

Predictively Modeling Social Text

William W. Cohen¹

Carnegie Mellon University

Social interactions facilitated by computer are increasingly common, and have an increasing impact on society. Since many of these interactions are recorded, social media now provides a rich new source of data for sociologists and others seeking to understand how communities function. At the same time, we are now beginning to understand that social effects (in contrast to effects driven by purely rational individual decision-making) have a huge impact in many areas, ranging from economic markets to political decision-making to scientific progress. These are some of the factors driving the increasing interest in techniques for mining social media.

In this talk we consider the problem of modeling textual social media. Text comprises a large fraction of social media, and is much easier to store and analyze than images or video. In particular, there is a long history of work in analyzing natural-language text, including a great deal of work on processing text from nearly unrestricted domains - notably newswire text. Social media text, however, differs from newswire text in a number of important ways, which raise interesting technical issues in attempting to analyze it.

First, newswire text has one commonly-understood purpose - namely, communicating factual information about recent events - and hence there has been broad agreement about the goals of analysis - namely, extracting factual information from text. In contrast, social media has *no clear single purpose*, leading to a wide variety of possible goals for researchers to pursue.

Second, newswire text is intended to be read by a particular set of people, who are assumed to have a particular type of background knowledge. In contrast, social media text is often *targeted at a particular community of readers* - often readers that share a large amount of background knowledge with the writer. This means that statements can be made that can be easily interpreted by the reader, even though they are highly ambiguous in a broader context. To take a few examples, an entity name like “Bob” might have an obvious referent in a particular social setting, but be impossible to resolve outside of that setting; or a phrase like “of course” might be clearly understood as sarcasm in one setting, and clearly understood as agreement in another.

The second issue raises many technical questions about *how to best model communities in conjunction with the text* that they give rise to. In my talk, I will describe two key modeling tools - random-walk based graph analysis and probabilistic generative models of text - and discuss how these tools can be extended to model issues of crucial importance in social-media text, such as links between documents, links between community members, and the impact of shared external events.

The first issue raises many technical questions about *how to evaluate* such models. Ideally, a model should be both *understandable* - i.e., it should lend itself to useful visualizations of the data - and *predictive* - i.e., it should allow one to forecast future behavior. In my talk, I will describe several predictive tasks that we have used as to evaluate models of social-media text, including predicting the publication activity of scientists, predicting how people will interpret sequences of events, predicting the link activity of bloggers, and the predicting the commenting activity of members of large community blog sites. Importantly, quantitative differences between different models can be seen on predictive tasks, even when visualizations based on these models are, subjectively, hard to distinguish; furthermore, the relative ordering of different models on predictive tasks is usually similar across a range different tasks.

In more detail, I will discuss a number of distinct modeling tasks. I will describe a probabilistic model, based on latent Dirichlet allocation (LDA) [1] and stochastic block models [2], that jointly models the contents of, and the connections between, a corpus of linked documents [3, 4]. We show that there are significant differences in how well these models can perform certain predictive tasks—notably, predicting which outgoing links will appear in a blog post. The same techniques can also be used to model scientific literature, and have the same relative performance on predicting which citations will appear in a scientific paper. We will then explain how another extension of this model can be used to predict a different type of user activity—predicting which community members will comment on a blog post [5].

We next turn to the issue of modeling *time*, and more specifically, the impact of external events on a corpus of social media. In particular, we explore whether one can reliably detect external events from a time-tagged stream of blog postings. We present several alternative probabilistic models for this task, and show that some of them are nearly as accurate as human annotators; however, this is a less strong statement than one might think, since there is a surprisingly large amount of disagreement among human annotators as to what constitutes a significant event [6]. To address this, we also consider a *semisupervised topic model*, in which users are allowed to provide a partial description of the timeline of events they consider important [7].

Finally we consider an alternative way of detecting events—using random-walk based proximity measures on a heterogeneous graph containing documents, words, and timestamps. We show that nearly-as-accurate predictive results can be obtained with computationally less expensive techniques. We also show that certain novel predictive tasks can be readily addressed with graph similarity measures involving scientific literature [8].

Acknowledgements

The talk describes joint work with Amr Ahmed, Andrew O. Arnold, Ramnath Balasubramanian, Matthew Hurst, Frank Lin, Ramesh Nallapati, Noah A. Smith, Eric Xing, and Tae Yano.

References

1. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. *Journal of machine Learning Research* **3** (2003) 993–1022
2. Airodi, M., Blei, D., Xing, E., Fienberg, S.: Mixed membership stochastic block models for relational data, with applications to protein-protein interactions. In: *Proceedings of International Biometric Society-ENAR Annual Meetings*. (2006)
3. Nallapati, R., Cohen, W.W.: Link-PLSA-LDA: a new unsupervised model for topics and influence of blogs. In: *International Conference for Weblogs and Social Media*. (2008)
4. Nallapati, R.M., Ahmed, A., Xing, E.P., Cohen, W.W.: Joint latent topic models for text and citations. In: *Proc. of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (2008)
5. Yano, T., Cohen, W.W., Smith, N.A.: Predicting response to political blog posts with topic models. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. (2009) 477–485
6. Balasubramanyan, R., Lin, F., Cohen, W.W., Smith, N.A., Hurst, M.: From episodes to sagas: Understanding the news by identifying temporally related story sequences (poster). In: *Proc. of ICWSM-2009*. (2009)
7. Balasubramanyan, R., Cohen, W.W., Hurst, M.: Modeling corpora of timestamped documents using semisupervised nonparametric topic models. In preparation (2009)
8. Arnold, A., Cohen, W.W.: Information extraction as link prediction: Using curated citation networks to improve gene detection. In Liu, B., Bestavros, A., Du, D.Z., Wang, J., eds.: *WASA*. Volume 5682 of *Lecture Notes in Computer Science*, Springer (2009) 541–550