Networks & Sentiment - Project Proposal

Jan Overgoor & Evan Rosen

October 21, 2010

1 The Problem

The way news gets presented is seldom neutral, especially in the editorial domain of blogs. Using a large corpus of hyperlinked blog posts, we plan to investigate the way in which subjectivity and sentiment associated with an item of news can effect its propagation. An immediate question is then what constitutes a unit of news. While an obvious choice is to use text delimited by punctuation marks, as in [Leskovec et al., 2009], we are also interested in exploring the role of the context in which quotes are used. Thus, we hope to demonstrate that some type of emotional variation and selection operates on either quotes themselves, or quotes along with their context. First, we plan to test for the correlation of various dimensions of sentiment with overall quote popularity. Second, we wish to probe the causality of this relationship by tracking the interactions of various quote variants over time in an attempt to give a more detailed picture of the mechanism by which sentiment-laden quotes rise (or don't rise) to popularity.

2 The Data

We plan to use a data set consisting a large collection of mainstream news web sites and blogs, similar to that collected in the MemeTracker project [Leskovec et al., 2009] but with the actual webpage text included as well. This amounts to a large set of text les which, for each page, record the URL, the date of publication, any outgoing URL links, any quotations, and the original html from the page with the location of any quotations marked. The date of publication, in this context, corresponds to the time at which the page was pushed to an RSS feed.

3 The Method

To answer the first question laid out above, we need to somehow score each snippet according to the incidence of sentiment rich phrases. To do this, we plan to build a variety of sentiment lexicons using a sentiment propagation algorithm over a lexical graph built from web data [Velikovich et al., 2010, Godbole et al., 2007]. Once we have created these lexicons we intend to score

each quote simply summing the polarity of each word and normalizing by length. For quote context we plan to use a weighted sum over some context window in which the weight of each word decays as a function of its distance from the beginning or end of the quote. With a score for each snippet, we can then compute the correlation with quote popularity. This whole process will be repeated many times with various seed sets in our sentiment propagation algorithm, in as part of our search for the relevant dimensions of sentiment.

To answer the second question, regarding dynamics among quote variants, we first need to determine the set of quotes which may be appropriately called variants of one another. We plan to use exactly the same approach as the original MemeTracker paper which involves building and partitioning a weighted phrase graph.

To demonstrate that sentiment is responsible for popularity, we then need to show that the particular quote within a quote cluster which achieves the greatest popularity, is in some way dependent on features of its sentiment. The methodological approach is to say that if we control for informational content by comparing quotes in the same cluster, the differences we see in popularity can be attributed to non-informational properties like sentiment. We expect that the quotes which are popular within a given quote cluster, towards the beginning of news story, will be slightly different from the popularity of those present later on. The comparison of sentiment between these two extremes is another important way to demonstrate a causal relationship between sentiment and popularity.

4 Evaluation

For the first question, we can evaluate the correlation between sentiment and popularity using standard correlation metrics. It might prove valuable to create small data set labelled with sentiment, such that we can test whether our sentiment scoring system is working the way we expect it to, independent of any new findings about the data. Unfortunately, evaluation of the second step is less straight forward, as we do not know exactly what to expect in terms of the spread of sentiment across quote variants. One option would be to compute the correlation between within-cluster popularity and sentiment scores for each time step. If this correlation increased over time, we could take it be a strong indication that our causal hypothesis is correct.

5 Related Questions

There are a couple of other hypotheses we could test about the relation between sentiment and the propagation of information. We discuss a number of possible research directions here that we might quickly look into when time allows. For all situations we assume to have some function S that takes a piece of text and returns a value that represents the sentiment contained in the text.

How does network structure affect sentiment?
Does the popularity of a quote affect the sentiment that is generally associated with the quote? It is possible that the method of presentation becomes more subjective when a quote becomes more popular. One way of testing this would be to map out the average sentiment in the context of instances of a quote cluster against the cumulative volume of the quote cluster. A positive result could result into questions about the incentives to write about quotes. One could also try to make claims about the difference in the subjective use of language in mainstream media in blogs. The intuitive assumption that blogs generally use more subjective content could be affirmed, or a contrasting result could show nice features disproving the objectivity of mainstream media.

How does sentiment affect the network?
To start we can look into the correlation between the sentiment content of text and any features of the associated quote cluster networks. We could correlate sentiment with in- and out degrees of documents, the density of the resulting quotation network, ect. A very interesting connection to look into would be the relation between the co-occurrence of quotes between web pages and one of the pages having a hyperlink to to other. How do the quote cluster networks and their hyperlinked sub-networks relate to each other? For highly opinionated content an application of sentiment would be to try to classify the hyperlinks between documents (containing quotes) as pointing to documents supporting your viewpoint or criticizing it. Also, one could cluster the occurrences of a quote by the different perspectives (opinions) about the quote, based on the sentiment that is used to portray the quote.

What is the appropriate operational unit of news?
There is a clear trade-off between the amount of content that is used to represent something and the computational weight of doing so. It would be nicest to use all data available to do things like search and information extraction, but on a very big data set this is often computationally intractable. The approach taken by [Leskovec et al., 2009] is to take quotes, text between quotation marks, as the basic unit to track and reason about. If we can successfully apply sentiment extraction to disambiguate occurrences of quotes and reason about their behavior in the network, we can infer things about where the relevant content is in the domain of quotes. Also, with the appropriate application of sentiment extraction it could be possible to create a simple yet effective representation of the occurrence of the quote that includes more information about the context in which the quote occurs.

6 Deliverable

Ultimately, we hope to create a final report which summarizes the various correlations between sentiment content and popularity, which also provides a detailed

picture of the way in which quote variants with the same informational content vary over time as function of sentiment. Central to our project is the examination of the causal effect of sentiment upon popularity and we want to build a story which best tests this theory. On a more conceptual level, we would like to use these results to comment on the fundamental nature of the news; shifting the focus away from its traditional conception as either an objective or even manipulative source of information, to more of a community of self-replicating entities whose properties arise out of internal features of sentiment, such as emotional variation and selection.

References

[Godbole et al., 2007] Godbole, N., Srinivasaiah, M., and Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

[Leskovec et al., 2009] Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506.

[Velikovich et al., 2010] Velikovich, L., Blair-Goldensohn, S., Hannan, K., and McDonald, R. (2010). The viability of web-derived polarity lexicons. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785.