Natural Language Processing for Social Media

Atefeh Farzindar and Diana Inkpen
(NLP Technologies Inc., Université de Montréal; and University of Ottawa)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 30), 2015, xix+146 pp; paperbound, ISBN 978-1-62705-388-4; ebook, ISBN 978-1-62705-389-1; doi 10.2200/S00659ED1V01Y201508HLT030, $55.00

Reviewed by
Annie Louis
University of Essex

Today, social media refers to a wide range of Web sites and Internet-based services that allow users to create content and interact with other users. Some of these tools, such as multi-party chats, discussion forums, blogs, and online reviews, have been a focus of natural language processing (NLP) research for quite some time now. But within the last decade, NLP work has expanded rapidly to cover an immense variety of new social media content—microblogs such as Twitter, social networks such as Facebook, comments on news articles, captions on user-contributed images such as on Flickr, and forums dedicated to specialized topics and needs (e.g., health and online education). Simultaneously, many other research communities are carrying out work using social media data—information science, information retrieval, network science, social media analytics, social science, psychology, and corpus linguistics. Today, a large number of businesses are also centered on, or benefit from, analytics performed on social media. Given these myriad research and commercial interests in the social media domain, we are at a time where we should seek to clearly understand what role NLP has in the field of social media analysis, both in terms of the key and interesting language questions, as well as contributions NLP can make to the research carried out in other fields. In this context, this short book by Farzindar and Inkpen is timely and exciting, filling an obvious gap.

This book is targeted towards researchers who have a background in natural language processing and machine learning, and want to learn about research questions, tasks, and techniques related to processing of social media texts. Chapter 1 introduces social media, and highlights its large scale and continuous growth. This chapter also describes the challenges associated with processing data from social media (large volume, non-standard language, short length of the texts, and a need to separate useful information from extraneous content), and overviews applications that benefit from social media analysis. This chapter is brief; computer science researchers will already have a general understanding of these issues, and can use the chapter to peruse some of the details. The rest of the chapters can be logically divided into two parts containing two chapters each, and the book ends with a conclusion chapter and a short case study.

Chapters 2 and 3 form the larger part of the book's content, and are also the core sections where NLP methods are discussed. Chapter 2 focuses on how preprocessing of the text can be done. This problem includes tasks such as tokenizing the text into words, normalizing non-dictionary words, part-of-speech (POS) tagging, and syntactic

doi:10.1162/COLI_r_00270

parsing. The need for each task is discussed, together with an overview of work that has developed specialized methods for social media–type texts. Also included are useful discussions about how each of these modules is typically evaluated. The methods are discussed rather briefly; they are pointers to work rather than descriptions of the models or techniques. An exception is a detailed example given for dialect identification, where the model, data, experiments, and comprehensive results are presented. Similar detailed work is presented at a few other chosen points in the book. Newcomers to NLP technology will appreciate a listing of NLP tools in this chapter, and those new to social media processing will find pointers to tools developed specifically for social media texts. Chapter 3 focuses on information extraction problems such as geolocation identification, opinion mining, event extraction, and NLP applications such as machine translation and automatic summarization. Although this chapter is called "semantic analysis" and the term "semantics" is used at several points in the book, the authors acknowledge that the term refers to any intelligent processing of the text, rather than an attempt to create a representation of the text's meaning. Again, detailed examples are presented for geolocation, opinion classification, and machine translation. The usefulness of these examples is mixed, and will be discussed later in this review. Otherwise, the methods and techniques are motivated and described briefly, similar to the previous chapter.

The second part of the book, Chapters 4 and 5, discusses applications that benefit from processing of social media texts, and how data can be obtained for research. A wide spectrum of applications are described including financial, political, and defense interests. Chapter 5 summarizes existing data sets and evaluation methods together with discussions about how useful data can be gleaned from large volumes, and how privacy can be maintained during social media research. The conclusion chapter reiterates the key challenges of social media processing, and brings up questions that are yet to be addressed by the NLP community. The book ends with a brief case study on a social media–based event-monitoring application that combines many of the tasks and technologies discussed earlier in the book.

The authors should be commended for taking up a subject in which NLP has a huge impact. At the same time, as a nascent area, this topic is worthy of discussion within our community to identify the main challenges, the techniques that are consistently useful, as well as the scope for further work. In this regard, this book provides a valuable introduction to the field. For a short book, it covers a number of dimensions to social media language analysis—preprocessing, information extraction, as well as applications. Newcomers to the social media domain will find the book a good reference point to research done in the area. The highlighting of challenges in this domain and differences from NLP work typically carried out on news and other single-author and edited text are very useful. Another handy element is a continuous focus on explaining how methods for various NLP problems are evaluated.

At the same time, although a broad set of topics is discussed, some key areas of social media processing are ignored. One is the use of social media as a tool for language understanding. On one side, researchers are excited about using social media as a large-scale source of conversation data on which questions about language can be asked. Recent work in this area has created new insights into language variation (Doyle 2014; Eisenstein 2015b), phonology (Eisenstein 2013, 2015a), and entrainment (Danescu-Niculescu-Mizil, Gamon, and Dumais 2011; Doyle, Yurovsky, and Frank 2016), to name a few. Another line of work is richer conversation understanding through language processing beyond lexical and syntactic levels. Work on dialog act tagging (Kim,

Cavedon, and Baldwin 2010; Ritter, Cherry, and Dolan 2010), topic structure (Ramage, Dumais, and Liebling 2010; Peng, Wang, and Dredze 2014), as well as phrasing (Niculae and Danescu-Niculescu-Mizil 2014; Tan, Lee, and Pang 2014) on social media conversations fall under this category and yet are not covered in this book. Even on the word- and sentence-level, the book does not cover research carried out in the related web, social media, and social network communities. Many of the models in these areas carry out similar word-level processing but ask different sets of research questions. There are some references to this line of work but the majority of the content remains unexplored. Finally, there is heavy focus on Twitter or microblogs in comparison with other forms of social media. Some sections such as summarization or machine translation entirely focus on Twitter. Although it may be possible that there are logical divisions of social media text types and that tweets represent a class of short texts, such reflections and explanations are not provided in the book.

Going beyond coverage, a more fundamental question that newcomers as well as current NLP researchers want to ask is what is the role of natural language processing for social media. It is clear that the tasks involved in preprocessing content (normalization, POS tagging, and parsing) are clearly in the realm of natural language processing, and so are tasks such as named entity recognition, opinion mining, and event extraction. However, the book does not do a good job of distinguishing what new techniques are required due to distinguishing properties of the social media domain. Some of the tasks appear to be solvable by adding annotated data from this domain to existing NLP corpora, and retraining the models. Other problems appear to be solved by supervised techniques with word-level features, sometimes combined with social media and social network metadata. It is not even clear from the book whether syntactic parsing, and so forth, are useful in current NLP work on social media. A reflection on the types of techniques that work well so far, and insights into new types of models proposed for social media, and an emphasis on methods addressing special properties, for example, streaming algorithms, would have been worthwhile additions. Clarifying these questions would have been an important desired property of a book on NLP and social media, and this is the main place where the book falls short.

The book is also not self-sufficient for understanding the details of social media processing. There is no clear discussion of particular models, especially the ones specific to social media. For example, normalization of non-dictionary words is a task clearly new and motivated by social media texts, but only a very brief section is given to its discussion and devoid of any models or techniques. Often multiple paragraph descriptions of a study are given but without enough details, equations, or technical explanations to actually understand the approach. As a result, not much assimilation of the topic is possible at the end of many sections of the book. Rather, a reader may only use these descriptions as a pointer for finding work to read further and understand. Another undesirable property of this book is a heavy focus on research carried out by the authors of the book themselves. The detailed examples in the book where comprehensive explanations of data, models, and results are given are all the authors' own work with their collaborators. At least from the book, it is not clear why these models should be described in detail compared with other tasks and the work of other researchers. The case study is also on a company associated with one of the authors. For a book focusing on such a wide topic, a biased focus could have been avoided.

Still, overall, this book will be useful as a reference point for work currently done, and a starting point for further discussions on social media and NLP work.

References

Danescu-Niculescu-Mizil, Cristian, Michael Gamon, and Susan Dumais. 2011. Mark my words!: Linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, pages 745–754, Hyderabad.

Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 98–106, Gothenburg.

Doyle, Gabriel, Dan Yurovsky, and Michael C. Frank. 2016. A robust framework for estimating linguistic alignment in twitter conversations. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 637–648, Montreal.

Eisenstein, Jacob. 2013. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 11–19, Atlanta, GA.

Eisenstein, Jacob. 2015a. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19:161–188.

Eisenstein, Jacob. 2015b. Written dialect variation in online social media. In Charles Boberg, John Nerbonne, and Dom Watt, editors, *Handbook of Dialectology*. Wiley.

Kim, Su Nam, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 862–871, Cambridge, MA.

Niculae, Vlad and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2008–2018, Doha.

Peng, Nanyun, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 674–679, Baltimore, MD.

Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, (ICWSM)*, Washington, DC.

Ritter, Alan, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 172–180, Los Angeles, CA.

Tan, Chenhao, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 175–185, Baltimore, MD.

Annie Louis is a lecturer in the School of Computer Science and Electronic Engineering, University of Essex. She is interested in natural language processing and machine learning techniques to understand both how people use language and build language-related technology. In particular, she creates models that infer how documents and conversations are structured, and the impressions they make on the readers and listeners. She is also interested in applying such models to automatically assess the quality of documents, summarize and generate text, and improve search of social media content. Louis' e-mail address is aplouis@essex.ac.uk.