ABSTRACT
The Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) released in 2014 states that warming of the climate system is “unequivocal,” and it is “extremely likely” that human influence has been the dominant cause. However, public perceptions of anthropogenic climate change have varied widely, and indeed may have been significantly influenced by a disproportional set of non-scientific opinion makers. While the statistics of extremes such as heat waves and heavy rainfall have been scientifically attributed to climate change, such attributions are not possible for single extreme events. Nevertheless, articles in social science and climate journals, including Nature Climate Change, have suggested that exposure to extreme weather events can directly influence opinions about climate change. Greenhouse-gas reduction policies, resilience to natural hazards, and adaptation decisions ultimately rely significantly on having adequate public support, but conducting real-time surveys of public perceptions can be difficult, expensive, and occasionally even impossible. The role of the micro-blogging site Twitter (http://twitter.com) has turned the Web into a major repository of topical comments, and hence a potential source of information for social science research. This paper attempts to understand whether Twitter data mining can complement and supplement insights about climate change perceptions, especially how such perceptions may change over time upon exposure to climate related hazards. A combination of techniques drawn from text mining, hierarchical sentiment analysis and time series methods is employed for this purpose. Future research is motivated in these areas, while potential pitfalls are discussed.

1. INTRODUCTION
Despite scientific consensus [14], climate change remains a politically polarizing topic. A recent Op-Ed article 1 in the New York Times claimed: “Here’s a scary fact about America: We’re much more likely to believe that there are signs that aliens have visited Earth (77 percent) than that humans are causing climate change (44 percent).”

The IPCC’s Special Report on Extremes [4] and the Fifth Assessment Report [14] attempted to relate the statistics of weather or hydrological extremes to human induced global warming. The current state of the science does not usually permit relating individual extremes to climate change. However, as indicated in Nature Climate Change [13], public opinions are known to be disproportionately influenced by exposure to extreme events. In addition, national and international policies, news events, and even email leaks appear to influence public perceptions.

Public perceptions of climate change have been tracked through carefully designed manual and localized surveys [13]. Here we ask the question whether mining the social media may provide an alternate source of tracking public opinions on climate change. While manual surveys will likely remain indispensable, automated or semi-automated surveys of the social media may offer complementary and supplementary benefits. The findings from social media surveys may help sharpen manual surveys, while the latter may lead to more focused surveys on the social media. In situations where manual surveys may be difficult, social media surveys may offer a first-order assessment of public sentiment.

The role of the micro-blogging site Twitter has turned the Web into a major repository of comments on many topics and a potential source of information for social science research. Twitter’s core function allows users to post short messages, or tweets, which are up to 140 characters long and allows several ways for users to communicate with each other or express their opinions about a specific object. There is a significant prior literature which have attempted to relate Twitter sentiment insights to particular events [15, 9, 1]. The proof-of-concept results presented here is a step toward understanding the value of social media mining for tracking climate change opinions.

Several effort have been placed on detecting public perception on climate change [13, 7, 2, 3]. However, none of the previous work has utilized the widely available comment in-
Our major sentiment analysis is based on the tweets excluding re-tweeted tweets. This is because we assume the sentiment behind re-tweeted tweets is hard to detect and analyze. We extracted a total of 285,026 tweets posted in English that are not re-tweeted. Although Twitter is a big collection of comment information and text data, a major disadvantage of Twitter data is that tweets do not have labels. Previous work on sentiment analysis with Twitter data employed emoticons such as “:-)” and “:-(” to form a training set for sentiment classification [12]. In our work, we manually labeled the Twitter data and classified them into subjective and objective groups. Within the subjective group, we further distinguish them into positive and negative classes. Subjective tweets mean that the tweets express users’ opinions or emotions regarding climate change; whereas, objective tweets are normally news regarding climate change or the title of an article. We label the tweets which believe in climate change, are concerned about it, or express the opinion that actions need to be taken to mitigate climate change as positive ones. In contrary, the negative tweets include opinions that do not believe in climate change, and think it is just a scam. We have five people who worked on labeling the Twitter data, and choose the label which have at least three people who agree (i.e., we use the majority label). Overall, we have collected 1050 objective tweets and 1500 subjective tweets, and within the 1500 subjective tweets, we have 1000 positive tweets and 500 negative tweets.

3. APPROACH

In this paper, we analyze the data hierarchically, where we first apply subjectivity detection to distinguish subjective tweets from the objective ones in the entire corpus and then perform sentiment analysis only within the subjective tweets.

We represent each tweet with a bag-of-words representation. Because each tweet is short, we use binary word indicators as our feature representation. We pre-process our data as follows: We lowercased all letters (strip casing off all words), tokenized (convert the string to a list of tokens based on whitespace and remove punctuation marks), removed rare words ([5] suggests that words occurring two or fewer times may be removed, since these words are unlikely to be present to aid in future classifications), removed stopwords and frequent words, and reduced each word to its stem (removing any prefixes and suffixes).

We explored two classification methods for sentiment text classification: Naïve Bayes [11] and Support Vector Machines (SVMs) [16]. Naïve Bayes is a generative classifier, whereas a support vector machine is an example of a discriminative classifier. We chose Naïve Bayes and SVM in this study because both have been proven to perform well on text classification tasks. In addition to these two methods, we also performed feature selection on our Twitter data. We found that feature selection is important because each tweet is typically very short, where each message is not allowed to exceed 140 characters, making a bag-of-word feature representation (with dimensionality equal to the number of words in the Twitter dictionary) for each sample tweet to be very sparse.

Feature Selection. We initially have $D = 1300$ features (words). Not all of these features will be important for the classification task. Furthermore, our problem is quite sparse; even sparser than typical document classification tasks. Thus, feature selection will be helpful. Feature
selection algorithms are defined by the criterion utilized for evaluating features and the search strategy. Searching all $2^D$ possible feature subsets is intractable. Here, we apply a simple search strategy by simply scoring each feature individually. There are numerous ways of evaluating or scoring features. [5] compares various feature selection metrics and their impact on the performance of classifiers. In our work, we use the chi-squared metric, which is a common statistical test that measures divergence from the distribution expected if one assumes the feature occurrence is actually independent of the class value. The formulation of the chi-squared measure is: $X^2(D, f, c) = \sum_{f \in \{0,1\}} \sum_{c \in \{0,1\}} \frac{(N_{f,c} - \beta_{f,c})^2}{\beta_{f,c}}$, where $e_f = 1$ means the document contains term $f$, and $e_f = 0$ means the document does not contain term $f$. $e_c = 1$ represents the document is in class $c$ and $e_c = 0$ represents the document is not in class $c$. $E$ is the expected frequency when the assumption that the presence of feature $f$ and class $c$ is independent is satisfied. Higher value of $X^2$ indicates that the hypothesis of independence is incorrect. We then rank order the features based on this score.

To determine the model order, meaning the number of features to keep, we measure the classification performance on a held-out validation set. We use both \textit{macro} F measure and accuracy as performance measures. F1 measure is defined as $F_1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$, where recall measures the ratio of the number of true positive cases to the number of all observed positive cases, and precision measures the ratio of the number of true positive cases to the number of all predicted positive cases by classifier. There are two methods for averaging the F-measure over a collection of 2-class classification problems. One is the \textit{macro} averaged F-measure, which is the traditional arithmetic mean of the F-measure computed for each problem. Another is the \textit{micro} averaged F-measure, which is an average weighted by the class distribution. Since we are interested in average performance across different classifications, so we focus on \textit{macro} averaged F-measure.

4. SENTIMENT ANALYSIS

In this section, we report the results on our sentiment analysis on mining the Twitter climate change data. We first randomly select one-fifth of entire labeled tweets as valida-

![Figure 2: 10-Fold Cross-Validation for Each Method, Varying the Number of Features](image)

Table 1: Candidate Models

<table>
<thead>
<tr>
<th>Subjectivity</th>
<th>Algorithm</th>
<th>No. Features</th>
<th>Accuracy</th>
<th>F1 measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjectivity</td>
<td>SVM</td>
<td>30</td>
<td>0.7535</td>
<td>0.8025</td>
</tr>
<tr>
<td>Subjectivity</td>
<td>Naive Bayes</td>
<td>100</td>
<td>0.7581</td>
<td>0.8215</td>
</tr>
<tr>
<td>Subjectivity</td>
<td>SVM</td>
<td>800</td>
<td>0.7548</td>
<td>0.8292</td>
</tr>
</tbody>
</table>

Model Selection.

We perform feature selection for both SVM and Naive Bayes classifiers. We rank ordered the features based on the chi-squared scoring described in the previous section and evaluated the performance of these two classifiers for varying number of features and evaluated the classifiers’ performance on both tasks, \textit{subjective vs objective} and \textit{positive vs negative}, based on accuracy and F1-measure using 10-fold cross-validation on the training set. The results are shown in Figure 2. The performances of both algorithms vary significantly with different number of features. As the feature size increases, both methods have serious over-fitting problem. The reason for this observation is that: firstly, tweets are relatively short compared to other documents leading to very sparse feature vectors in high dimension; secondly, the training data size is relatively limited for such high dimension. The result confirms our premise of the importance of feature selection. With small number of features, the two algorithms perform well in predicting sentiment. After careful comparison of the result, we selected a few set of candidate models to compare. We report the performance results of these candidate models on our held-out validation set in Table 4. For the subjectivity detection task, Naive Bayes has the highest accuracy performance using 1000 features compared to the SVM results. But the F1 measure of this Naive Bayes classifier is slightly less than that of SVM. Because, the performance of both classifiers are almost the same, we prefer to use SVM because it uses much fewer features to avoid the over-fitting problem. We chose SVM with 30 features compared to 20 features because it has better performance on both accuracy and F1 measure. Similarly, for the polarity sentiment task, we select SVM with 100 features. Because their performance are almost similar, we pick the model with fewer features.

Prediction and Event Detection.

With the selected subjectivity detection and sentiment polarity algorithm, we extract the \textit{subjective} tweets from our entire climate change related tweets which have been divided into subgroups based on day. We, then, predict the sentiment \textit{polarity} on the subjective tweets as reported daily to calculate the percentage of \textit{positive} and \textit{negative} sentiments. The daily percentage of subjective and objective tweets are shown in Figure 3 (top), and the percentage of \textit{positive} and
The subjective and objective percentages present large variability as we move along the time axis. This variability is influenced by many factors, such as the news, articles published on that specific day or the occurrence of any event. Because of these confounding factors, it is not easy to detect any major change or event using the subjective and objective percentages. It would be quite beneficial to climate sentiment studies if we can detect whether the sudden change in Twitter sentiment regarding climate change are related to major climate events or extreme weather conditions. We, thus, focus on the sentiment polarity percentages. We analyze the sentiment polarity percentage trend by tracking the mean and standard deviation calculated from a fixed-size sliding window for each time point, and plot the z-score normalization as a function of time [8]. In [8], they calculate the z-score normalization for each of the six normalized moods scores, which are in the range of $[0, 1]$, from POMS scoring (Profile of Mood States, a well-established psychometric instrument). The z-score normalization can be calculated as follows: $m_z = \frac{m - \bar{m}(\theta(i, \pm k)) \pm \sigma(\theta(i, \pm k))}{\bar{m}(\theta(i, \pm k)) \pm \sigma(\theta(i, \pm k))}$, where $\bar{m}(\theta(i, \pm k))$ and $\sigma(\theta(i, \pm k))$ represent the mean and standard deviation of the time series within the local $[i, \pm k]$ $m$-neighborhood for a specific day. $m$ is a normalized mood score.

In this work we have the negative percentage data expanded for 67 days, we consider it as the normalized negative mood score $m$ in the above approach. For example, higher percent of negative sentiment for a specific day represents the higher mood score for the negative sentiment. Using a sliding window size of 7 days (3 days before and 3 days after), we can derive the z-score normalization to detect short-term fluctuations of public negative sentiment as a result of particular short-term events. The result is shown in Figure 4. We are interested in looking at the point which has a z-score close to or near ±2. We observe that we can relate several climate change related events to the sudden fluctuation of negative sentiment. On Nov. 1st, because of President Obama’s Executive order on climate change, the significant increase in negative sentiment about climate change (day 31) can be detected from the graph. Decrease in negative sentiment about climate change can be observed on Nov. 11th (day 40). This is probably because of the occurrence of the destructive Typhoon Haiyan in the Philippines.

Figure 3: Subjectivity Detection and Sentiment Polarity Prediction

Negative tweets are shown in Figure 3 (bottom). The day presented in the graph is from October 3, 2013 to December 12, 2013 (excluding November, 21, 22, 23 and 24).

The subjective and objective percentages present large variability as we move along the time axis. This variability is influenced by many factors, such as the news, articles published on that specific day or the occurrence of any event. Because of these confounding factors, it is not easy to detect any major change or event using the subjective and objective percentages. It would be quite beneficial to climate sentiment studies if we can detect whether the sudden change in Twitter sentiment regarding climate change are related to major climate events or extreme weather conditions. We, thus, focus on the sentiment polarity percentages. We analyze the sentiment polarity percentage trend by tracking the mean and standard deviation calculated from a fixed-size sliding window for each time point, and plot the z-score normalization as a function of time [8]. In [8], they calculate the z-score normalization for each of the six normalized moods scores, which are in the range of $[0, 1]$, from POMS scoring (Profile of Mood States, a well-established psychometric instrument). The z-score normalization can be calculated as follows: $m_z = \frac{m - \bar{m}(\theta(i, \pm k)) \pm \sigma(\theta(i, \pm k))}{\bar{m}(\theta(i, \pm k)) \pm \sigma(\theta(i, \pm k))}$, where $\bar{m}(\theta(i, \pm k))$ and $\sigma(\theta(i, \pm k))$ represent the mean and standard deviation of the time series within the local $[i, \pm k]$ $m$-neighborhood for a specific day. $m$ is a normalized mood score.

In this work we have the negative percentage data expanded for 67 days, we consider it as the normalized negative mood score $m$ in the above approach. For example, higher percent of negative sentiment for a specific day represents the higher mood score for the negative sentiment. Using a sliding window size of 7 days (3 days before and 3 days after), we can derive the z-score normalization to detect short-term fluctuations of public negative sentiment as a result of particular short-term events. The result is shown in Figure 4. We are interested in looking at the point which has a z-score close to or near ±2. We observe that we can relate several climate change related events to the sudden fluctuation of negative sentiment. On Nov. 1st, because of President Obama’s Executive order on climate change, the significant increase in negative sentiment about climate change (day 31) can be detected from the graph. Decrease in negative sentiment about climate change can be observed on Nov. 11th (day 40). This is probably because of the occurrence of the destructive Typhoon Haiyan in the Philippines.

5. CONCLUSION

Traditionally, the attitudes, knowledge, and opinions of citizens and key decision-makers have been studied through relatively expensive and logistically challenging survey techniques, but more recently scientists and many other groups have begun to exploit the vast amounts of information available in social media platforms. This paper presents proof of concept results to suggest that mining social media data, exemplified here through Twitter accounts, can be a valuable way to yield insights on climate change opinions and societal response to extreme events.

Our work points to new opportunities for leveraging and developing knowledge discovery methods, such as opinion mining and time series techniques using social media platforms, for social science research and informing urgent societal priorities. Specifically, we used classical sentiment analysis algorithms in detecting and tracking opinions regarding climate change from Twitter feeds. In addition to measuring overall patterns and trends in climate-related sentiments, we detected a connection between short-term fluctuations in negative sentiments and major climate events. We found that major climate events can have a result in sudden change in sentiment polarity, but considering the variation in sentiment polarity shows that there is still significant uncertainty in overall sentiment. We used Twitter data to illustrate how the opinions of Twitter users can change over time and in the aftermath of specific events, but similar approaches may be extended to other publicly available information and social media platforms. While Twitter users may not represent all societal groups, its large and rapidly growing popularity supports information exchange among roughly 50 million U.S. citizens (250 M users globally), government agencies, political leaders, activist organizations, and other influential opinion-makers.
6. REFERENCES


Acknowledgments

Ganguly is funded by NSF Expeditions in Computing Grant # 1029711. Scyphers is an NSF SEES fellow and funded by NSF OCE-1215825. Dy is funded by NSF IIS-0915910.