Tweetin' in the Rain: Exploring societal-scale effects of weather on mood

Paper ID 146

Abstract

There has a been significant recent interest in using the aggregated sentiment from social media sites to understand and predict real-world phenomena, ranging from the stock market to political polls to the box office success of movies. However, the data from social media sites also offers a unique and—so far—unexplored opportunity to study the impact of external factors on aggregated sentiment, at the scale of a society. For example, can well-studied disorders like seasonal affective disorder be observed at large scale?

Using Twitter-specific sentiment extraction methodology, we 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. Overall, we observe that aggregate sentiment can be predicted as positive/negative with an ROC area of 0.78, indicating high accuracy.

Introduction

There has been significant recent interest in using the sentiment, in aggregate, of postings on online social media sites like Twitter in order to measure and predict real-world events. For example, recent work has explored predicting the stock market (Eric Gilbert 2010; Bollen, Mao, and Zeng 2010), forecasting the success of movies at the box office (Asur and Huberman 2010), and replacing traditional political polling (O'Connor et al. 2010; Tumasjan et al. 2010) with data taken from Twitter.

However, the data from social media sites also offers a unique and—so far—unexplored opportunity to study the impact of external factors on aggregated 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 (Mersch et al. 1999) patterns. Unfortunately, these studies have been limited in scale by their methodology; they often rely on repeated surveys, and the largest of these studies examine only a few hundred subjects. As a result, most studies are also limited to examining the effect of a single variable (e.g., temperature) on sentiment. Thus, it remains unclear (a) whether the individual-level patterns translate into population-wide trends, (b) if so, which of the variables dominate the population-wide signal, and (c) how multiple variables interact to influence sentiment.[1]

In this paper, we take the first steps towards understanding the influence of weather and time on the aggregated 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. Due to the massive scalability required to process such large data sets in near-realtime, most existing approaches measure sentiment using lists of positive/negative words and phrases; we do as well. Using a corpus of over 1.5 billion messages, we automatically create a sentiment-scored word list for a large set of tokens based on the co-occurence of each token with emoticons (Read 2005; Pak and Paroubek 2010). We demonstrate that our resulting list has high accuracy, based on a comparison to manually rated messages from Amazon Mechanical Turk.

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 (rather than simply looking at variable correlations) allows us to capture potentially complex, non-linear interactions between different vari-

[1] Recent work (Dodds et al. 2011) has observed that patterns of aggregate sentiment do appear to exist in online social networking services, but focuses only on temporal patterns.

ables. Overall, we find that our machine learning techniques can predict the aggregate sentiment with an ROC area over 0.78, indicating high accuracy.

Additionally, using machine learning allows us to explore the dependence between variables that is used to make predictions. For example, we find strong interdependence on the predicted sentiment between the temperature and humidity, matching common intuition. 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, rather than unique, new information.

Background

Twitter is a "micro-blogging" service that allows users to multicast short messages (called *tweets*). Each user has a set of other users (called *followers*) who receive their messages. The follow relationship in Twitter is directed, and requires authorization from the followee only when the followee has elected to make their account private. Each tweet can only be up to 140 characters in length. The default setting in Twitter is to allow all tweets to be publicly visible; at the time of our data collection, we found that only 8% of users elected to make their account private.

Twitter data

We obtained data from Twitter using the Twitter API from August 15–September 1, 2009 (Cha et al. 2010). Using a cluster of 58 whitelisted machines, we iteratively requested information about each user, including their profile, their followers, and their tweets.[2] In total, we obtained information on 54,981,152 in-use accounts connected together by 1,963,263,821 follow links, and a total of 1,516,115,233 tweets.[3]

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 through September 1, 2009. Doing so allows us to ensure that we have a sufficient number of tweets per location and time period. Using only tweets issued in 2009 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. The location is an optional self-reported string; we found that 75.3% of the publicly visible users listed a location. In order to turn the user-provided string into a mappable location, we use the Google Maps API. Beginning with the most popular location strings (i.e, the strings provided by the most users), we query Google Maps with each location string. If Google Maps is able to interpret a string as a location, we receive a latitude and longitude as a response. We restrict our scope to users in the U.S. by only considering response latitudes and longitudes that are within the U.S.. In total, we find mappings to a U.S. longitude and latitude for 246,015 unique strings, covering 3,279,425 users (representing 8.8% of the users who list a location).

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 map each of the 246,015 latitudes and longitudes into their respective U.S. county. We then consider only the counties that are part of the 20 largest U.S. metropolitan areas as defined by the U.S. Census Bureau (U.S. Census Metropolitan Areas and Components). Unless otherwise stated, our analysis for the remainder of this paper is at the metropolitan-area level.

Weather data

In order to collect weather data, we use Mathematica's WeatherData package (Mathematica WeatherData Package). In brief, the WeatherData package aggregates weather data from the National Oceanic and Atmospheric Administration, the U.S. National Climatic Data Center, and the Citizen Weather Observer Program. For each of the 20 metropolitan areas, we collected the cloud cover percentage, humidity, temperature, precipitation, and wind speed for every hour period from 00:00:00 on January 1, 2009 until 00:00:00 on September 1, 2009 (the same period as our tweets cover).

Measuring Sentiment

Sentiment analysis is a well-studied topic, with much recent work focusing on leveraging sentiment expressed on Twitter to predict real-world phenomena. In this section, we detail our sentiment inference methodology and present an evaluation of its accuracy.

Background

In order to estimate the sentiment of users on Twitter, we examine the content of their tweets. Ideally, we would like to use existing sentiment analysis techniques (Turney 2002; Pang, Lee, and Vaithyanathan 2002; Liu 2006). However, there are a few unique characteristics of Twitter that make natural language processing (NLP)-based techniques not directly applicable. First, the amount of data we have (multiple terabytes, when uncompressed) requires an extremely efficient approach. Unfortunately, many NLP techniques are simply not fast enough to make the analysis feasible. Second, 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. As a result, NLP algorithms trained on proper English text do not work as well when applied to Twitter messages.

[2]Twitter's userids are numerically assigned, allowing us to enumerate each user.

[3]This study was conducted under Northeastern University Institutional Review Board protocol #10-03-26.

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; Wilson, Wiebe, and Hoffmann 2005; Hu and Liu 2004). Unfortunately, existing lists present a number of challenges when used on Twitter data: First, Twitter messages are limited to 140 characters, causing users to often abbreviate words; these lists rarely include such abbreviations. Second, Twitter users often use nelogisms and acronyms (e.g., OMG, LOL) and Twitter-specific syntax (e.g., hashtags like #fail) when expressing sentiment. Existing lists do not include or account for such acronyms. 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; Pak and Paroubek 2010). In brief, we consider only tweets that contain exactly one of the emoticons `:)`, `:-)`, `:(`, `:-(`, as the emoticons 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). We then look at the tokens that occur in these tweets, and calculate the fraction of time each token appears with one of the positive emoticons. This results in a token list with a weighting for each token, where the weighting indicates the propensity for the token to appear in positive-emoticon-tagged tweets.

In more detail, we start with the collection of all tweets. We first narrow ourselves to English tweets by only considering tweets that have at least 75% of the tokens (delimited by spaces) appearing in the Linux `wamerican-small` English dictionary.[4] This narrows our tweet collection to 591,406,152 tweets, which are the tweets we wish to infer the sentiment for.

In order to construct our token list, we derive an initial set of clearly positive and negative tweets by extracting the tweets with exactly one of the four emoticons above; this results in 15,668,367 tweets with a positive emoticon and 5,237,512 tweets with a negative emoticon (a ratio almost 3-to-1). We then tokenize the tweets on spaces (ignoring hashtags, usernames, and URLs) resulting in 277,137,071 occurrences of 937,905 unique tokens. We ignore any token that did not appear at least 20 times, giving us 275,193,529 occurrences of 75,065 unique tokens. To create our token list, we calculate the relative fraction of times the token occurs with a positive emoticon and use this as the token's score. For example, the token `relaxing` occurred in 39,584 tweets with positive emoticons and 3,439 tweets with negative emoticons, giving `relaxing` a score of 0.9201.

[4]The is a standard list of 50,252 words that is used by spell-checking programs on the Linux platform.

Similar to previous lists, we calculate the sentiment of a tweet by looking for occurrences of listed tokens, taking the average on the individual token sentiment scores to be the sentiment score of the entire tweet. In more detail, if a tweet contains n tokens that are present in the token list and their sentiment scores are $\{v_1, v_2, \ldots, v_n\}$ and the frequency of each of these tokens in the tweet is $\{f_1, f_2, \ldots, f_n\}$, the sentiment score of the tweet is calculated by the weighted mean of the scores

$$V_{tweet} = \frac{\sum_i^n v_i f_i}{\sum_i^n f_i} \quad (1)$$

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. We find that the AMT results showed a strong inter-respondent Pearson correlation of 0.784,[5] which is in line with the results from other studies using AMT respondents (Peng and Park 2011). 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 rating for our 1,000 tweets. We find the two to have a correlation coefficient of 0.651, demonstrating that our sentiment inference methodology is close to human ratings.

We make this resulting token list, as well as the code necessary to generate a similar list from a different set of input tweets, available to the research community at `http://socialnetworks.ccs.neu.edu`.

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. Doing so has the advantages of capturing potentially complex, non-linear interactions between input variables that would be missed if we simply looked for pairwise variable correlations. Below, we first detail our machine learning approach before evaluating the effectiveness of sentiment prediction and examining the relative importance of input variables.

Decision trees

To convert our problem to one that is amenable for machine learning, it is necessary to aggregate tweets

[5]This represents the correlation between each rating and the average of the other nine ratings for the same tweet.

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 from January 1, 2009–September 1, 2009 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 chose to use bagged decision trees (Breiman 1996) as our machine learning algorithm for several reasons. First, trees can handle all attribute types and missing values. Second, the split predicates in tree nodes provide an explanation why the tree made a certain prediction for a given input. Third, bagged trees are among the very best prediction models for both classification and regression problems (Caruana and Niculescu-Mizil 2006). Fourth, they are perfectly suited for explanatory analysis because they work well with fairly little tuning. Fifth, bagged trees can be easily trained in parallel, and querying the trees for predictions can be parallelized as well.

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. For each predictor, we build 1,000 decision trees, each on an independent bootstrap sample, and take the overall average prediction of these 1,000 trees to be the overall prediction. For each tree, we generate a training set for that tree by selecting randomly from the training set 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.

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. It is worth noting that this approach can easily work with other positive-versus-negative thresholds, or even use multiple levels on the scale from negative to positive. However, a full exploration is beyond the scope of this paper.

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. Values below 0.5 indicate an inverse ranking, or one where periods with more positive sentiment tend to be ranked lower than periods with more negative sentiment. 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, so it is comparable across different experimental setups and schemes. In general, an A' of 0.7 or higher is viewed as providing good predictive value.

Variable class	Area Under ROC Curve
Season (S)	0.5998
Geography (G)	0.6555
Time (T)	0.7274
Weather (W)	0.7378

Table 1: Area under the ROC curve for different classes of input variables. The climate-based variables (captured by W) and periodic variations (captured by T) show the strongest predictive value, while the other variables all provide useful predictions.

Input variables

In order to predict the sentiment on Twitter, we examine four different classes of input variables. 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).

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 indepen-

dently, 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 presented in Table 1.

We note two interesting observations from this experiment. 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 2, 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 model. A popular way of doing so are partial dependence plots (Panda, Riedewald, and Fink 2010; Hastie, Tibshirani, and Friedman 2009; Hochachka 2006; Friedman 2001; Hooker 2004; Linton and Nielsen 1995), which visualize partial dependence functions. A partial dependence function for a given multi-dimensional function $f(\mathbf{X})$ (where $\mathbf{X}$ is a vector

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 2: Area under the ROC curve for different combinations classes of input variables. All variables show an increase in predictive power, peaking at an ROC of 0.7857 for the combination of all four variable classes.

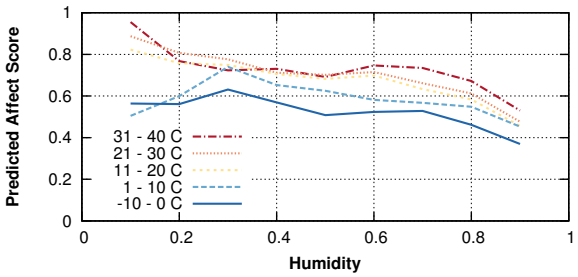

Figure 1: Partial dependence plot of predicted sentiment score from the all-variable bagged tree, based on different combinations of humidity and temperature. As humidity increases the predicted sentiment score decreases (with a more pronounced effect at higher temperatures), matching intuition.

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.

In brief, the method works as follows: suppose we are interested in studying the interaction of input variables i_m and i_n, among the entire set of input variables $\{i_1, i_2, ...i_k\}$. For each element (a, b) in the cross product of all values of i_m and i_n, we create a new input data set with every value of i_m replaced with a and every value of i_n replaced with b. We feed this data set into the predictor, and take the average predicted aggregate sentiment of all data points to be the predicted aggregate sentiment at $i_m = a$ and $i_n = b$. Repeating this method for all values of i_m and i_n provides a high-level overview of how i_m and i_n interact to affect the resulting aggregate sentiment prediction.

Figures 1 and 2 examine different pairwise combinations of variables, examining the interaction of humidity and temperature, and day of week and hour of day, respectively. Many of the trends observed match intuition about the effect of external variables on sentiment: For example, in Figure 1, 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. Moreover, Figure 2 shows clear diurnal and weekly patterns of sentiment which match strongly with previously observed patterns (Stone et al. 1996; Larsen and Kasimatis 1990; Dodds et al. 2011; Macy 2010).

Important variables Next, we explore the relative importance of individual input variables in predicting aggregate sentiment. We previously explored the relative predictive power of variables classes (e.g., weather and time), but we now take a closer look at the variables within the classes. To do so, we use the common back-

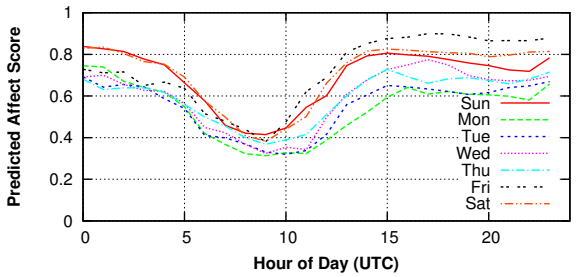

Figure 2: Partial dependence plot of predicted sentiment score based on combinations of day and hour. Note that all times are UTC, so the trough corresponds to 2:00am (EST) and 11:00pm (PST).

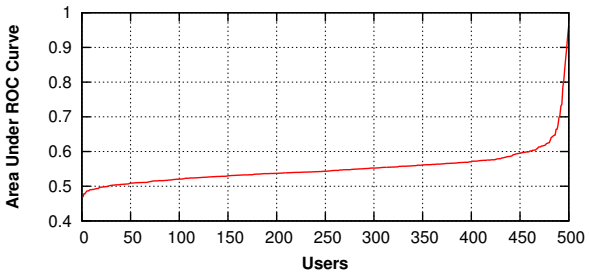

Figure 3: Area under the ROC curve for sentiment predictors built for each of the 500 users we consider. The users are ordered by area under the ROC curve.

wards variable elimination method (Kohavi and John 1997), which starts with the all-variable tree and simply greedily removes the variable that causes the lowest drop in predictive power. The last few remaining variables are the most important.

Table 3 presents the results of this experiment. The table shows that there is not a significant drop in the prediction accuracy while eliminating the first 45 variables (the overall drop in performance is less than 1%). This is likely due to the non-independence of the variables (e.g., dropping the 3-hour humidity still leaves the 2-hour humidity and 6-hour humidity variables). However, the last five remaining variables all demonstrate significant drops in predictive power, suggesting that short-term temporal patterns (hour-of-day and day-of-week), geographic patterns (city), and climate-based patterns (72-hour temperature) dominate.

Predictions for individual users As a final point of evaluation, we examine whether the predictability of sentiment applies to individual users as well (as opposed to only users in aggregate). In other words, can we build a custom predictor for certain individual users and obtain high ROC area for the sentiment of those users' individual tweets? In order to have sufficient data points for an individual, we consider only users who have at least 2,000 tweets in our data set. This leaves us with 11,920 users; to make the analysis below feasible, we select a random sample of 500 of these users.

Step	Variable	Area Under ROC Curve
0	*All*	0.7857
⋮	⋮	⋮
46	Day of month	0.7806
47	Temp. (72h)	0.7751
48	Day of week	0.7532
49	City	0.7376
50	Hour of day	0.6581

Table 3: Order of elimination of variables, showing the five most important variables and A' value before removal.

We repeat the same methodology as above, splitting each users' tweets into a test and training set and creating a bagged tree predictor for each user. Note that instead of predicting the sentiment value of aggregated tweets, we are instead predicting the sentiment value of individual tweets. Also note that we do not include the geography (G) variable, as each user only has a single value for this variable. We then calculate the area under the ROC curve for each user when using all weather (W), time (T), and season (S) variables.

The distribution of the area under the ROC curve for our 500 users is shown in Figure 3. We observe that, on average, the predictability is close to random; the median of the area under the ROC curve values is 0.543. The relatively poor predictability is not surprising, given that we are training our predictor on a much smaller data set with correspondingly much higher variance than the aggregated case. However, there are a few users who show more predictable sentiment: 2.0% of the users possess an area under the ROC curve of 0.7. As future work, we plan on exploring the characteristics of the few users who do show significant sentiment predictability.

Discussion

We now turn to examine a few points of discussion brought up by our analysis in this section.

Not all variables independent As we observed above, there is a high degree of inter-correlation among the variables we consider. For example, knowing the geographic area provides significant information about the weather, and knowing the weather provides significant information about the season. As a result, it is non-trivial to produce variable-specific patterns, as such patterns are very likely to be influenced by other variables. We continue to explore the minimal set of variables that provides the strongest predictive power, looking to see whether an "orthogonal basis" of variables exists.

Using other input variables In our analysis so far, we have focused on input variables including time, season, geography, and climate, primarily due to data

availability. However, our approach can easily be extended to include other variables into the machine learning predictor, such as stock market prices, unemployment rates, and the outcome of sporting events. Including such variables has the potential to aid psychologists and sociologists in the study of population-wide patterns of sentiment.

Related Work

We now detail related work in sentiment analysis, the factors that affect sentiment, and the use of sentiment on Twitter.

Sentiment Analysis

With the emergence of online activities and the growth of virtual communities, determining the sentiment of users is becoming a more attractive mechanism for predicting real-world phenomena. The most common sentiment analysis methods can be divided into two main categories: lexicon-based methods and machine learning-based methods. Lexicon-based methods, such as (Wilson, Wiebe, and Hoffmann 2005; Bradley and Lang 1999; Hu and Liu 2004; Kim and Hovy 2004), calculate the sentiment of the text using a list of words with predefined sentiment scores. Machine learning-based methods unsurprisingly use various machine learning techniques do the classification. Usually, the machine learning algorithm is trained on manually labeled training sets (Turney 2002; Pang, Lee, and Vaithyanathan 2002; Barbosa and Feng 2010; Pang and Lee 2005; Dave, Lawrence, and Pennock 2003), but there have also been approaches to labeling data based on emoticons (Pak and Paroubek 2010). Our approach is one of the first to combine the two types of methods, which enables us to create a larger, more accurate and more comprehensive word list that is better suited to the characteristics of online communication (e.g. the use of emoticons, common abbreviations like LOL, etc).

Effects on sentiment

The effect of weather on sentiment is a well studied topic in psychology. The change of seasons (and specifically the lack of sunshine) can be the cause of different symptoms of depression (Mersch et al. 1999). There have also been studies looking at both positive and negative effects of weather (Denissen et al. 2008; Keller et al. 2005; Rohan and Sigmon 2000). In addition to the known effect of climate on sentiment, researchers found daily (Stone et al. 1996), weekly (Larsen and Kasimatis 1990) and seasonal (Rohan and Sigmon 2000) patterns in the variation of sentiment. We are the first to examine all of these factors in combination.

Leveraging data from a microblogging site like Twitter provides additional benefits. For example, the short status updates on Twitter makes users more likely to frequently report on their status. This results in a broader sample of text, both in the number of subjects and the frequency of measurements, than the small sample research designs that are commonly used in psychology.

Applying Twitter data

The evolution of sentiment analysis has made it possible to examine many aspects of Twitter that are related to sentiment. For example, it has enabled researchers to do studies regarding political sentiment (Tumasjan et al. 2010) and public health (Paul and Dredze 2011), as well as to compare sentiment on Twitter to data gathered from polls (O'Connor et al. 2010). Several works use sentiment analysis to make predictions about the stock market (Bollen, Mao, and Zeng 2010; Eric Gilbert 2010), box-office success (Asur and Huberman 2010), and election outcomes (Tumasjan et al. 2010). Because so many results build on sentiment analysis, it is important to examine the predictability of sentiment itself by studying the effect of hidden factors that are known to influence sentiment in real life.

Conclusion

There has been significant recent interest in using the aggregate sentiment from social media sites like Twitter to try to predict real-world phenomena. However, the aggregated sentiment also offers a unique and—so far—unexplored opportunity to study the effect of external factors on aggregate sentiment, at the scale of a society. In this paper, we took steps in this direction. We first demonstrated that by leveraging the tweets themselves, we can automatically create a token list that is tailored to the peculiarities of Twitter. We make this list and necessary code available to the community, allowing other algorithms to take advantage of these improvements in their predictions.

We then examined the patterns of sentiment that result when using this new list. 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, indicating high prediction accuracy. 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.

References

Asur, S., and Huberman, B. 2010. Predicing the future with social media. `http://arxiv.org/abs/1003.5699`.

Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In *COLING*.

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

Bollen, J.; Mao, H.; and Zeng, X.-J. 2010. Twitter mood predicts the stock market. In *ICWSM*.

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*.

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW*.

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. `http://arxiv.org/abs/1101.5120`.

Eric Gilbert, K. K. 2010. Widespread worry and the stock market. In *ICWSM*.

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*.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning, Second Edition*. Springer.

Hochachka, W. M. 2006. Data-mining discovery of pattern and process in ecological systems. *J. Wildlife. Man.* 71(7).

Hooker, G. 2004. *Diagnostics and extrapolation in machine learning*. Ph.D. Dissertation, Stanford University, Department of Computer Science.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *KDD*.

Keller, M. C.; Fredrickson, B. L.; Ybarra, O.; Cote, S.; Johnson, K.; A., J. M.; Conway; and Wager, T. 2005. A warm heart and a clear head: The contingent effects of weather on human mood and cognition. *Psy. Sci.* 16(5).

Kim, S.-M., and Hovy, E. H. 2004. Determining the sentiment of opinions. In *COLING*.

Kohavi, R., and John, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97(2).

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

Linton, O., and Nielsen, J. P. 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82(1).

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

Macy, M. 2010. Answers in search of a question. In *New Directions in Text Analysis Conference*.

Mathematica weatherdata package. `http://reference.wolfram.com/mathematica/ref/WeatherData.html`.

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*.

Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LERC*.

Panda, B.; Riedewald, M.; and Fink, D. 2010. The model-summary problem and a solution for trees. In *ICDE*.

Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*.

Peng, W., and Park, D. H. 2011. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In *ICWSM*.

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).

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*.

Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL*.

U.s. census metropolitan areas and components. `http://www.census.gov/population/estimates/metro-city/99mfips.txt`.

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

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*.