OMG, I Have to Tweet That!
A Study of Factors that Influence Tweet Rates

Emre Kıcıman
Microsoft Research
emrek@microsoft.com

Abstract

Many studies have shown that social data such as tweets are a rich source of information about the real world including, for example, insights into health trends.

A key limitation when analyzing Twitter data, however, is that it depends on people self reporting their own behaviors and observations. In this paper, we present a large scale quantitative analysis of some of the factors that influence self reporting bias. In our study, we compare a year of tweets about weather events to ground truth knowledge about actual weather occurrences. For each weather event we calculate how extreme, how expected, and how big a change the event represents. We calculate the extent to which these factors can explain the daily variations in tweet rates about weather events. We find that we can build global models that take into account basic weather information, together with extremeness, expectation and change calculations to account for over 40% of the variability in tweet rates. We build location specific (*i.e.*, a model per each metropolitan area) models that account for an average of 70% of the variability in tweet rates.

Introduction

Many recent studies have shown that large-scale social media analysis is a rich source of information about real-world events and trends. For example, Paul and Dredze (2011) analyze Twitter to better understand trends in health, such as what drugs are used to treat common ailments. Others have used similar analyses of social media to predict box office ratings of movies and detect occurrences of earthquakes (Asur and Huberman 2010; Sakaki, Okazaki and Matsuo 2010). Each of these studies, in essence, is treating social media as a signal to measure the relative real-world occurrence of events. For example, if more people tweet about flu symptoms this week as compared to last week then we will believe the real-world occurrence of the flu has increased.

A critical challenge to using social data to compare and contrast real-world events and trends, however, is the bias introduced by the self-reported nature of social media. It is commonly understood that the frequency of discussion about events on social media does not directly reflect the true frequency of event occurrence in the real-world, but instead is a complex function of individual experience combined with what each individual believes would interest their friends or followers (Java et al. 2007; Naaman, Boase and Lai 2010). As an extreme example, very few people ever tweet about simple occurrences such as breathing and drinking water, and yet these are obviously much more frequent than highly tweeted events such as natural disasters and celebrity sightings.

The implications of this self-reporting bias are significant. Researchers can assume only that this bias is constant for a given kind of event, but can say little about the bias across different kinds of events. As a result, researchers study trends in an event's occurrence across time and across geographies but can make few inferences about the relationship between distinct events through social media analysis.

Unfortunately, this prevents us from addressing many interesting questions: Do more people enjoy playing ball at the neighborhood park or jogging around the track? (Or are people simply more likely to tweet about ball games because it's a group activity?) Are STDs a smaller health problem than other diseases? (Or do people avoid tweeting about STDs out of embarrassment?) Are people who play video games more or less likely to watch a movie in the cinema? Each of these is a question that we might fruitfully address using social media data; and the answers have the potential to influence important decisions being made by individuals, governments, corporations and other organizations about priorities, policies and spending. Unfortunately, each of these questions requires us to

analyze the relative frequencies and relationships in people's self-reporting behavior and cannot be answered without a better quantitative understanding of self-reporting bias: what is it about an event that makes it more or less "tweetable"?

This paper provides a first large-scale, quantitative analysis of some of the factors that influence self-reporting bias by comparing a year of tweets about weather events in cities across the United States and Canada to ground-truth knowledge about actual weather occurrences. In our study, we focus on three potential factors that seem likely to affect tweet rates: How extreme is the weather? How expected is the weather given the time-of-year? How much did the weather change?

We find that we can build global models that take into account basic weather information, together with extremeness, expectation and change calculations to account for over 40% of the variability in tweet rates. We build location-specific (*i.e.*, a model per each metropolitan area) models that account for an average of 70% of the variability in tweet rates.

It is worth noting that our goal is not yet to find techniques or analyses to compensate or reverse-engineer any self-reporting bias in social media. Instead our primary contribution is to present a methodology for the quantitative analysis of such bias, so that we can now begin to understand the extent to which it exists and the factors that influence it.

In the rest of this paper, we first review related work, followed by a description of the processing and preparation of our Twitter and weather report datasets. We then present our experiments and results, discuss open issues and conclude.

Related Work

Over the last several years, much research has focused on understanding real-world events and trends through studying of the digital footprints of human activity. Earlier work began with studies of search engine query logs, such as Google Flu Trends analysis of query logs to detect rates of influenza (Ginsberg et al. 2008). This use of search query logs is well characterized in Goel et al.'s (2010a) study of the predictive value of search query logs compared to domain-specific data sets, and Goel et al (2010b)'s study on predicting consumer behavior through search queries.

The use of large-scale analysis of digital activities to predict or characterize real-world phenomena has extended to analysis of social media, such as Twitter and Facebook. In addition to the passive analysis of social media to understand health trends, predict box office returns, and detect earthquakes, Sheth (2009) proposes modeling humans as being actively in-the-loop of a *citizen sensor* network. Similar analyses are now being made of location-based social networks—also reliant on self-reporting of location via check-ins—to better understand patterns of human mobility for traffic forecasting, urban planning and other applications (Cheng et al. 2011). All of these works demonstrate a need for a better understanding of the self-reporting bias inherent in social media.

While not studied in the context of social networks, self-reporting bias has been studied in other contexts. Donaldson and Grant-Vallone (2002) report on self-reporting bias in the context of organizational behavior research, and suggest that not properly accounting for such bias can lead to misleading empirical results. In this context, self-reported data is argued to be suspect because it is subject to response biases as well as method biases (Podsakoff, Mackenzie, Lee and Podsakoff 2003). While a common technique to mitigate the biases is to gather self-reported data through the experience sampling method (*e.g.*, through active polling of a respondent through mobile phones, as in Intille et al. 2003), such techniques are not applicable in passive analysis of social media.

Data Preparation

We derive our weather-related social media data set from a full archive of 12 months of Twitter captured between Jun 1, 2010 and Jun 30, 2011—because of data corruption, we are not using Twitter data captured during Mar. 2011. From this data set, we compute the daily rate of weather-related tweets in 56 different metropolitan areas. We compare these tweet rates to weather features of extremity, expectation and change calculated from historical weather data provided by the National Oceanic and Atmospheric Administration of the United States.

In this section, we describe how we identified weather-related tweets and associated them with specific cities, as well as how we compute features about ground-truth weather information.

Identifying Weather-related Tweets

For our analysis in this paper, we are interested in discovering the rate of weather-related tweets that occurred per-day across metropolitan areas. We will not attempt to distinguish tweets about different kinds of weather.

We start by filtering the full archive of tweets for tweets that contain at least 1 weather-related word from a list of 179 weather-related words and phrases. This list was built by hand from weather glossaries, augmented with synonym data derived from search queries (*e.g.*, queries clustered by co-clicks) and a previous LDA analysis of tweets. This first step extracts a super-set of weather-related tweets (~130M messages).

Next, we use a supervised learning technique to build a classifier for weather-related tweets. Two annotators hand labeled a random sample of 2000 tweets as either being or not being weather-related. Our labeling criteria specified that a tweet should be likely to be a report or comment on current weather conditions. We use 1800 labeled tweets for training our classifier and 200 labeled tweets for validation. Table 1 shows example weather-related and non-weather-related tweets.

Table 1: Example weather-related tweets and non-weather related tweets from our super-set selection. The original weather-related term is shown in bold.

Tweet Text	Weather-related?
Woke up to a **sunny** 63F (17C) morning. It's going to be a good day :)	Yes
Japan, Germany **hail** U.N. Iran sanctions resolution+	No
The **rainy** season has started.	Yes
The inside of our house looks like a **tornado** came through it.	No

We use a simple classifier that estimates the probability of a tweet being weather related as

$$\frac{1}{|T|}\sum_{t\in T} P(weather|t)$$

where T is the set of features derived from the tweet text, $P(weather|t)$ is the empirical probability of a tweet being weather-related given that it contains a feature t. We calculate $P(weather|t)$ empirically from our labeled training data using +1 smoothing:

$$P(weather|t) = (1 + C(weather|t))/(1 + C(t))$$

where $C_{weather}(t)$ and $C(t)$ are, respectively, the count of weather-related tweets containing the feature t and the count of all tweets containing the feature t in our training data. To generate features for a tweet, we first apply a simple stemmer to the text that removes all -s and -ing suffixes from words. Then we generate a feature t to represent every unigram token in the stemmed tweet text. We also generate a feature t for each pair of tokens co-occurring in the tweet. With these features, our classifier distinguishes weather-related tweets by learning, for example, that the words "seasons", "outside", and "humidity" are correlated with weather-related tweets, whereas words such as "health" and "clear" are signals that tweets are not weather-related. Using co-occurrences of tokens allows the classifier to learn that the word "heat" co-occurring with "score", "play" and other basketball related words is unlikely to be weather related (These tweets are instead more likely mentioning the Miami Heat, a basketball team).

Our classifier achieves an F-Score of 0.83, with a precision of 0.80 and recall of 0.85. After using our classifier to filter our super-set of tweets, we are left with approximately 71M tweets classified as being weather-related. We also experimented with variations of this classification technique, including using a more sophisticated Porter stemmer, and using n-gram features instead of co-occurrences. We found the Porter stemmer to provide minimal improvement, and found features based on token co-occurrences to be significantly better than bi-gram and tri-gram features.

Identifying the Location of Tweets

In order to compare a weather-related tweet to ground-truth weather information, we must identify the general geographic area that the tweet is referencing. Since only a small percentage of tweets are explicitly geo-coded, we use the textual user-provided location field in a user's Twitter profile to identify the region-of-interest for a tweet.

Unfortunately, the user-provided profile locations are in general not easily interpretable (Hecht et al. 2011). The same location may be referred to using multiple names (*e.g.*, "New York City", "Manhattan", "NYC"), or identified at different granularities (*e.g.,* a broad name such as "NY/NJ", a medium-granularity name such as "Brooklyn, NY" or a fine-grained name such as "Bushwick, Brooklyn"). In some cases, the location is a nickname for a location (*e.g.,* "D{M}V" for the DC-Maryland-Virginia area, or "The Big Apple" for New York City) or a nonsensical phrase (*e.g.,* "everywhere" or "none of your business").

Our desire is to normalize the textual and sometimes arbitrary user-provided location information into concrete geo-coded coordinates that are more useful in analyses. To do so, we analyze 1 month of the full Twitter archive to find tweets that include both a user-provided location field and explicitly geo-coded coordinates. We use this subset of tweets to learn a mapping from user-provided location fields to latitude-longitude coordinates.

We first capture the distribution of geo-coded points associated with each unique location field in our data-set. Through manual inspection, we find that the median geo-coded point provides the most accurate mapping to the location field. Alternatives, such as selecting the center of a Gaussian inferred from the geo-coded points, are often inaccurate due to outliers in the location fields.

In addition, we discard location fields that cover too broad an area or were too ambiguous (based on the distribution of geo-coded points associated with a field), or had too little support in our data set. Then, we merge location fields with similar geo-mappings together to

create clusters for roughly metropolitan-sized areas. Table 2 show a sample of the location fields clustered together, as well as locations that we discarded as being too ambiguous. In total, we are able to infer a geo-coding mapping for over 13,000 location fields and over 2,700 distinct clusters around the world.

We apply this location mapping to our set of weather-related tweets and discard locations that do not have a sufficiently high number of tweets (~100s of tweets per day), and limit our analysis to locations occurring in the United States, to match the availability of weather stations in our dataset. To better compare tweet rates across different locations, we normalize the rate of weather-related tweets in a particular location by the median count of weather-related tweets made in that location. The result is a dataset of the daily weather-related tweet rate for 56 metropolitan areas in the United States, with over 8M weather-related tweets during our 12 month analysis.

Table 2: Sample of discovered Twitter location clusters

Location cluster	Example members
New York	"NYC", "Yonkers", "manhattan," "NY,NY", "Nueva York", "N Y C", The Big Apple"
Los Angeles	"Laguna beach", "long beach", "LosAngeles,CA", "West Los Angeles, CA", "Downtown Los Angeles", "LAX"
Filtered out due to ambiguity (large area)	"World", "everywhere", "USA", "California", …

Historical Weather Data

Our historical weather data consists of several datasets aggregated and distributed by the National Oceanic and Atmospheric Administration (NOAA) of the United States. Our primary weather data set is derived from hourly weather reports that include temperature, wind, atmospheric pressure, precipitation, and other occurrence information for discrete events (*e.g.*, tornados, fog, and thunderstorms). In some of our calculations of expectation, we also use the 30-year weather norms provided by NOAA. This norms dataset provides statistical information (average and std.dev.) about temperature and precipitation for every day of the year.

From the hourly weather reports, we calculate daily summaries that include the daily minimum, mean and maximum observed temperatures, maximum observed wind speeds, daily precipitation rates, snowfall, visibility, cloud height, etc. In total, we collect 9 continuous measures of daily weather at a location, as well as a list of discrete weather events that occurred. For each daily summary of weather data at a location, we calculate how extreme and expected the weather is along the given dimension, as well as how much it has changed in comparison to previous days.

Expectation: Expectation captures how normal the observed weather is at a location, given the context of the time-of-year. For this, we use the 30-year weather norms dataset provided for temperature and precipitation data, and calculate where the observed weather falls in the percentile distribution for the given day. Unfortunately, the weather norms dataset does not include norms for other weather measures or for weather events.

This measure is intended to identify events that are interesting and "tweetable" because they are unusual given the time context of when they occur, though they might be considered to be normal events in other contexts. For example, a warm winter day might be considered unusual and thus interesting, even though the temperature would be considered normal in the spring. Note that because we do not have weather forecast data (only weather measurements), we do not attempt to capture the surprise when a weather forecast turns out to be incorrect.

Extremeness: Our measure of how extreme the weather is on a particular day is intended to capture where the day's weather lies between being, for example, the hottest or the coldest day of the year. We calculate extremeness at multiple time scales, comparing weather observations to the previous 1 month, 3 months, 6 months and 12 months of data.

This measure of extremeness is intended to identify events that are interesting and "tweetable" because they occur infrequently, even though they might be expected given the context. For example, a hot day in mid-June might be expected because of the summer season, and yet still be considered extreme because it is so much hotter than other days throughout the year.

Change: Our measure of change captures how different the observed weather data is from previous days' weather. For each day, we compare the given measure to the weather from 1 day, 3 days and 7 days earlier.

This measure is intended to identify events that are interesting and "tweetable" because of their contrast with recent weather history. For example, a slightly colder day in the spring might not be unexpected given historical norms, nor extreme given the annual weather information, but might still be an interesting event if it suddenly following a series of warm days.

Calculating these abstract features of extremeness, expectation and change serves two purposes. First, it acts as a normalization of weather data across locations. While

110°F may be a record-breaking high temperature in some locations, it might be just a normal summer day in other locations. Secondly, it provides a simpler abstraction intended to bridge the underlying multitude of weather with more intuitive notions of what may make an event more interesting to followers and friends in a social network.

Analysis and Results

Tweet Rates and Weather Reports

In this subsection, we take a basic, "zoomed in" look at the kind of correlation we are looking for between tweet rates and weather reports. Figure 1 shows the weather-related tweet rate and daily temperature for San Diego, CA during the period of Sep. 1-Oct. 15, 2010. Here, we see two peaks in the tweet rate, on the hottest day in this period (Sep. 27) and the first thunderstorm of the season (Sep. 30). We can verify that these peaks of tweets are related to the weather events by inspecting the words being used in tweets. Here, we clearly see the relationship: on Sep. 27, the top tokens being tweeted are "weather", "heat", "temperature", "hotter" and "hottest". A few days later, on Sep. 30, the top tokens being tweeted are "rain", "weather", "lightning" and "thunder".

While this example demonstrates a relationship between the content of weather-related tweets and the actual weather, it hints at a more complex relationship between the weather and the *rate* at which there are tweets about the weather. For example, the local peak in daily temperature on Sep. 3 seems to have little impact on the tweet rate. In the rest of this analysis, we will try to tease out some of this more complex relationship.

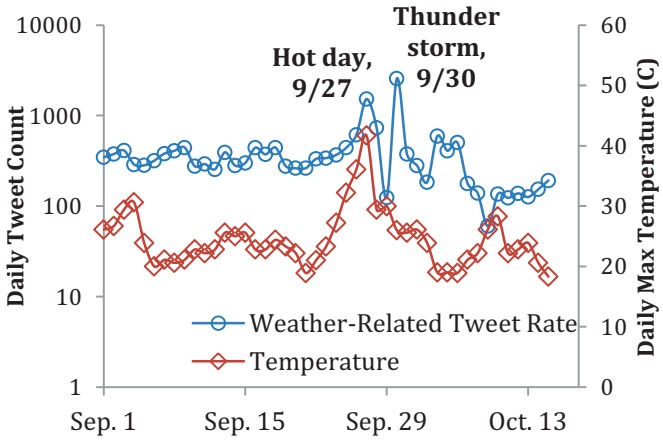

Figure 1: Weather-related Tweet rate and temperature in San Diego, CA from Sep. 1-Oct 15, 2010

Linear Regression

To more deeply explore the relationship between a set of weather-derived features and the daily rate of weather-related tweets, we learn a linear model:

$$r = a + \sum_i b_i w_i$$

where r is the rate of weather-related tweets, and w_i are features derived from the ground-truth weather observations. The w features include the daily weather observations, expectation, extremeness, and change features. We also include the absolute value and squared value of these features. We fit the model parameters a and b_i using a least-squares regression with L2 regularization. We solve the model using an off-the-shelf implementation of an L-BFGS algorithm provided in Microsoft Solver Foundation, a library for mathematical programming and optimization.

Unless otherwise noted, we fit a global model across all our data (~365 days * 56 cities). For each model, we report the R^2 correlation between the predicted value $\hat{r}$ and the observed weather-related tweet rate, $\bar{r}$. Table 3 summarizes the results of the global models we build throughout the rest of this section.

Correlating Basic Weather Data and Tweet Rates

As a baseline, we first build a model to predict weather-related tweet rates per city based solely on observed

Table 3: R^2correlations for our various global models

Model Features	Parameter	R^2 correlation
Basic weather	-	0.30
Expectation Only	-	0.12
Expectation+Basic	-	0.33
Extremeness Only	1 mo. comparison	0.20
	3 mo. comparison	0.24
	6 mo. comparison	0.26
	12 mo. comparison	0.30
	All	**0.36**
Extremeness+Basic	All	**0.40**
Change Only	1 day	0.21
	3 day	0.23
	7 day	0.21
	All	**0.28**
Change+Basic	All	0.35
Basic+Extremeness+ Change	All	**0.43**
Expectation+ Extremeness+Change	All	**0.42**
Basic+Expectation+ Extremeness+Change	All	**0.43**

weather data (*e.g.*, temperature, precipitation but not derived features of extremeness, and expectation).

We find that, across all our data points, the R^2 correlation for this basic model is 0.30. That is, 30% of the variability in daily tweet rates can be accounted for by variations in basic weather data. Inspecting the learned model parameters, we find that the regression model has placed much of its weight on parameters related to the maximum temperature, wind speed, snowfall and precipitation. Other features, including the visibility, cloud height, and pressure receive much less weight.

Figure 2 shows the per-location R^2 correlations for this model. We see that this simple model serves some locations well—at the high end, our models have over a 0.45 correlation with Washington, D.C. and Knoxville, TN. However, the model completely fails in predicting tweet rates for several other cities, achieving a negative correlation rate for 4 locations, including San Francisco and Puerto Rico.

Correlating Expectation and Tweet Rates

Next, we study the relationship between our measures of expectation, how normal weather is given the location and time-of-year context, and tweet rates. We build linear regression models using expectation and basic weather data together, as well as expectation data alone. In the latter case, we find an R^2 value of only 0.12, and in the latter, an R^2 value of 0.33. This indicates that our expectation measure adds little information about likely tweet rates beyond what is already contained in basic weather data. This implies that the likelihood of tweeting is not affected strongly by how normal the weather is given the time of year.

Correlating Extremeness and Tweet Rates

To study the relationship between extremeness and tweet rates, we build several independent models to explore the effect of calculating extremeness over the 1-month, 3-month, 6-month and 12-month time-frame affects the correlation of our models with tweet rates (Table 3). Overall, we find that extremeness can independently explain more of the variation in weather-related tweet rates than basic weather alone. Calculating extremeness over a longer period (12-month) is more informative than calculating it over shorter time periods (6 mo. or 3 mo.). We also find there is a large overlap between the information content of extremeness and basic weather—together their R^2 correlation is only 0.40.

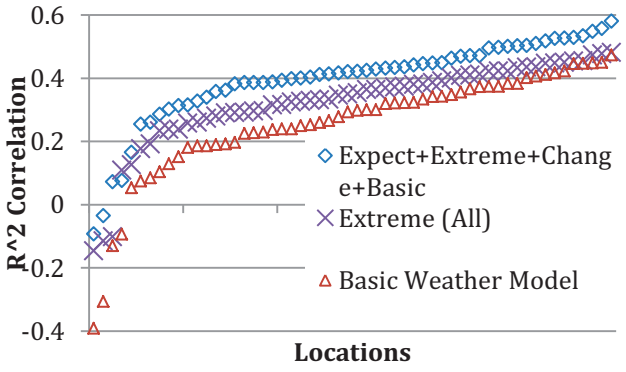

Figure 2: The distribution of R^2 correlations across locations for 3 global models

Correlating Delta Change and Tweet Rates

Finally, we analyze the relationship between delta-change in weather and tweet rates. As in our study of extremeness, we measure delta-change over multiple time-scales, including change over 1 day, 3 days and 7 days. Table 3 shows the R^2 correlation resulting from our experiments. Overall, we see that there is little difference in the amount of information gained from building these delta-change models over different time scales. Each model's R^2 correlation is approximately 0.22. The three models in our experiments do combine, however, to provide a higher R^2 correlation of 0.28.

Combining Extremeness, Expectation, and Delta Change Models

Combining our models, we find that our derived features of extremeness, expectation and delta change provide an R^2 correlation of 0.42, and adding basic weather information provides little additional benefit.

Per-Location Models

Throughout our analyses, we find that there is a high variance in the R^2 correlations between our global models predictions and individual location's weather tweet rates, as shown in Figure 2. Moreover, we see that it is the same locations that have consistently lower or higher correlation scores than average, regardless of the model used. A preliminary investigation of the general Twitter behaviors in, geographic location of, and weather at these locations did not find any clear explanations for the differences in model performance.

In Table 4, we show how our models behave when trained not globally but separately per location. We find that these models, trained on individual locations (*i.e.*, metropolitan areas), are significantly more accurate, with average R^2 correlations of over 0.70, indicating that there are likely to be additional location-specific characteristics that our global models could take into account to improve performance. Figure 3 shows the distribution of R^2 correlation across locations. Compared to the data in

Figure 2, we see that not only have the R^2 correlations improved, but the slope of the curve has also flattened, indicating that the best and worst modeled locations are now more similar. Inspecting the underlying data, we see that while there are still cities modeled more poorly than others, they are no longer the same locations (*e.g.*, Washington DC and Puerto Rico) that were being poorly modeled with our global models. We are currently investigating this further to see how we can learn from the behavior of these local models to improve our global modeling.

Table 4: Mean R^2 Correlation of local models

Model Features	R^2 correlation
Basic weather	0.45
Expectation+Basic	0.70
Extreme+Basic	0.70
Change+Basic	0.71

Discussion

Additional Factors Likely to Effect Tweet Rates

In addition to the factors of extremeness, expectation and delta-change, there are several other major factors that we did not account for in this study.

- **Sentiment:** The sentiment about an event is likely to have a significant impact on the likelihood of tweeting about it. Garcia and Schweitzer (2011) found in their analysis of Amazon product reviews have shown that product reviews have a skewed sentiment --- for example, positive reviews are more likely to be very positive than mildly positive. A similar relationship between expressed sentiment and the likelihood of people tweeting about a real-world event could be an important factor in interpreting social media.

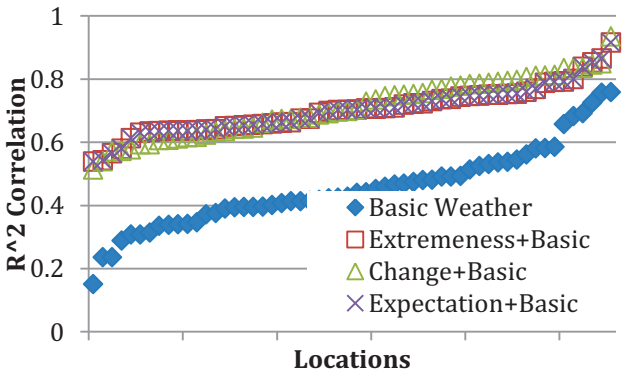

Figure 3- The Distribution of R^2 correlation across locations for local models

- **Privacy concerns, embarrassments and safety:** There are many events, such as medical issues and crime reporting, where there is likely to be a disincentive to publicly report an otherwise interesting event. We believe that the impact of such privacy-related concerns is important to understanding the "tweet-ability" of events in general. However, our specific selection of the weather domain, with little if any privacy concerns, as our first ground-truth comparison did not provide an opportunity to study such concerns.
- **Population segments**: In their study, Naaman, Boase and Lai (2010) found that some Twitter users ("me-formers") are significantly more likely to tweet about mundane activities and events, as compared to others (*i.e.*, "informers"). Modeling how many individuals are tweeting, which population segment, and their Twitter behaviors may be important.
- **Mobile devices:** The market penetration and usage of mobile devices may have an impact on the relative frequency of *in situ* tweets in a metropolitan area, and thus affect the likelihood that Twitter users report on weather events that they experience.
- **Time-of-Day, day-of-week, holiday, and other effects of time:** Our underlying weather features represented the weather to the granularity of a single day, and do not differentiate among weather events that occur during working hours, lunch-time, evening hours, commuting hours, etc. Nor do we discriminate among the different days when a weather event might occur (*e.g.*, weekend vs. weekday). It is plausible that weather events that occur when people are likely to be outdoors, or events that interact or interfere with periodic or one-time activities would be more likely to be noticed and reported in social media.

Other Sources of Bias

In this paper, our examination of reports of objective events (weather) studies self-reporting bias introduced when a user must choose whether or not to tweet. However, this study does not address other sources of bias. For example, we do not study the bias introduced by social pressure to conform or the bias introduced by privacy concerns. Nor do we address the issue of population bias introduced by focusing on a population of Twitter users instead of the public as a whole. Furthermore, we do not study the effect of bias due to individual characteristics—*e.g.*, an individual's propensity to tweet about mundane vs. extraordinary events—or specific social network

characteristics; or due to external factors (*e.g.*, priming effects). We leave these and other factors to be studied in future work.

Other Sources of Ground Truth

This paper is the beginning of a broader investigation into the properties of real-world events and trends that make them more or less likely to be discussed in social media. In addition to improving our understanding of when and why people tweet about weather events, we are also planning to gather and analyze additional ground-truth data from other domains, such as sports events and concerts.

In the process, we expect we will have an opportunity to test the influence of additional factors, such as sentiment, as well as investigate to what degree, if any at all, our findings may be applicable across domains. In this context, it is worth mentioning that studies of method variance in behavioral sciences have been found to depend heavily on constructs and domains (Cote and Buckley 1987).

Conclusions

In this paper, we compared ground-truth data about a year of weather data across the United States to weather-related tweets. We studied the correlation between daily tweet rates and the expectation, extremeness, and the change in observed weather. We find that we can build global models that take into account basic weather information, together with extremeness, expectation and change calculations to account for over 40% of the variability in tweet rates. We can build location-specific (*i.e.*, a model per each metropolitan area) models that account for an average of 70% of the variability in tweet rates.

Of the three factors we studied—expectation, extremeness, and the change in weather—we found that extremeness provided the most value in accounting for the variability of Tweet rates. The N-day change in weather was the second most important factor, and expectation based on time-of-year provided relatively little value. In the future, we plan to further investigate the relationship between these and other underlying factors of events and their associated tweet rates across domains.

This is, to our knowledge, the first large-scale quantitative analysis of the correlation between features of real-world events and the biases of their representation in social media.

References

Asur, S; and Huberman, B.A., 2010. Predicting the Future with Social Media. In *IEEE/WIC/ACM Intl. Conf. on Web Intelligence and Intelligent Agent Technology.*

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D.Z. 2011. Exploring Millions of Footprints in Location Sharing Services. In *ICWSM'11.*

Cote, J. A.; and Buckley, R. 1987. Estimating trait, method and error variance: Generalizing across 70 construct validation studies. In *Journal of Marketing Research, 24.*

Donaldson, S.I.; Grant-Vallone, E.J. 2002. Understanding Self-report Bias in Organizational Behavior Research. In *Journal of Business and Psychology, 17(2).*

Garcia, D.; Schweitzer, F. 2011. Emotions in Product Reviews – Empirics and models. In *Proc. of IEEE Intl. Conf. on Social Computing.*

Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M.; and Brilliant, L. 2008. Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014.

Goel, S.; Hofman, J.M.; Lahaie, S.; Pennock, D.M.; Watts, D.J. 2010. What Can Search Predict? In *WWW'10.*

Goel, S.; Hofman, J.M.; Lahaie, S.; Pennock, D. M.; Watts, D.J. 2010. Predicting consumer behavior with Web Search. In *PNAS.*

Hecht, B.; Hong, L.; Suh, B.; Chi., E. 2011. Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles. In *CHI'11.*

Intille, S. S.; Tapia, E. M.; Rondoni, J.; Beaudin, J.; Kukla, C.; Agarwal, S.; Bao, L.; and Larson, K. 2003. Tools for Studying Behavior and Technology in Natural Settings. In *Ubicomp'03.*

Java, A.; Song, X.; Finin, T.; Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proc. of the Joint 9th WEBKDD and 1st SNA KDD Workshop.*

Naaman, M.; Boase, J.; and Lai, C-H. 2010. Is it Really About Me? Message Content in Social Awareness Streams. In *CSCW'10.*

Paul, M.; and Dredze, M. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *ICWSM'11.*

Podsakoff, P. M.; MacKenzie, S. B.; Lee, J-Y; and Podsakoff, N. P. 2003. Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. In *Journal of Applied Psychology 88(5).*

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW'10.*

Sheth, A. 2009. Citizen Sensing, Social Signals, and Enriching Human Experience. In *IEEE Computer 13(4):87 92.*