Measuring and predicting sentiment on Twitter

Aniko Hannak
College of Computer Science
Northeastern University
ancsaaa@ccs.neu.edu

Eric Anderson
Department of Psychology
Northeastern University
anderson.er@husky.neu.edu

Lisa Feldman Barrett
Department of Psychology
Northeastern University
l.barrett@neu.edu

Sune Lehmann
Department of Informatics
Technical University of Denmark
slj@imm.dtu.dk

Alan Mislove
College of Computer Science
Northeastern University
amislove@ccs.neu.edu

Mirek Riedewald
College of Computer Science
Northeastern University
mirek@ccs.neu.edu

ABSTRACT

There has a been significant recent interest in using the aggregate sentiment from social media sites like Twitter to predict real-world phenomena, ranging from the stock market to political polls to the box office success of movies. Due to the massive scalability required to analyze such large amounts of data in near-realtime, most approaches use simple techniques like word lists in order to measure the sentiment of individual messages. However, the accuracy and coverage of such approaches remains unknown, and the resulting patterns of sentiment (which ultimately serve as input to the prediction algorithms) have yet to be closely studied.

Thus, in this paper, we take a closer look at the techniques for measuring sentiment and the resulting patterns, focusing on Twitter and making two contributions. First, we find that existing word lists tend to have poor accuracy and low coverage (meaning there are many tweets they are unable to classify). We therefore use the tweets themselves to create a much more comprehensive word list and demonstrate that this newly created list significantly outperforms existing ones. Second, we examine the patterns of sentiment, exploring whether the well-studied patterns of individuals translates into population-wide trends. We find that aggregate sentiment follows distinct temporal, seasonal, geographic, and climate-based patterns, and that machine learning approaches are able to predict the aggregate sentiment with high accuracy. Our results can inform existing algorithms and suggest that many of the variations in aggregate sentiment are part of repetitive patterns, rather than unique, new information.

1. 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 [3, 21], forecasting the success of movies at the box office [1], and replacing traditional political polling [29, 39] with data taken from Twitter. Due to the massive scalability required to process such large data sets in near-realtime, most of these approaches measure sentiment using lists of positive/negative words and phrases, looking for occurrences in the online postings. While the results of these studies have been impressive, there has yet to be a direct validation of the accuracy of the inferred sentiment that underlies their predictive power.

In addition, psychologists have studied the sentiment of individuals (often referred to as "affect"), and found surprising daily [38], weekly [25], seasonal [37], geographic [27], and climate-related [27] patterns. However, 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 affect. Thus, it remains unclear (a) whether the individual-level patterns translate into population-wide trends, (b) which of the variables dominate the population-wide signal, and (c) how multiple variables interact to influence affect.[1]

In this paper, we take the first steps towards addressing these concerns by focusing on the methods for scalably measuring aggregate sentiment and the resulting patterns. Specifically, we center our analysis on two primary questions:

- First, how accurate are current methods for inferring sentiment using word lists? How close do they come to human-rated scores? Are there ways of constructing more accurate and tailored word lists?
- Second, to what extent do the known patterns of individual affect translate into population-wide patterns? How accurately can the aggregate sentiment itself be predicted?

We focus our results on a collection of over 1.5 billion

[1]Very recent work [11] has observed that patterns of aggregate sentiment do appear to exist in online social networking services, but focuses only on temporal patterns.

messages taken from Twitter, due to its massive popularity and the fact that most users allow their messages to be publicly-visible.

We first examine existing word list-based approaches to inferring sentiment. Given the unique grammar, syntax, and conventions of online services like Twitter, the accuracy of using such lists to measure sentiment is unclear. To evaluate the accuracy of sentiment inference, we create a test set of 1,000 tweets, and each tweet was manually rated by 10 respondents from Amazon Mechanical Turk for sentiment. We find the existing lists suffer from both relatively low accuracy and low coverage (the fraction of messages that can be rated). To address these issues, we construct a new list with positive/negative values learned from the tweets themselves [30, 36], and demonstrate both higher accuracy and significantly greater coverage.

We then examine whether known patterns of individual affect 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 variables. Overall, we find that our machine learning techniques can predict the aggregate sentiment with an ROC area over 0.78, and that the hour of the day, day of the week, geographic region, and recent temperature provide the most information. Our results can inform existing algorithms that make predictions using aggregate sentiment, and suggest that many of the variations in aggregate sentiment are part of repetitive patterns, rather than unique, new information.

The remainder of the paper is organized as follows. Section 2 provides background on Twitter and on the data sets we collect and examine. Section 3 explores the accuracy and coverage of existing word list-based approaches, and develops a new, better performing, list. Section 4 looks at the patterns of sentiment that result when using this new list, leveraging machine learning techniques to predict sentiment. Section 5 details related work and Section 6 concludes.

2. 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; we found that only 8% of users elected to make their account private.

2.1 Twitter data

We obtained data from Twitter using the Twitter API from August 15–September 1, 2009 [7]. 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 2009 tweets leaves us with 1,369,833,417 tweets (90.3% of the entire data set).

2.2 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 [43]. Unless otherwise stated, our analysis for the remainder of this paper is at the metropolitan-area level.

2.2.1 Limitations

We now briefly discuss potential limitations of our location inference methodology. First, it is worth not-

[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.

ing that Google Maps will also interpret locations that are at a granularity coarser than a U.S. county (e.g., "Texas"). We manually removed these, including the mappings of all 50 states, as well as "United States" and "Earth." Second, users may lie about their location, or may list an out-of-date location. Third, since the location is per-user (rather than per-tweet), a user who moves from one city to another (and updates his location) will have all of his tweets considered as being from the latter location. Despite these potential problems, as we demonstrate later in this paper, we are still able to detect a strong signal in the data.

2.3 Weather data

In order to collect weather data, we use Mathematica's WeatherData package [28]. 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).

3. MEASURING SENTIMENT

Sentiment analysis is a well-studied topic, with much recent work focusing on leveraging sentiment expressed on Twitter to predict real-world phenomeona including the stock market [3], movie box office receipts [1], and political polls [29]. In this section, we detail our affect inference methodology and present an evaluation of its accuracy.

3.1 Background

In order to estimate the affect of users on Twitter, we examine the content of their tweets. Ideally, we would like to use existing sentiment analysis techniques [23, 33, 40]. 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.

As a result, most prior Twitter sentiment analysis work has focusing on *word lists*, containing a set of words with a sentiment score attached to each. Examples of such lists include the Affective Norms for English Words (ANEW) list [6], the list by Wilson, Wiebe, and Hoffmann (WWH) [42], and the list by Hu and Liu (HL) [16]. Existing approaches generally look for occurrences of listed words in tweets, taking the average on the individual word sentiment scores to the be sentiment score of the entire tweet. In more detail, if a tweet contains n words that are present in the word list and their valence scores are $\{v_1, v_2, \ldots, v_n\}$ and the frequency of each of these words in the tweet is $\{f_1, f_2, \ldots, f_n\}$, the valence score of the tweet is calculated by the weighted mean of the scores

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

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 words), the fraction of tweets that contain at least one listed word is often small. Fourth, using a word list to extract sentiment is a simple approach that is likely to make incorrect affect assessments on certain tweets. Thus, it is surprising that there has yet to be an analysis of the accuracy of this methodology, as it is unclear how well this naïve approach works, or how accurate the various lists are relative to each other.

3.2 Methodology

In order to address the challenges above, we construct a Twitter-specific word list by using the tweets themselves. In brief, we consider only tweets that contain exactly one of the emoticons `:)`, `:-)`, `:(`, `:-(`, as the emoticons often represent the true sentiment of the tweet [41]. We then look at the tokens (words) that occur in these tweets, and calculate the fraction of time each token appears with one of the happy emoticons. This results in a word list with a weighting for each word, where the weighting indicates the propensity for the word to appear with happy-emoticon-tagged tweets. It is important to note that a similar methodology has been proposed in prior work [30, 36], but has not yet been evaluated at scale on a large text corpus, nor has its accuracy been measured.

In more detail, we start with the collection of all 1.7 billion 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. This

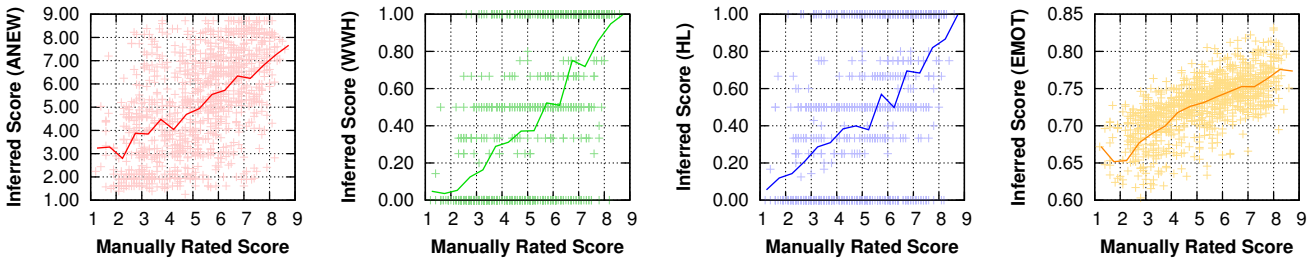

Figure 1: Scatterplot of inferred affect score for various lists versus the average AMT affect score. The solid line corresponds to the average of buckets of size 0.5. While all lists show a correlation, the EMOT list demonstrates a much tighter distribution.

narrows our tweet collection to 591,406,152 tweets. We only consider tweets that have exactly one of the four emoticons above; this results in 15,668,367 tweets with a happy emoticon and 5,237,512 tweets with a sad 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. Finally, we ignore any token that did not appear at least 20 times, giving us a final list with 275,193,529 occurrences of 75,065 unique tokens.

To create our token list (referred to as EMOT), we calculate the relative fraction of times the token occurs with a happy emoticon and use this as the token's score. For example, the token `relaxing` occurred in 39,584 tweets with happy emoticons and 3,439 tweets with sad emoticons, giving `relaxing` a score of 0.9201. A sample of a few of the elements of this list are provided in Table 1.

Relative to the existing word lists, the advantages of this approach are four-fold. First, the creation of the list is automated from the Twitter data alone; creating word lists such as the ANEW list [6] required manually surveying people. Second, the resulting list is substantially larger than existing lists, giving much greater coverage. Third, the list is specific to Twitter, automatically capturing abbreviations, nelogisms, and Twitter-specific syntax. Fourth, our methodology can easily be extended to support other languages (by simply using a different dictionary during the filtering phase) or the language and customs used for specific topics (by simply narrowing the list of tweets to those containing a given hashtag of interest).

Token	Occurrences Happy	Sad	Score
`:((`	540	1,415	0.2761
`sore`	10,803	23,372	0.3161
`offices`	1,555	537	0.7434
`table`	10,513	3,215	0.7658
`charming`	1,938	258	0.8821

Table 1: Sample of tokens and inferred sentiment scores from our token list. Note that non-affective tokens `offices` and `table` have scores close to the overall fraction of happy tweets of 0.749.

3.3 Evaluation

We now examine the accuracy of our inferred token list and compare our accuracy to existing word lists. To do so, we create a list of manually, human-rated tweets using Amazon Mechanical Turk (AMT) [2]. We created AMT tasks (called *human intelligence tasks* or *HITs*), paying the Turk users $0.10 to rate the affect of 10 tweets. A screenshot of the HIT that we used is shown in Figure 2. The text and response input used in the HIT was modeled after the survey used in the creation of the ANEW [6] word list.

We created a test set consisting of 1,000 tweets. In order to ensure variety on this relatively small test set, we selected 125 random tweets with affect scores in each of the ranges {1–2, 2–3, ..., 8–9} when scored using the ANEW list. 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 calculated an average AMT affect score for each tweet. We found that the AMT results showed

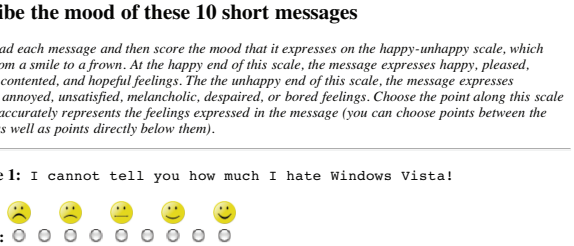

Figure 2: Screenshot of the AMT task for rating tweets. The instructions are very similar to those used in the original ANEW study [6]. Each task consisted of rating 10 tweets.

Word List	Correlation Pearson	Correlation Spearman	Hit Rate
ANEW	0.5233	0.5266	39.6%
WWH	0.5784	0.5744	41.4%
HL	0.4810	0.4764	49.2%
EMOT	**0.6510**	**0.7046**	**89.9%**

Table 2: Pearson and Spearman correlation coefficient area when comparing automatically inferred scores to manually rated scores. Also shown is the hit rate on random tweets, or the fraction of tweets each list is able to score. Our EMOT list shows superior performance to existing lists with more a significantly increased hit rate.

a strong inter-respondent Pearson correlation of 0.784,[4] which is in line with the results from other studies using AMT respondents [34].

We now evaluate the accuracy of the various word lists using this list of 1,000 AMT affect-scored tweets. Figure 1 shows the scatterplot of the automatically inferred affect score when using each of the four word lists versus the average AMT affect score; the dark line depicts the average value of buckets of size 0.5. All of the word lists show a correlation with the AMT score, with the EMOT word list demonstrating a significantly"tighter" distribution. To quantify the relative performance of the various word lists, we calculate two measures, shown in Table 2. First, we calculate the Pearson correlation coefficient between the inferred and AMT scores, measuring the tendency of the inferred score to increase when the AMT score increases. All show a positive correlation, with EMOT outperforming the others by a substantial margin. Second, we compare the rank ordering of the various word lists by calculating Spearman correlation coefficient. Again, the EMOT list demonstrates substantially improved performance.

As a final point of evaluation, we calculate the *hit rate* of the word lists, defined by the fraction of all tweets which the word list is able to score (i.e., the fraction of tweets that contain at least one word on the list). To do so, we selected a random subset of 350,000 tweets and scored each tweet with all four lists. Also shown in Table 2, the EMOT list can score almost 90% of the tweets, while no other list can score over 50%. This improvement is largely due to the much greater size of the EMOT list, a property afforded by the list creation methodology.

We make this resulting word 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`.

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

4. PREDICTING SENTIMENT

With our Twitter-specific word list in hand, we now turn to examine the patterns of sentiment that exist. In particular, psychologists have found distinct daily [38], weekly [25], seasonal [37], geographic [27], and climate-related [37] patterns of individual affect. However, these studies have been limited in scale by their methodology, as they often rely on repeated surveys. The largest of these studies examine only a few hundred subjects. These studies are also generally limited to studying the effect of a single variable on affect. As a result, it is unclear whether the individual-level patterns translate into population-wide trends, which of the variables dominate the population-wide signal, and how multiple variables interact to influence affect.

To understand the patterns of sentiment on Twitter, 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.

4.1 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 all tweets into hour-long buckets according to both time and geography. In more detail, for each of the 20 largest 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 the sentiment score for the bucket.

We chose to use bagged decision trees [4] 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 [8]. 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 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 test set with replacement [12]. 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 real-valued number (the average of all tweet sentiments) to a binary happy (1) or sad (0) value. The cutoff for the happy/sad 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 happy-versus-sad thresholds, or even use multiple levels on the scale from sad to happy. However, a full exploration is beyond the scope of this paper.

4.2 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 a our predictor ranks two periods in their true relative order [13]. 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.

4.3 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).

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

Table 3: 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.

4.4 Results

We now turn to examine the effectiveness of 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.

4.4.1 Prediction performance

We constructed 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 presented in Table 3.

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 4, 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

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 4: 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.

0.5. This result indicates that the well-studied patterns of individual affect do indeed result in trends of aggregate sentiment, and can even be predicted with high accuracy.

4.4.2 *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 [14,15,17,18,24,31], which visualize partial dependence functions. A partial dependence function for a given multi-dimensional function $f(\mathbf{X})$[5] 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.

[5]Note that $\mathbf{X}$ is a vector of multiple input variables.

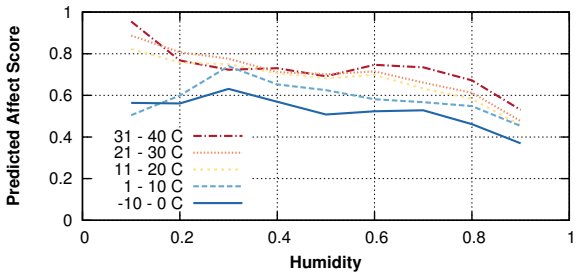

Figure 3: Partial dependence plot of predicted affect score from the all-variable bagged tree, based on different combinations of humidity and temperature. As humidity increases the predicted affect score decreases (with a more pronounced effect at higher temperatures), matching intuition.

Figures 3 and 4 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 affect: For example, in Figure 3, as the humidity increases, the predicted affect 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 affect when the temperature is higher. Moreover, Figure 4 shows clear diurnal and weekly patterns of sentiment which match strongly with previously observed patterns [11, 25, 26, 38].

4.4.3 *Important variables*

As a final point of evaluation, 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 backwards variable elimination method [19], 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 5 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

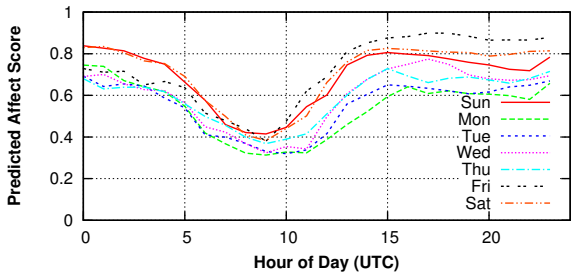

Figure 4: Partial dependence plot of predicted affect score based on combinations of day and hour. Note that all times are UTC, so the trough corresponds to between 2:00am (EST) and 11:00pm (PST).

patterns (72-hour temperature) dominate.

4.5 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 in Section 4.4.3, 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.

Predictions using aggregate sentiment Recall that a significant amount of recent work has explored using aggregate sentiment to predict real-world phenomena. Our work here has demonstrated that aggregate sentiment is exceedingly regular, and can be predicted to a significant degree. Thus, it is likely to be enlightening to examine whether previously-explored predictors work well when provided with our input variables alone (instead of aggregate sentiment). If their predictive performance holds up, it suggests that aggregate sentiment is not what underlies their predictive power; if not, it suggests that existing approaches are able to separate out the repeating patterns of sentiment and leverage on the heretofore-unpredictable variations.

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

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.

5. RELATED WORK

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

5.1 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 [6,16,20,42], 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 [5,9,32,33,40], but there have also been approaches to labeling data based on emoticons [30]. 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).

5.2 Effects on sentiment

The effect of weather on affect 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 [27]. There have also been studies looking at both positive and negative effects of weather [10, 22, 37]. In addition to the known effect of climate on affect, researchers found daily [38], weekly [25] and seasonal [37] 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 number of subjects and frequency of measurements, than the small sample research designs that are commonly used in psychology.

5.3 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 [39] and public health [35], as well as to compare sentiment on Twitter to data gathered from polls [29]. Several works use sentiment analysis to make predictions about the stock market [3, 21], box-office success [1], and election outcomes [39]. Because so many results build on sentiment analysis, it is important to examine the predictability of affect itself by studying the effect of hidden factors that are known to influence affect in real life.

6. 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. While most of the research so far has focused on the predicted phenomena themselves, in this paper, we took a closer look at the sentiment inference algorithms that underlies these approaches' predictive power. Specifically, we examined a popular sentiment technique that leverages positive/negative word lists for inferring sentiment. Due to the massive amount of data that is being produced, this technique represents a scalable and efficient approach for inferring sentiment.

However, the accuracy and coverage of the word list approaches remains unknown. We found that existing lists only cover between 39% and 49% of tweets; this is largely because the lists are generally manually created, are not site-specific, and do not account for abbreviations, neologisms, and unique syntax. We demonstrated that by leveraging the tweets themselves, we can automatically create a new word list that has substantially increased accuracy and coverage. We make this list and necessary code available to the community, allowing other algorithms to take advantage of these improvements in their predictions.

Additionally, we 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 over 78% predictable. These results can inform existing algorithms, and suggest that many of the variations in aggregate sentiment are part of repetitive patterns, rather than unique, new information.

7. REFERENCES

[1] S. Asur and B. Huberman. Predicing the Future with Social Media. Mar. 2010. `http://arxiv.org/abs/1003.5699`.

[2] Amazon Mechanical Turk. `http://www.mturk.com/`.

[3] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *ICWSM*, Washington, D.C., May 2010.

[4] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, Aug. 1996.

[5] L. Barbosa and J. Feng. Robust Sentiment Detection on Twitter from Biased and Noisy Data. *COLING*, Beijing, China, Aug. 2010.

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

[7] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM*, Washington, D.C., May 2010.

[8] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. *ICML*, Pittsburgh, PA, June 2006.

[9] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *WWW*, Budapest, Hungary, May 2003.

[10] J. J. A. Denissen, L. Butalid, L. Penke, and M. A. G. van Aken. The effects of weather on daily mood: A multilevel approach. *Emotion*, 8:662–667, Oct. 2008.

[11] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. Jan. 2011. `http://arxiv.org/abs/1101.5120`.

[12] T. G. Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40(2):139–157, July 2000.

[13] J. Fogarty, R. S. Baker, and S. E. Hudson. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *GI*, Vancouver, Canada, May 2005.

[14] J. H. Friendman. Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, 29:1189–1232, 2001.

[15] G. Hooker. Diagnostics and extrapolation in machine learning. Ph.D. Thesis, Stanford University, Department of Computer Science, May 2004.

[16] M. Hu and B. Liu. Mining and summarizing customer reviews. *KDD*, Seattle, WA, Aug. 2004.

[17] T. Hastie, R. Tibshairni, and J. Friedman. *The Elements of Statistical Learning, Second Edition.* Springer, 2009.

[18] W. M. Hochachka. Data-mining discovery of pattern and process in ecological systems. *J. Wildlife. Man.*, 71(7):2427–2437, 2006.

[19] R. Kohavi and G. H. John. Wrappers for Feature

Subset Selection. *Artificial Intelligence*, 97(2):273–324, 1997.
[20] S.-M. Kim and E. H. Hovy. Determining the Sentiment of Opinions. *COLING*, Geneva, Switzerland, Aug. 2004.
[21] E. G. K. Karahalios. Widespread Worry and the Stock Market. *ICWSM*, Washington, D.C., May 2010.
[22] M. C. Keller, B. L. Fredrickson, O. Ybarra, S. Cote, K. Johnson, J. M. A., Conway, and T. Wager. A warm heart and a clear head: The contingent effects of weather on human mood and cognition. *Psy. Sci.*, 16(5):724–731, May 2005.
[23] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer-Verlag, New York, NY, 2006.
[24] O. Linton and J. P. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93–100, 1995.
[25] R. J. Larsen and M. Kasimatis. Individual differences in entrainment of mood to the weekly calendar. *J. Per. & Soc. Psych.*, 58(1):164–171, Jan. 1990.
[26] M. Macy. Answers in Search of a Question. *New Directions in Text Analysis Conference*, Cambridge, MA, May 2010.
[27] P. P. A. Mersch, H. M. Middendorp, A. L. Bouhuys, D. G. M. Beersma, and R. H. van den Hoofdakker. Seasonal affective disorder and latitude: a review of the literature. *J. Aff. Disord.*, 53(1):35–48, Apr. 1999.
[28] Mathematica WeatherData Package. `http://reference.wolfram.com/mathematica/ref/WeatherData.html`.
[29] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM*, Washington, D.C., May 2010.
[30] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. *LERC*, Valletta, Malta, May 2010.
[31] B. Panda, M. Riedewald, and D. Fink. The Model-Summary Problem and a Solution for Trees. *ICDE*, Long Beach, CA, Mar. 2010.
[32] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ACL*, Ann Arbor, MI, June 2005.
[33] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. *EMNLP*, Philadelphia, PA, July 2002.
[34] W. Peng and D. H. Park. Generate Adjective Sentiment Dictionary for Social Media Sentiment Analysis Using Constrained Nonnegative Matrix Factorization. *ICWSM*, Barcelona, Spain, July 2011.
[35] M. J. Paul and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. *ICWSM*, Barcelona, Spain, July 2011.
[36] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *ACL*, Ann Arbor, MI, June 2005.
[37] K. J. Rohan and S. T. Sigmon. Seasonal mood patterns in a northeastern college sample. *J. Aff. Disord.*, 59(2):85–96, Aug. 2000.
[38] A. A. Stone, J. M. Smyth, T. Pickering, and J. Schwartz. Daily Mood Variability: Form of Diurnal Patterns and Determinants of Diurnal Patterns. *J. App. Soc. Psych.*, 26(14):1286–1305, July 1996.
[39] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*, Washington, D.C., May 2010.
[40] P. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *ACL*, Philadelphia, PA, July 2002.
[41] C. Vogel and J. Janssen. Emoticonsciousness. *Multimodal Signals: Cognitive and Algorithmic Issues, LNAI 5398*, chapter 2, pages 271-287, Anna Esposito, Amir Hussain, Maria Marinaro, and Raffaele Martone, eds., Springer-Verlag Publishers, Berlin, Germany, 2009.
[42] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. *HLT/EMNLP*, Vancouver, Canada, Oct. 2005.
[43] U.S. Census Metropolitan Areas and Components. `http://www.census.gov/population/estimates/metro-city/99mfips.txt`.