Tweetin' in the Rain: Exploring Societal-scale Effects of Weather on Mood

Aniko Hannak[†] **Eric Anderson**[†] **Lisa Feldman Barrett**[†]
Sune Lehmann[‡] **Alan Mislove**[†] **Mirek Riedewald**[†]

[†]Northeastern University [‡]Technical University of Denmark

Abstract

There has been significant recent interest in using the aggregate sentiment from social media sites to understand and predict real-world phenomena. However, the data from social media sites also offers a unique and—so far—unexplored opportunity to study the impact of external factors on aggregate sentiment, at the scale of a society. Using a Twitter-specific sentiment extraction methodology, we the explore patterns of sentiment present in a corpus of over 1.5 billion tweets. We focus primarily on the effect of the weather and time on aggregate sentiment, evaluating how clearly the well-known individual patterns translate into population-wide patterns. Using machine learning techniques on the Twitter corpus correlated with the weather at the time and location of the tweets, we find that aggregate sentiment follows distinct climate, temporal, and seasonal patterns.

Introduction

There has been significant recent interest studying the aggregate sentiment of postings on online social media sites like Twitter. Researchers have identified interesting patterns of aggregate sentiment (Dodds et al. 2011; Golder and Macy 2011) and have used aggregate sentiment to measure and predict real-world events, including the stock market (Gilbert and Karahalios 2010), the box office success of movies (Asur and Huberman 2010), and political polls (O'Connor et al. 2010).

However, the data from social media sites also offers a unique and—so far—unexplored opportunity to study the impact of external factors on aggregate sentiment, at the scale of a society. For example, psychologists have studied the sentiment of individuals (hedonic feelings of pleasantness; referred to in the psychological literature as "affect" (Barrett and Bliss-Moreau 2009)) and found surprising daily (Stone et al. 1996), weekly (Larsen and Kasimatis 1990), seasonal (Rohan and Sigmon 2000), geographic (Mersch et al. 1999), and climate-related (Denissen et al. 2008) patterns. Unfortunately, these studies have been limited in scale by their methodology; they often rely on repeated surveys, study a single factor (e.g., temperature), and only consider up to a few hundred subjects. Thus, it remains unclear whether the observed individual-level patterns translate into population-wide trends, which of the factors dominate the population-wide signal, and how multiple factors interact to influence sentiment.

In this paper, we take the first steps towards understanding the influence of weather and time on the aggregate sentiment from Twitter. We first use a Twitter-specific methodology for inferring the sentiment of Twitter messages (tweets) that is able to handle the unique grammar, syntax, abbreviations, and conventions of Twitter. Using a corpus of over 1.5 billion messages, we automatically create a large sentiment-scored word list based on the co-occurrence of each token with emoticons.

We then examine whether known patterns of individual sentiment result in population-wide patterns of aggregate sentiment. Specifically, we treat the detection of patterns as a machine learning problem, with a goal of trying to predict the aggregate sentiment given input variables such as time of day, season, and weather. Using machine learning allows us to capture potentially complex, non-linear interactions between different variables. Overall, we find that our machine learning techniques can predict the aggregate sentiment with ROC area over 0.78, indicating high accuracy. Our results can inform existing algorithms that make predictions using aggregate sentiment, and suggest that many of the previously-observed variations in aggregate sentiment are part of repetitive patterns trigged by external factors.

Background

We obtained Twitter data from 2006–2009, covering 54,981,152 accounts and a total of 1,516,115,233 tweets (Cha et al. 2010). Because the number of tweets grew dramatically as Twitter became more popular, for the remainder of this paper, we focus only on tweets issued between January 1, 2009 and September 1, 2009; this leaves us with 1,369,833,417 tweets (90.3% of the entire data set).

Geographic data

To determine geographic information about users, we use the self-reported *location* field in the user profile.[1] The location is an optional self-reported string; we found that

[1]The geo-tagging of individual tweets was not widely deployed at the time of our data collection.

75.3% of the publicly visible users listed a location. We query Google Maps with each unique location, and receive a latitude and longitude as a response for the locations that Google is able to interpret.

To correlate our Twitter data with weather information, we aggregate the users into U.S. metropolitan areas. Using data from the U.S. National Atlas and the U.S. Geological Survey, we restrict our scope to the 20 largest U.S. metropolitan areas as defined by the U.S. Census Bureau. The resulting data set covers 1,391,902 users (2.79% of all users) and 110,082,511 tweets (8.03% of all tweets).

Weather data

In order to collect weather data, we use Mathematica's WeatherData package. In brief, the WeatherData package aggregates and publishes weather data from the National Oceanic and Atmospheric Administration. For each metropolitan area, we retrieve the cloud cover percentage, humidity, temperature, precipitation, and wind speed for every hour for the same period as our tweets.

Measuring Sentiment

To estimate the sentiment of users on Twitter, we examine the content of their tweets. Ideally, we would like to use existing sentiment inference algorithms (Liu 2006). However, due to the strict length requirement on tweets (140 characters), most Twitter messages often contain abbreviations and do not use proper spelling, grammar, or punctuation. Thus, algorithms trained on proper English text do not work as well when applied to Twitter messages.

As a result, most prior Twitter sentiment analysis work has focused on *token lists*, containing a set of tokens (words) with a sentiment score attached to each (Bradley and Lang 1999). Unfortunately, using existing lists on Twitter data has a number of drawbacks: First, the 140 character limit often causes users to abbreviate words, and the lists rarely contain abbreviations. Second, Twitter users often use neologisms and acronyms (e.g., OMG, LOL) and Twitter-specific syntax (e.g., hashtags like #fail) when expressing sentiment; these also rarely appear on existing lists. Third, due to the limited size of existing lists (to the best of our knowledge, the largest list contains only 6,800 tokens), the fraction of tweets that contain at least one listed token is often small.

Methodology

In order to address the challenges above, we construct a Twitter-specific token list by using the tweets themselves (Read 2005). First, we narrow ourselves to English tweets by only considering tweets that have at least 75% of the tokens appearing in the Linux `wamerican-small` English dictionary. Next, we derive an initial set of clearly positive and negative tweets by considering only tweets that contain exactly one of the emoticons[2] `:)`, `:-)`, `:(`, `:-(`; this results in 15,668,367 tweets with a positive emoticon and 5,237,512 tweets with a negative emoticon. We tokenize these tweets on spaces (ignoring hashtags, usernames, and URLs, as well as any token that did not appear at least 20 times), giving us 75,065 unique tokens. Finally, to create our sentiment-scored token list, we calculate the relative fraction of times each token occurs with a positive emoticon and use this as the token's score. Thus, each token's score ranges between 0 and 1, and indicates propensity for the token to appear in positive-emoticon-tagged tweets.

Similar to previous lists, we calculate the sentiment score of a tweet by looking for occurrences of listed tokens, taking the weighted average on the individual token sentiment scores to be the sentiment score of the tweet.

Evaluation

We now examine the accuracy of inferring the sentiment of tweets with our token list. To do so, we create a list of manually, human-rated tweets using Amazon Mechanical Turk (AMT) by paying Turk users $0.10 to rate the sentiment of 10 tweets. The text and response input used in the HIT was modeled after surveys from previously used (Bradley and Lang 1999) lists. We create a test set consisting of 1,000 tweets. Each tweet was rated by 10 distinct individuals physically located in the United States, for a total of 10,000 individual ratings. Based on these 10 ratings, we calculate an average AMT sentiment score for each tweet. We then examine the Pearson correlation between the average of the human ratings and our token list-based rating; we find the two to have a correlation coefficient of 0.651, demonstrating that our sentiment inference methodology is strongly correlated with human ratings.

Sentiment Patterns

With our Twitter-specific word list in hand, we now turn to examine the patterns of sentiment that exist. To do so, we treat the problem as a machine learning problem, with the goal of predicting aggregate sentiment.

Decision trees

To convert our problem to one that is amenable for machine learning, it is necessary to aggregate tweets together (since predicting the sentiment of an individual tweet without any knowledge of the tweet content is remarkably hard). Thus, we aggregate the tweets into hour-long buckets for each of the metropolitan areas. In more detail, for each of the 20 metropolitan areas we consider, we aggregate tweets into hourly buckets, taking the average of the sentiment of all tweets to be the sentiment score for the bucket. This results in 5,832 hour-long buckets for each metropolitan area.

We choose to use bagged decision trees (Breiman 1996) as our machine learning algorithm for several reasons: they can handle all attribute types and missing values, they can provide an explanation why the tree made a certain prediction for a given input, they are among the very best prediction models for both classification and regression problems (Caruana and Niculescu-Mizil 2006), and they can be easily trained and queried in parallel.

[2]We choose to use emoticons as they often represent the true sentiment of the tweet (Vogel and Janssen 2009) and often match the underlying sentiment of the writer (Derks, Bos, and von Grumbkow 1997).

For each experiment, we first create a training set consisting of 66% of the input data, and reserve the remainder of the input data as a test set. We build our predictor by constructing 1,000 decision trees. Each tree is trained on an independent bootstrap sample of the training set, drawn with replacement (Dietterich 2000) (this is a common method in machine learning that results in better predictions and excludes the appearance of random variables as important ones). Finally, we take the average prediction of the 1,000 trees to be the overall prediction.

To simplify the creation of trees, the input sentiment score for the training and test set is reduced from a rational number (the average of all tweet sentiments) to a binary positive (1) or negative (0) value. The cutoff for the positive/negative division for each experiment is chosen to be the median of the union of the training and test sets, meaning an equal number of input data points are labeled with 1 and 0.

Measuring prediction accuracy

In order to measure the accuracy of sentiment prediction, we require a way to compute the likelihood that the predictor ranks time periods with more positive sentiment higher than time periods with more negative sentiment. To do so, we use the metric *Area under the Receiver Operating Characteristic (ROC) curve* or A'. In brief, this metric represents the probability that our predictor ranks two periods in their true relative order (Fogarty, Baker, and Hudson 2005). Therefore, the A' metric takes on values between 0 and 1: A value of 0.5 represents a random ranking, with higher values indicating a better ranking and 1 representing a perfect ordering of the sentiment scores. A very useful property of this metric is that it is defined independent of the functional shape of the distribution of the true sentiment scores. In general, an A' of 0.7 or higher is viewed as providing good predictive value.

Input variables

In order to predict the sentiment on Twitter, we provide four different classes of variables as input to the predictor. First, we examine geography (G) by considering the metropolitan area. Thus, the G input variable takes on one of 20 values, one for each metropolitan area. Second, we examine the season (S) by considering the month. This variable is intended to capture any long-term season variable in sentiment, and can take on one of nine values (since our input data only covers January–September). Third, we examine the time (T) by considering the day-of-month, day-of-week, and hour-of-day. These variables together are intended to capture short-term periodicity in sentiment.

Fourth, we examine the effect of climate by examining weather (W). The weather variables we include consist of humidity, cloud cover, precipitation, temperature, and wind speed. Additionally, because weather may have compounding effects, we include historic weather information by providing the average of each weather variable for the past 1, 2, 3, 6, 12, 24, 48, 72, and 96 hours. Thus, there are 45 distinct weather variables (five variables, each averaged over nine time periods).

Variable classes	Area Under ROC Curve
G, S	0.6585
W, S	0.7427
T, S	0.7450
W, G	0.7561
T, W	0.7724
G, T	0.7753
W, G, T, S	0.7857

Table 1: Area under the ROC curve for different combinations of input variables.

Results

We now turn to examine the effectiveness of bagged decision trees when trying to predict sentiment. We begin by examining each of the four input data variables classes separately, before examining trees built using combinations of the variables. Doing so allows us to understand the relative contribution of each of the variable classes.

Prediction performance We construct bagged trees with each of the input variable classes independently, and measured their performance on the test set. As before, we measure performance using the A' metric, which can be interpreted as capturing the probability that the tree correctly orders each pair of test records. The results of this experiment are a A' of 0.5998 for Season (S), 0.6555 for Geography (G), 0.7274 for Time (T), and 0.7378 for Weather (W).

We make two interesting observations. First, all four variable classes show a ROC area significantly greater than 0.5. This indicates that all four have predictive value, even when viewed independently of other variables, when predicting the aggregate sentiment of tweets. Second, the relative magnitude of the ROC area provides guidance as to the predictive power of each of the variable classes. Clearly, the time and weather variables provide the greatest amount of information, suggesting that daily/weekly and climate-based patterns exist.

Next, we examine the performance of trees produced by combinations of variable classes. Presented in Table 1, the results demonstrate that, as expected, the predictive performance of the trees increases as more variables are added. In particular, once all variable classes are used when training the tree, the A' value of the resulting tree is 0.7857—substantially higher than 0.5. This result indicates that the well-studied patterns of individual sentiment do indeed result in trends of aggregate sentiment, and can even be predicted with high accuracy.

Complex interactions Recall that our motivation for using a machine learning approach was to be able to capture potentially complex, non-linear prediction dependencies between input variables. For example, humidity may serve as a useful predictor of sentiment, but only if the temperature is above a certain threshold. To better explore such trends, we now take a closer look into the bagged tree built using all input variables.

It is generally challenging to visualize a multidimensional function, including those encoded by a machine learning

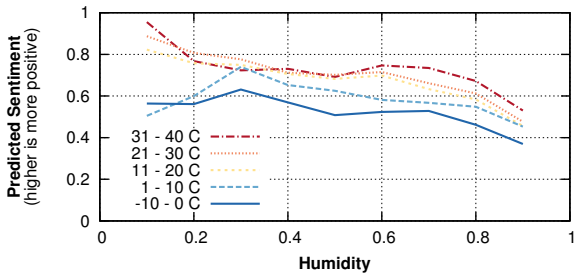

Figure 1: Partial dependence plot of predicted sentiment score from the all-variable bagged tree, based on different combinations of humidity and temperature. As humidity increases the predicted sentiment score decreases (with a more pronounced effect at higher temperatures).

model. A popular way of doing so are partial dependence plots (Panda, Riedewald, and Fink 2010), which visualize partial dependence functions. A partial dependence function for a given multi-dimensional function $f(\mathbf{X})$ (where $\mathbf{X}$ is a vector of multiple input variables) represents the effect of some of the input variables on $f(\mathbf{X})$ after accounting for the average effects of all the other input variables on $f(\mathbf{X})$. Partial dependence plots on appropriately chosen variable combinations can also be used for visualizing variable interactions captured by a model.

Figure 1 examines the interaction of humidity and temperature. The trends observed match intuition about the effect of external variables on sentiment: as the humidity increases, the predicted sentiment score decreases for all values of temperature. However, this decrease is especially pronounced at higher temperatures, suggesting the humidity has a much more profound effect on sentiment when the temperature is higher.

Conclusion

There has been significant recent interest in the aggregated sentiment from social media sites like Twitter. In this paper, we examined the effect of external factors on sentiment. We found that the well-studied dependence on time of day, season, location, and climate appear as population-wide trends, allowing the aggregate sentiment itself to be predicted with an ROC area of 0.78. These results can inform existing algorithms, and suggest that many of the previously observed variations in aggregate sentiment are part of repetitive patterns, rather than unique, new information.

Acknowledgements

We thank Fabricio Benevenuto and Meeyoung Cha for their assistance in gathering the Twitter data used in this study. We also thank the anonymous reviewers for their helpful comments, and Alper Okcan for his assistance with the machine learning algorithms. This research was supported in part by NSF grants CNS-1054233 and IIS-1017793, and an Amazon Web Services in Education Grant.

References

Asur, S., and Huberman, B. 2010. Predicing the Future with Social Media. In *WI*.

Barrett, L. F., and Bliss-Moreau, E. 2009. Affect as a Psychological Primitive. *Exp. Soc. Psy.* 41.

Bradley, M. M., and Lang, P. J. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Breiman, L. 1996. Bagging Predictors. *Machine Learning* 24(2).

Caruana, R., and Niculescu-Mizil, A. 2006. An empirical comparison of supervised learning algorithms. In *ICML*.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*.

Denissen, J. J. A.; Butalid, L.; Penke, L.; and van Aken, M. A. G. 2008. The effects of weather on daily mood: A multilevel approach. *Emotion* 8.

Derks, D.; Bos, A. E.; and von Grumbkow, J. 1997. Emoticons and social interaction on the Internet: the importance of social context. *Computers in Human Behavior* 23(1).

Dietterich, T. G. 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning* 40(2).

Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. A.; and Danforth, C. M. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS One* 6(12).

Fogarty, J.; Baker, R. S.; and Hudson, S. E. 2005. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In *GI*.

Gilbert, E., and Karahalios, K. 2010. Widespread Worry and the Stock Market. In *ICWSM*.

Golder, S. A., and Macy, M. W. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333(6051).

Larsen, R. J., and Kasimatis, M. 1990. Individual differences in entrainment of mood to the weekly calendar. *J. Per. & Soc. Psych.* 58(1).

Liu, B. 2006. *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer-Verlag.

Mersch, P. P. A.; Middendorp, H. M.; Bouhuys, A. L.; Beersma, D. G. M.; and van den Hoofdakker, R. H. 1999. Seasonal affective disorder and latitude: a review of the literature. *J. Aff. Disord.* 53(1).

O'Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *ICWSM*.

Panda, B.; Riedewald, M.; and Fink, D. 2010. The Model-Summary Problem and a Solution for Trees. In *ICDE*.

Read, J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *ACL*.

Rohan, K. J., and Sigmon, S. T. 2000. Seasonal mood patterns in a northeastern college sample. *J. Aff. Disord.* 59(2).

Stone, A. A.; Smyth, J. M.; Pickering, T.; and Schwartz, J. 1996. Daily Mood Variability: Form of Diurnal Patterns and Determinants of Diurnal Patterns. *J. App. Soc. Psych.* 26(14).

Vogel, C., and Janssen, J. 2009. *Emoticonsciousness*. Springer-Verlag Publishers. chapter 2, 271–287.