Machine Learning for Social Media Analytics

Prem Melville, Vikas Sindhwani, Richard D. Lawrence, Estepan Meliksetian,
Yan Liu, Pei-Yun Hsueh, Claudia Perlich
IBM T.J. Watson Research Ctr., Yorktown Heights, NY 10598
{pmelvil, vsindhw, ricklawr, emelikse, liuya, pshsueh, perlich}@us.ibm.com

The rise of the blogosphere has empowered the average consumer to influence the public perception and profitability of brands. As such, marketing organizations need to be aware of what people are saying in influential blogs, how the expressed opinions could impact their business, and how to extract business insight and value from these blogs. This has given rise to the emerging discipline of *Social Media Analytics*, which draws from Social Network Analysis, Machine Learning, Data Mining, Information Retrieval, and Natural Language Processing. The automated analysis of blogs and other social media raises several interesting challenges, which we address below.

Seeking Relevance: The immediate objective is to effectively filter down the vast blogosphere from millions to the thousands of blogs most relevant to the brand/product being monitored. We want not only the blogs that directly talk about a specific product, but also those about competing products and potential new customers. Given the complexity of this task, simple keyword searches are inadequate; and instead, we explore text-based and network-based techniques. Based on the content of blogs, we label some posts as *relevant* or *irrelevant*, and treat relevance filtering as a standard text classification problem. An alternative to such text-based techniques is relevance filtering via graph based methods. Known relevant blogs may be viewed as positive labels on a large partially labeled blog graph. Then assuming that links encode similar degrees of relevance, we can apply a classical label diffusion procedure – *snowball sampling* being a crude instance of this. A better approach is a combination of these methods we call *focused snowball sampling*, which iteratively crawls from links in blogs deemed to be relevant by a text classifier. Given a crawl of the entire blogosphere and negative examples, one may also use graph transduction [6] or graph kernel based classification techniques. Further improvement can also be made using classification models that combine graph structure and text content [7].

Influence and Authority: Having identified a subset of relevant blogs, it is then useful to determine the most authoritative bloggers in this space. These are the experts or mavens whose opinions catch on most rapidly. It is important to identify this set of bloggers, since any negative sentiment they express could spread far and wide. In addition to authorities, there are influential bloggers who are very well connected, who are most responsible for the spread of information in the blogosphere. When presented with a large number of posts relevant to a topic, ordering them by the blogger’s influence assists in information triage – since these are the posts most likely to be read by others. Given that we have a network of directed edges indicating the links between posts/blogs, we can apply measures of prestige from Social Network Analysis. For instance, the authority of a blog can be characterized in terms of PageRank based on the number and authority of other blogs that link to it, while the influence of a blog can be measured by Flow Betweenness, that captures the degree to which the blog contributes to the flow of information between other bloggers. There has been a lot of recent work studying influence and the diffusion of information in social networks [1][2], which are important in furthering our understanding of the dynamics of communication in networks; however, they do not directly give us measures of influence and authority in graphs. It would be useful to follow the analyses done in these papers to derive something actionable, such as a better measure of blog influence. Given that there are several alternative approaches to measure a node’s rank in a graph (*Degree Centrality, Closeness Centrality*, etc.) how do we determine which measure to use? Answering this question is non-trivial since the notions of authority and influence vary based on the application, and how the network was generated. For the purposes of marketing we are interested in bloggers who influence the thinking, and subsequently the content blogged by others. If a blogger is indeed influential, we would expect his ideas to propagate to other blogs. Based on this, we propose objectively comparing candidates influence measures on the task of predicting user content generation. A measure of influence is helpful if it enables one to select a blog (or set of blogs) that more accurately predicts future discussion in the blogosphere.

Sentiment Detection: Considering that it's virtually impossible to read all user-generated content, it has become crucial to automatically identify negative (and positive) sentiment in blogs to enable rapid response. The main challenge in doing this is that the expression of sentiment tends to be domain specific, and the set of domains to monitor change often. Thus we require sentiment classifiers that can rapidly adapt to new domains with the minimum of supervision. Treating sentiment detection as a text classification task has made it possible to adapt to domains, provided we have enough training examples in the target domain. However, supervision for a sentiment classifier can be provided not only by labeling documents, but also by labeling words. For instance, labeling a word such as "atrocious" as negative is one way to express our prior belief of the sentiment associated with it. It is possible to learn from such labeled words in conjunction with labeled documents in our proposed framework of Dual Superversion. We have demonstrated that by labeling a few words and few documents we can learn an accurate model that requires less supervision than labeling only documents or only words. We have demonstrated the generality of our approach on different domains, and with two different families of techniques: one based on *pooling multinomials* [3] and the other based on semi-supervised graph transduction methods [4]. The burden of labeling data can be further reduced by exploiting active learning in the Dual Supervision setting [8]. Even though there are expressions of sentiment that are domain-specific, there is still a large amount of overlap in how positive and negative emotion is conveyed across domains. This enables the use of *transfer learning* to adapt a classifier trained in one domain to a new domain with little to no labeled data in the target domain [5].

Emerging Topics: Given the constant chatter on the blogosphere, much insight can be gleaned by examining patterns of what people are blogging about to find emerging areas of discussion. One commonly used approach to capture this notion of *hot* topics is to identify frequently occurring key-phrases. Such an approach may identify that "Barack" and "Obama" are words that frequently appear together, and as a phrase they are mentioned many times in political blogs. However, this fact may not be interesting to report, if the phrase is always frequently mentioned. Instead, if we compare the relative frequency of occurrence of phrases today to the occurrences over the past week, we are likely to identify more currently relevant phrases, such as "public healthcare." Such methods can bring to our attention that certain entities are being extensively discussed; however, from a marketing perspective we also want to know what is being said about them. This requires going beyond key-phrases to identifying sets of posts that are discussing the same topic. A lot of research in the area of topic modeling with Latent Dirichlet Allocation, Probabilistic Latent Semantic Analysis and Non-negative Matrix Factorizations can be brought to bear here. Several recent papers have developed models of temporal evolution of topics in document streams [9], where the main idea is to construct topic models in different time-windows tying them together to maintain some form of temporal continuity. In this direction, we are currently exploring regularized non-negative matrix factorizations. An associated problem is that of *Guided Topic Evolution*, where we incorporate feedback from a user in an online fashion as to which topics should be tracked or discarded from the analysis.

REFERENCES

[1] M. Goetz, J. Leskovec, M. Mcglohon, and C. Faloutsos, "Modeling Blog Dynamics," *AAAI Conference on Weblogs and Social Media*, 2009.

[2] G. Kossinets, J. Kleinberg, and D. Watts, "The structure of information pathways in a social communication network," *KDD*, 2008.

[3] P. Melville, W. Gryc, and R. Lawrence, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification," *KDD*, 2009.

[4] V. Sindhwani and P. Melville, "Document-Word Co-Regularization for Semi-supervised Sentiment Analysis," *ICDM*, 2008.

[5] J. He, Y. Liu, and R. Lawrence, "Graph-based transfer learning", *CIKM*, 2009.

[6] X. Zhu, "Semi-supervised Learning literature survey", Tech Report 1530, Dept. of Comp. Sci., Univ. of Wisconsin, Madison.

[7] V.Sindhwani, "On Semi-supervised Kernel Methods", Doctoral Thesis, University of Chicago, 2007.

[8] V.Sindhwani, P. Melville, R. Lawrence, "Uncertainty Sampling and Transductive Experimental design for Active Dual Supervision", *ICML*, 2009.

[9] D. Blei and J. Lafferty, "Dynamic Topic Models", *ICML*, 2006.