Analysis and Visualization of Sentiment and Emotion on Crisis Tweets

Megan Torkildson, Kate Starbird, and Cecilia Aragon

University of Washington, Seattle, WA, USA
{mtorkild, kstarbi, aragon}@uw.edu

Abstract. Understanding how people communicate during disasters is important for creating systems to support this communication. Twitter is commonly used to broadcast information and to organize support during times of need. During the 2010 Gulf Oil Spill, Twitter was utilized for spreading information, sharing firsthand observations, and to voice concern about the situation. Through building a series of classifiers to detect emotion and sentiment, the distribution of emotion during the Gulf Oil Spill can be analyzed and its propagation compared against released information and corresponding events. We contribute a series of emotion classifiers and a prototype collaborative visualization of the results and discuss their implications.

Keywords: Sentiment Analysis, Twitter, Machine Learning

1 Introduction and Related Work

Many users turn to Twitter during times of crises to seek and relay information. A tweet transmits not only information, but often emotion. However, research to understand emotion in disaster-related tweets has been relatively unexplored. We seek to better understand how users communicate emotion through analyzing Twitter data collected during the 2010 Gulf Oil Spill.

The Gulf Oil Spill, which began on April 20, 2010, spanned 84 days, during which it evoked an emotional response on many levels, not just as a reaction to the human-induced ecological disaster, but also to negligence of BP and perceived inadequacies of the response efforts. Our dataset contains 693,409 tweets ranging from May 18 – August 22, 2010. All tweets contain "#oilspill," the prevalent hashtag for the event.

Analysis of large tweet corpora is scalable using machine learning. We created a series of classifiers to detect emotion using ALOE (Affect Labeler of Expression), created by Brooks et al. (2013). The tool trains a Support Vector Machine (SVM) classifier on labeled data. SVMs have previously been successful in detecting Twitter emotion in other projects, such as Roberts et al. (2012), who created a series of binary SVM emotion classifiers receiving F-measures, the weighted average of precision and recall, ranging from 0.642 to 0.740.

Analyzing emotion in text-based communication provides insight for understanding how people communicate during disasters. Emotion detection provides context information; for example, identifying tweets labeled as "fear" might support responders in assessing mental health effects among the affected population. Due to the large size of many disaster related datasets, machine learning can help scale analysis. Schulz et al. (2013), use 7-class and 3-class classification to achieve between 56.6% to 55.8% accuracy when trained on random data; however, these classifiers received between 24.4% and 39.5% accuracy when applied to tweets from Hurricane Sandy. This demonstrates the need for a series of emotion classifiers trained on a disaster dataset, due to its unique qualities. Binary classifiers allow for multiple labels, possibly providing better understanding.

2 Methods

We created a taxonomy of emotion based on Ekman's six basic emotions: joy, anger, fear, sadness, surprise, disgust (Ekman 1992). Through manual coding, we added "supportive" and "accusation," due to significant occurrences in this corpus. For the sentiment classifiers, we used the scheme: positive, negative, and neutral, intended to be mutually exclusive.

During coding, we disregarded all links because the majority could not be recovered automatically. Article and blog titles in tweets and retweets were used to help label tweets for emotion. Hashtags were also used to determine the emotion content—for example, the hashtag "#ihatebp" exhibits negative sentiment and anger. Emoticons were considered, however they were rarely used in the dataset.

Tweets were randomly sampled. If the author sent an additional tweet within the hour, it was used to help determine the emotional content of the labeled tweet. In total 5054 tweets were coded, 0.7% of the database.

Tweet Text	**Sentiment and Emotion Label**
#oilspill #bp Oil firms start spill response project – [URL]	Neutral No emotion
Note to BP: You do not own the Gulf of Mexico! Public has a right to know what you're doing to it. #oilspill [URL]	Negative Anger
BP should be fined for every single bird, fish, sea turtle and human hurt or killed by this disaster. #gulf #oilspill	Negative Disgust
What happens when energy resources deplete....? [URL] #oilspill #blacktide	Negative Fear

Table 1. Example tweets coded for emotion.

3 Visualization

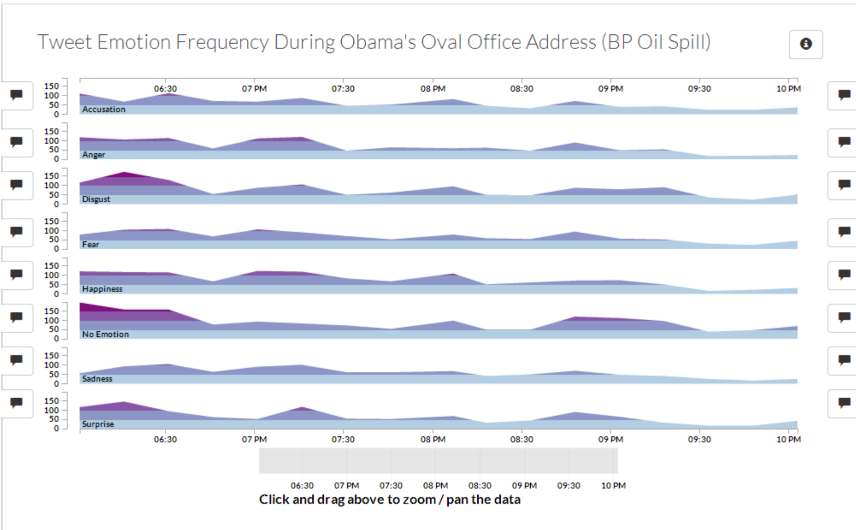

Figure 1. A screenshot of the Oil Spill Visualization. It displays the frequency of emotion detected in Twitter ranging from 2 hours before Obama's Oval Office Address (June 15, 2010 at 8:01 P.M. EDT) to 2 hours after.

To help better understand the coded data, we created a visualization that displays the frequency of emotion labels for events within the dataset. The goal was to support analysis of the emotional impact of events. The stacked area charts allow for easy comparison of values, facilitated by the colored bands. The top 25% of values, the time instances with the highest emotion frequency, have the highest color saturation. The coloring makes these peaks easily distinguishable. From this visualization, there is a large decrease in the number of disgust and accusation tweets after the time of Obama's speech.

Currently, the visualization's collaborative nature stems from its support for analysis. Users can reference the visualization and read example tweets to sensemake about the data. Further collaborative features are planned, such as shared views and annotations, to support the formation of hypotheses between researchers.

4 Results and Discussion

Table 2 shows classifier performance. The accuracy is appropriate compared to the small percentage of the coded dataset. The trend of low precision was caused by a high number of false positives. High imbalance within the dataset, such as a significantly low occurrence of positive tweets, also contributed to lower accuracy results. Additional coding of the dataset can improve these problems.

Code	Percent Occurrence (out of 5,054 tweets)	Precision	Recall	Accuracy
Positive	10%	0.56	0.46	91%
Negative	41%	0.71	0.69	76%
Neutral	49%	0.71	0.71	71%
Accusation	5%	0.03	0.53	52%
Anger	16%	0.21	0.55	64%
Disgust	13%	0.12	0.44	55%
Fear	9%	0.10	0.70	57%
Happiness	2%	0.03	0.58	51%
No emotion	38%	0.53	0.73	70%
Sadness	8%	0.15	0.78	56%
Supportive	5%	0.18	0.72	56%
Surprise	4%	0.06	0.59	53%

Table 2. Classifier performance for all 12 classifiers.

5 Conclusion and Future Work

Creating emotion classifiers trained on a disaster dataset will improve accuracy for this unique context. If applied to additional datasets, these classifiers may be more accurate than classifiers created for general tweets. The techniques used in this paper can be utilized for creating a taxonomy using ALOE to analyze emotion in additional Twitter datasets. The prototype visualization could also be used with different disaster datasets for sensemaking. In future work, we hope to further improve our accuracy by labeling additional data. We also plan to develop further collaborative features in the visualization and perform usability testing building on our previous experience creating collaborative visualization tools.

6 References

1. Brooks, M., Kuksenok, K., Torkildson, M., Perry, D., Robinson, J., Scott, T., Anicello, O., Zukowski, O., Harris, P., Aragon, C. (2013). Statistical Affect Detection in Collaborative Chat. *Proceedings of CSCW 2013*, 317-328.
2. Ekman, Paul. (1992). An argument for basic emotions. *Cognition & Emotion* 6.3-4, 169-200.
3. Roberts, K., Roach, M., Johnson, J., Guthrie, J., & Harabagiu, S. (2012). Empatweet: Annotating and detecting emotions on Twitter. *Proceedings of the LREC*, 3806-3813.
4. Schulz, A., Thanh, T., Paulheim, H., Schweizer, I. (2013). A Fine-Grained Sentiment Analysis Approach for Detecting Crisis Related Microposts. *Proceedings of the 10th International ISCRAM Conference*, 846-851.