Large-Scale Sentiment Analysis for News and Blogs (System Demonstration)

Namrata Godbole* namratagodbole@gmail.com
Manjunath Srinivasaiah* manj.blr@gmail.com
Steven Skiena◇ skiena@cs.sunysb.edu

*Google Inc., New York NY, USA

◇Dept. of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400, USA

1. Introduction

News can be good or bad, but it is seldom neutral. Although full comprehension of natural language text remains well beyond the power of machines, the statistical analysis of relatively simple sentiment cues can provide a surprisingly meaningful sense of how the latest news impacts important entities.

Here we demonstrate our large-scale sentiment analysis system for news and blog entities built on top of the *Lydia* text analysis system [1]. We determine the public sentiment on each of the hundreds of thousands of entities that we track, and how this sentiment varies with time. We encourage the reader to study our historical sentiment analysis for your favorite news entities at *http://www.textmap.com* and view daily sentiment at *http://www.textmap.com/sentiment*.

The results of our sentiment analysis correlate very well with historical events:

- *The Popularity of U.S. President George W. Bush* – Gallup/USA Today conducts a weekly opinion poll of about 1,000 Americans to determine public approval of their President. Figure 1 presents two time series; both our George W. Bush sentiment index (in blue) and the Gallup poll's presidential public approval rating (in red). Both time series are measured weekly, with the values reported in terms of z-scores (deviations from the mean). They show a strong positive correlation (coefficient 0.372) over a two year interval.

- *The Downfall of Enron* – Enron collapsed dramatically from one of the most respected U.S. corporations into bankruptcy over the last quarter of the year 2001. This sharp decline is captured in Enron's sentiment time series, shown in red in Figure 2. The blue line presents our *subjectivity* index, measuring the volume of subjective references concerning the entity. It identifies Enron as a subject of extreme passion in the news for a period of almost two years after its collapse. The grey line denotes the relative volume of references to Enron.

2. Evaluation

We are unaware of a similar *Lydia*-style entity sentiment analysis system which we can use as a gold standard to compare and verify our sentiment assessment. However, we can correlate the polarity and subjectivity scores to various independent measures of entity performance, such as sports team results and stock market indices. These experiments yield confirmation of the validity of our sentiment measures.

2.1 Baseball team evaluation

We predict that sentiment on the state of a sports team should be higher after a win than a loss. To test this prediction, we collected the outcomes of every Major League Baseball game played between July 2005 to May 2006. These performance time-series can be correlated with the polarity and subjectivity scores with different lead/lag time intervals to study the impact of performance on sentiment.

Our results show a significant spike in sentiment correlation with a lag of +1 day, which reflects when newspapers report the match results. Interestingly, the sentimental impact of each game has a substantial half-life, hanging on for more than a week before disappearing.

2.2 Stock Index vs. World Sentiment Index

Our daily time series of the relative occurrences of positive and negative words over all news text gives us a measure of the "happiness" of the world. We reason that world sentiment is closely related to the state of the economy, which is generally reflected by the stock market. To test this hypothesis, we correlate our global index to the Dow Jones stock index. The indices show a correlation coefficient of +0.41 with a time lag of 1 day, as expected given reportage delays.

2.3 Seasons vs. World Sentiment Index

Finally, we present a plot of the world sentiment index against the seasons of the year, shown in Figure 3. Observe that the volatility in world sentiment is considerably reduced during the summer months, as most of the industrial world takes its summer vacations. There also seem to be other periodic seasonal flows in sentiment. Interestingly, the lowest time point on the graph is not the period of the World Trade Center attack (September 2001) but rather April 2004, reflecting the Madrid train bombings, the start of insurgency in Iraq, and the breaking of the Abu Ghraib prison story.

References

[1] Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. In: Proc. Int. Conf. Weblogs and Social Media (ICWSM 07). (2007)

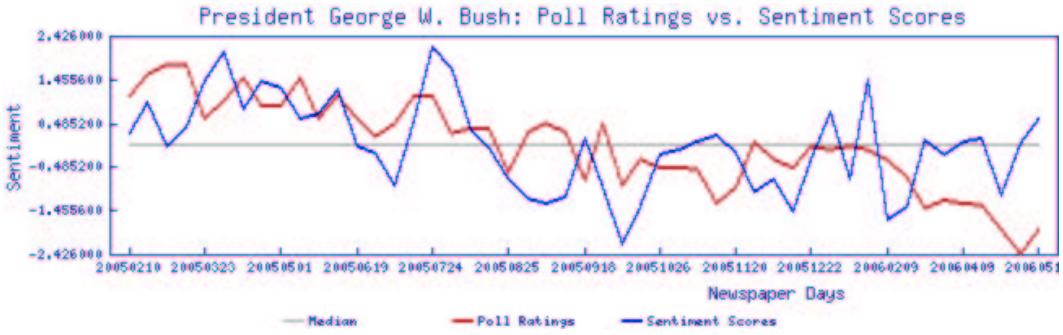

Fig. 1: *President Bush: Gallop poll opinion ratings (red) vs. our news sentiment index (blue).*

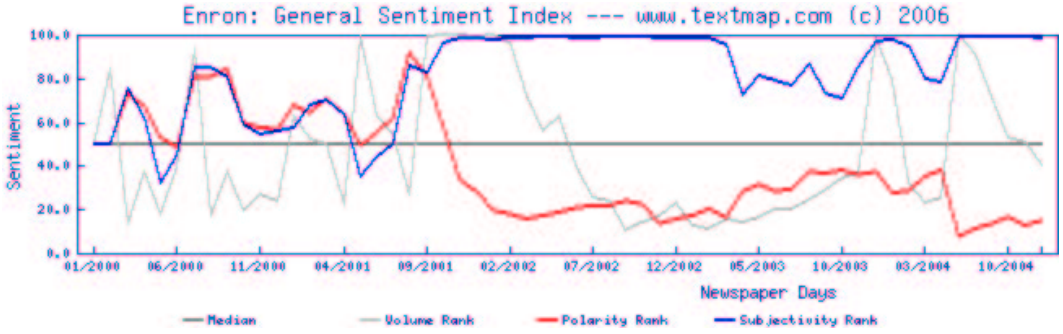

Fig. 2: *The Collapse of Enron, captured by our news sentiment (red) and subjectivity (blue) indices. The relative volume of references to Enron is shown in grey.*

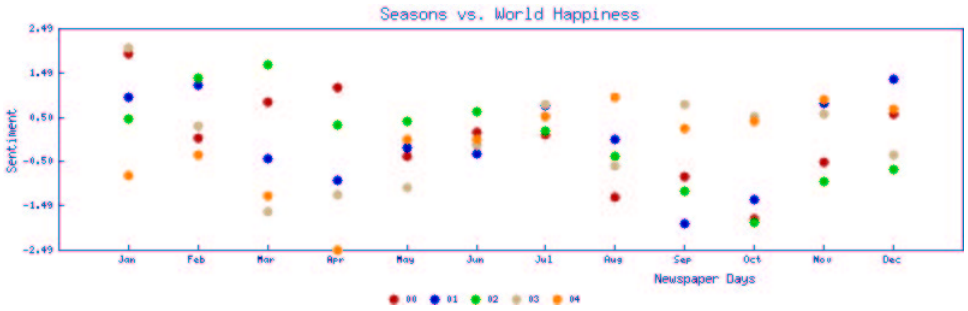

Fig. 3: *Calendar effects on our world sentiment index.*