Social Media: Friend or Foe of Natural Language Processing?

Timothy Baldwin
The University of Melbourne, VIC 3010, Australia
tb@ldwin.net

Abstract

In this talk, I will outline some of the myriad of challenges and opportunities that social media offer for natural language processing. I will present analysis of how pre-processing can be used to make social media data more amenable to natural language processing, and review a selection of tasks which attempt to harness the considerable potential of different social media services.

There is no question that social media are fantastically popular and varied in form — ranging from user forums, to microblogs such as Twitter, to social networking sites such as Facebook — and that much of the content they host is in the form of natural language. This would suggest a myriad of opportunities for natural language processing (NLP), and yet much of the applied research on social media which uses language data is based on superficial analysis, often in the form of simple keyword search. This begs the question: Are NLP methods not suited to social media analysis? Conversely, is social media data too challenging for modern-day NLP? Alternatively, are simple term search-based methods sufficient for social media analysis, i.e. is NLP *overkill* for social media? In exploring these questions, I attempt to answer the overarching question of whether social media data is the friend or foe of NLP.

I approach the question first from the perspective of what challenges social media language poses for NLP. The most immediate answer is the infamously free-form nature of language in social media, encompassing spelling inconsistencies, the free-form adoption of new terms, and regular violations of English grammar norms. Unsurprisingly, when NLP tools are applied directly to social media data, the results tend to be miserable when compared to data sets such as the Wall Street Journal component of the Penn Treebank. However, there have been recent successes in adapting parsers and POS taggers to social media data (Foster et al., 2011; Gimpel et al., 2011). Additionally, lexical normalisation and other preprocessing strategies have been shown to enhance the performance of NLP tools over social media data (Lui and Baldwin, 2012; Han et al., to appear). Furthermore, social media posts tend to be short and the content highly varied, meaning it is difficult to adapt a tool to the domain, or harness textual context to disambiguate the content. There is also the engineering challenge of real-time processing of the text stream, as much of NLP research is carried out offline with only secondary concern for throughput. As such, we might conclude that social media data is a foe of NLP, in that it challenges traditional assumptions made in NLP research on the nature of the target text and the requirements for real-time responsiveness.

However, if we look beyond the immediate text content of social media, we quickly realise that there are various non-textual data sources that can be used to enhance the robustness and accuracy of NLP models, in a way which is not possible with static text corpora. For example, simple information on the author of a post can be used to develop author-adapted models based on the previous posts of the same individual (at least for users who post sufficiently large volumes of data). Links in the post can be used to disambiguate the textual content of the post, whether in the form of URLs and the content contained in the target document(s), hashtags and the content of other similarly-tagged posts, thread-

26th Pacific Asia Conference on Language,Information and Computation pages 58–59

ing structure in web user forums, or addressee information and the content of posts from that individual. Simple timestamp information may provide insights into what timezone the user is likely to be based in, allowing for adjustment of language priors for use in language identification. User-declared metadata may also provide valuable information on the probable interpretation of a given post, e.g. knowing that a person is from Australia may allow for adjustment of lexical or word–POS priors. Multimodal content such as images or videos included in the post may also provide valuable insights into the likely interpretation for particular words. Social network information may also allow for user-specific adjustment of language priors of various types. In this sense, the rich context that permeates social media can very much be the friend of NLP, in providing valuable assistance in disambiguating content.

Turning to the question of why the majority of social media analysis makes use of simple language analysis such as word counts for a canned set of query terms, I suggest that the cause is largely because of the constraints imposed on the user by different social media APIs, and also the relative accessibility of such simple techniques, as compared to full-strength NLP. I go on to claim that "the tail has been wagging the dog" in social media research, in the sense that while impressive results have been achieved for particular application types, the choice of application has been constrained by what is achievable with relatively simple keyword analysis. For example, searching for keywords relating to earthquakes or influenza allows for impressive results to be achieved in earthquake detection or influenza outbreak analysis (Sakaki et al., 2010; Ritterman et al., 2009). However, this style of approach presupposes a highly-constrained, predetermined information need which is expressible in a small number of relatively unambiguous query terms. In applications such as trend analysis, the information need is more open-ended and it is unreasonable to expect that a static set of keywords will capture new trends. Even for highly-constrained information needs, there may not be a high-precision set of query terms which provide the necessary information. While it is certainly not the case that full-blown NLP is needed in all social media applications, it is equally not correct to say that NLP is overkill for all social media analysis. Rather, the emergence of more mature, robust NLP technologies tailored to social media data will enable new opportunities for social media analysis, earning new friends for NLP in the process.

References

Jennifer Foster, Ozlem Cetinoglu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proc. of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 893–901, Chiang Mai, Thailand.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 42–47, Portland, USA.

Bo Han, Paul Cook, and Timothy Baldwin. to appear. Automatically constructing a normalisation dictionary for microblogs. *ACM Transactions on Intelligent Systems and Technology*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.

Joshua Ritterman, Miles Osborne, and Ewan Klein. 2009. Using prediction markets and Twitter to predict a swine flu pandemic. In *Proceedings of the 1st International Workshop on Mining Social Media*, November.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. of the 19th International Conference on the World Wide Web (WWW 2010)*, pages 851–860, Raleigh, USA.