Latent Dirichlet Allocation

Introduction

This paper investigates techniques to identify the topics being discussed in one week of tweets from the Australian Twittersphere. Tweets posted in the week from Sunday 2 August to Saturday 8 August 2015 were extracted from a comprehensive dataset which captures all tweets by 2.8m Australian users on a continuing basis; The Tracking Infrastructure for Social Media Analysis (TrISMA) infrastructure (Bruns, Burgess & Banks et al., 2016). This resulted in a dataset of 5.5 million tweets. The selected week was chosen as part of a wider study looking at other metrics for the same week,

Suggested Citation (APA): Moon, Brenda. (2016, October 5-8). *A Topic Analysis approach to revealing discussions on the Australian Twittersphere*. Paper presented at AoIR 2016: The 17th Annual Meeting of the Association of Internet Researchers. Berlin, Germany: AoIR. Retrieved from http://spir.aoir.org.

including periodic patterns in tweets per day and a detailed @mention/retweet network for one day during the week. Having these other studies for the same dataset enables comparison of the results obtained by the different methods.

Bruns & Moe, 2014 suggest that most Twitter research has focussed on "the macro layer of Twitter communication: on the engagement with breaking news and other topics by participants in hashtag audiences" (p. 23-24), partly because it is methodologically difficult to move beyond this. The TrISMA dataset enables the selection of a dataset based on a date range, rather than being limited to keywords or hashtags. As a result, the dataset has not been focussed on a particular topic by the use of a keyword or hashtag, and contains tweets from all three layers of Twitter communication defined by Bruns & Moe (2014), not predominately the macro level of hashtag conversations.

Latent Dirichlet Allocation (LDA) is an unsupervised machine learning approach that identifies topics present in a text (Blei, Ng, and Jordan, 2003). Each text is considered to be a bag-of-words, that is only the frequency of the words in the text is used, not the word order or meaning of the words. A LDA topic is expressed as a Bayesian probability distribution across all the words in the corpus, with the contribution of each word towards that topic. Blei et al. (2003) defined LDA as:

> Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. (p. 996)

If the model is successful, the most important words in each topic should appear coherent to a human reader, "but the top few words in a topic only give a small sense of the thousands of the words that constitute the whole probability distribution" (Schmidt, 2012, p. 51). Each document in a corpus can be interpreted as having a mixture of topics which allows the identification of both the most significant documents for a topic and the prevalence of a topic in the corpus.

One measure used for the accuracy of LDA models is perplexity (Blei et al., 2003; J. Chang et al., 2009) and although it can be useful for comparing models or parameter options, it may not be an accurate measure of the quality of the topics, "there is no technical reason to suppose that held-out accuracy corresponds to better organization or easier interpretation" (Blei, 2012, p. 83). An alternative approach by J. Chang et al. (2009) uses human evaluation tasks assessing 'word intrusion' and 'topic intrusion' into the LDA identified topics as a way of measuring their coherence for a human reader, and this is the approach I have chosen to apply. I manually checked the highest likelihood words in each topic to see if they formed a recognisable, coherent topic instead of appearing random, and then check the topics that the model assigns to unseen documents (held out from training) by inspecting the documents to see what topics appeared in each tweet. I also investigate applying discrepancy functions for LDA developed by Mimno and Blei (2011) that "measure how well its statistical assumptions about the topics are matched in the observed corpus and inferred topics" (Mimno & Blei, 2011, p. 228).

Blei and Lafferty (2009) warn us about the interpretation of topic models:

> The topics and topical decomposition found with LDA and other topic models are not “definitive”. Fitting a topic model to a collection will yield patterns within the corpus whether or not they are “naturally” there. (And starting the procedure from a different place will yield different patterns!) Rather, topic models are a useful exploratory tool. (p. 17).

One of the most difficult decisions in LDA is the number of topics to select “choosing the number of topics is a persistent problem in topic modeling and other latent variable analysis” (Blei & Lafferty, 2009, p. 11). With too few topics, each topic will be broad and the most frequent words may not appear coherent, but conversely “as topics become more fine-grained in models with larger number of topics, they are less useful for humans” (J. Chang et al., 2009, p. 4).

The results of the topic analysis will be compared to the topics found with using bigram analysis in the periodic pattern study and the discussion clusters identified in the network analysis of the Wednesday tweets. By triangulating the results found by the different types of analysis we can get a better understanding of how each approach informs our understanding of the dataset.

References

Blei, D. M. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, *2*(1), 8–11.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (chap. Topic mode). Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Blei, D. M., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bruns, A., Burgess, J., Banks, J., Tjondronegoro, D., Dreiling, A., Hartley, J., … Sadkowsky, T. (2016). TrISMA: Tracking Infrastructure for Social Media Analysis. Retrieved from http://trisma.org/

Bruns, A., Burgess, J., & Highfield, T. (2014). A “Big Data” Approach to Mapping the Australian Twittersphere. In P.L. Arthur & K. Bode (Eds.), *Advancing Digital Humanities: Research, Methods, Theories* (pp. 113–129). Houndmills: Palgrave Macmillan.

Bruns, A., Moe, H. (2014). Structural Layers of Communication on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann, (Eds.), *Twitter and Society*. New York: Peter Lang.

Burgess, J., & Bruns, A. (2015). Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research after the Computational Turn. In G. Langlois, J. Redden, & G. Elmer (Eds.), *Compromised Data: From Social Media to Big Data* (pp. 93–111). New York: Bloomsbury Academic.

Chang, J., Gerrish, S., Boyd-Graber, J., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural information processing systems.* Vancouver, British Columbia.

Mimno, D., & Blei, D. (2011). Bayesian Checking for Topic Models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 227–237).

Rambukkana, N. (Ed.). (2015). *Hashtag Publics: The Power and Politics of Discursive Networks*. New York: Peter Lang.

Schmidt, B. M. (2012). Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities*, *2*(1), 49–65.