Social Media: A Microscope for Public Discourse

Leon Derczynski
Natural Language Processing Group
Department of Computer Science
University of Sheffield
leon@dcs.shef.ac.uk

Abstract

This abstract contains a summary and further reading to accompany the talk "Social Media: A Microscope for Public Discourse", given at the Digital Humanities Congress 2014.

Social media can be seen as a digital sample of all human discourse. We discuss the idiosyncracies and potential of this communication medium and present a mature software toolkit for social media study.

Although superficially social media can look like a seething tide of trivia, these seven hundred million openly-published daily messages have been shown to be rich in structured, salient signals. One can observe how relationships and groups form and dissipate in social groups. Displays of affect, social class, and tribe are frequently evident through choice of language (Hu et al., 2013). Reactions and attitudes towards events, movements and political ideas can be captured and recorded. Additionally, longitudinal analysis provides historical records for retrospective studies.

Much work in digital analysis of text has focused on news articles. While well-formed, covering a reasonably broad selection of topics, and often accurate, such text tends to have been written by middle-class working-age white men - leading to a narrowing of styles and ideas discussed (Eisenstein, 2013). This has serious knock-on effects: for example, language-based interactions with automated systems are more likely to be successful if one uses the language style the system was trained on. In contrast, Twitter - the model social media platform (Tufekci, 2014) - contains a broad and diverse range of authors. We discuss this bias and how it challenges digital tools.

As human knowledge expressed in language, the field of computational linguistics - the understanding and study of language using computational techniques - offers exciting prospects for the digital humanities. We finally introduce state-of-the-art, open software which focuses on processing and extracting information from Twitter (Bontcheva et al., 2013), and is part of the GATE language processing system developed at the University of Sheffield.

References

K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani. 2013. 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.

J. Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Y. Hu, K. Talamadupula, S. Kambhampati, et al. 2013. Dude, srsly?: The surprisingly formal nature of twitters language. *Proceedings of ICWSM*.

Z. Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.