Learning to Classify Emotional Content in Crisis-Related Tweets

Joel Brynielsson, Fredrik Johansson, Anders Westling
Swedish Defence Research Agency (FOI)
SE-164 90 Stockholm, Sweden
Email: firstname.lastname@foi.se

***Abstract*—Social media is increasingly being used during crises. This makes it possible for crisis responders to collect and process crisis-related user generated content to allow for improved situational awareness. We describe a methodology for collecting a large number of relevant tweets and annotating them with emotional labels. This methodology has been used for creating a training data set consisting of manually annotated tweets from the Sandy hurricane. Those tweets have been utilized for building machine learning classifiers able to automatically classify new tweets. Results show that a support vector machine achieves the best results (60% accuracy on the multi-classification problem).**

I. Introduction

During crises enormous amounts of user generated content, including tweets, blog posts, and forum messages is created, as documented in a number of recent publications, see, e.g., [1], [2], [3], [4], [5], [6]. Undoubtedly, large portions of this user generated content mainly consist of noise with very limited or no use to crisis responders, but some of the available information can also be used for detecting that an emergency event has taken place [1], understanding the scope of a crisis, or to find out details about a crisis [4]. That is, parts of the data can be used for increasing the tactical situational awareness. Unfortunately, the flood of information that is broadcast is infeasible for people to effectively find, organize, make sense of, and act upon without appropriate computer support [6]. For this reason, several researchers and practitioners are interested in developing systems for social media monitoring and analysis to be used in crises. One example is the American Red Cross' Digital Operations Center, opened in March 2012 [7]. Another example is the European Union FP7 security project Alert4All, having as an aim to improve the authorities' effectiveness of their alert and communication towards the population during crises [8], [9], [10]. In order to accomplish this, screening of new media is deemed important for becoming aware of how communicated alert messages are perceived by the citizens [11]. In this paper, we describe our methodology for collecting crisis-related tweets and tagging them manually with the help of a number of annotators. This has been done for tweets sent during the Sandy hurricane, where the annotators have tagged the emotional content as one of the classes $positive$ (e.g., happiness), $anger$, $fear$, or $other$ (including non-emotional content as well as emotions not belonging to any of the other classes). The tweets for which we have obtained a good inter-annotator agreement have been utilized in experiments with supervised learning algorithms for creating classifiers able to classify new tweets as belonging to any of the classes of interest. By comparing the results to those achieved when using a rule-based classifier we show that the used machine learning algorithms have been able to generalize from the training data and can be used for classification of new, previously unseen, crisis tweets.

The rest of this paper is outlined as follows. In Section II, we give an overview of related work. The used methodology is presented in Section III, where we describe how crisis-related tweets have been collected, selected using automated processing, and tagged manually by a number of annotators in order to create a training set. We also describe how a separate test set has been developed. In Section IV, we present experimental results achieved for various classifiers and parameter settings. The results and their implications are discussed in more detail in Section V. Finally, the paper is concluded in Section VI.

II. Related Work

The problem of sentiment analysis has attracted a lot of research in the last decade. One reason is probably the growing amounts of opinion-rich text resources made available due to the development of social media, giving researchers and companies access to the opinions of ordinary people [12]. Another important reason for the increased interest in sentiment analysis is the advances that have been made within the fields of natural language processing and machine learning. A survey of various techniques suggested for opinion mining and sentiment analysis is presented in [13]. A seminal work on the use of machine learning for sentiment analysis is the paper by Pang et al. [14], showing that good performance can be achieved for the problem of classifying movie reviews as either positive or negative.

Although interesting, the classification of movie reviews as positive or negative has limited impact on the security domain. However, the monitoring of social media to spot emerging trends and to assess public opinion is also of importance to intelligence and security analysts, as demonstrated in [15]. Microblogs such as Twitter pose a particular challenge for sentiment analysis techniques since messages are short (the maximum size of a tweet is 140 characters) and may contain sarcasm and slang. The utilization of machine learning techniques on Twitter data to discriminate between positive and

978-1-4673-6213-9/13/$31.00 ©2013 IEEE

negative tweets is evaluated in [16], [17]. Social media monitoring techniques for collecting large amounts of tweets during crises and classifying them with machine learning algorithms has become a popular topic within the crisis response and management domain. The use of natural language processing and machine learning techniques to extract situation awareness from Twitter messages is suggested in [4] (automatic identification of tweets containing information about infrastructure status), [5] (classification of tweets as positive or negative), and [6] (classification of tweets as contributing to situational awareness or not).

The main difference between our work and the papers mentioned above is that while most previous work focus on sentiment analysis (classifying crisis tweets as positive or negative), we focus on affect analysis or emotion recognition [18], i.e., classifying crisis tweets as belonging to an emotional state. This problem is even more challenging since it is a multinomial classification problem rather than a binary classification problem. We are not aware of any previous attempts to use machine learning for emotion recognition in crisis-related tweets. The use of affect analysis techniques for the security domain has however been proposed previously, such as the affect analysis of extremist web forums and blogs presented in [19], [20].

III. Methodology

Within the research project Alert4All we have discovered the need for automatically finding out whether a tweet (or other kinds of user generated content) is to be classified as containing emotional content [11]. Through a series of user-centered activities involving crisis management stakeholders [21], the classes of interest for command and control have been identified as $positive$, $anger$, $fear$, and $other$, where the first class contains positive emotions such as happiness, and the last class contains emotions other than the ones already mentioned, as well as neutral or non-subjective classifications. In the following, we describe the methodology used for collecting crisis-related tweets, selecting a relevant subset of those, and letting human annotators tag them in order to be used for machine learning purposes.

A. Collecting Tweets

The first step in our methodology was to collect a large set of crisis-related tweets. For this purpose we have used the Python package **tweetstream** [22] to retrieve tweets related to the Sandy hurricane, hitting large parts of the Caribbean and the Mid-Antlantic and Northeastern United States during October 2012. The **tweetstream** package fetch tweets from Twitter's streaming API in real-time. It should be noted that the streaming API only gives access to a random sample of the total volume of tweets sent at any given moment, but still this allowed us to collect approximately six million tweets related to Sandy during October 29 to November 1, using the search terms $sandy$, $hurricane$, and $\#sandy$. After automatic removal of non-English tweets, retweets and duplicated tweets, approximately 2.3 million tweets remained.

B. Annotation Process

After an initial manual review of the remaining collected posts, we quickly discovered that a large proportion of the tweets not unexpectedly belong to the category $other$. Since the objective was to create a classifier being able to discriminate between the different classes, we needed a balanced training data set, or at least a large number of samples for each class. This caused a problem since random sampling of the collected tweets most likely would result in almost only those belonging to the class $other$. Although this in theory could be solved by sampling a large enough set of tweets to annotate, there is a limit to how many tweets that can be tagged manually in a reasonable time (after all, this is the main motivation for learning such classifiers in the first place). To overcome this problem, we decided to based on manual inspection identify a small set of keywords which were likely to indicate emotional content belonging to any of the emotional classes $positive$, $fear$, or $anger$[1]. The list of identified keywords looks as follows:

- *anger*: anger, angry, bitch, fuck, furious, hate, mad,
- *fear*: afraid, fear, scared,
- *positive*: :), :-), =), :D, :-D, =D, glad, happy, positive, relieved.

Those lists were automatically extended by finding synonyms to the words using WordNet [23]. Some of the resulting words were then removed from the lists as they were considered poor indicators of emotions during a hurricane. An example of a word that was removed is "stormy," which was more likely to describe hurricane Sandy than expressing anger. By using the words in the created lists as search terms, we sampled 1000 tweets which according to our simple rules were likely to correspond to "positive" emotions. The same was done for "anger" and "fear," while a random sampling strategy was used to select the 1000 tweets for "other." In this way we constructed four files with 1000 tweets in each file. The way we selected the tweets may have an impact on the end results since there is a risk that such a biased selection process will lead to classifiers that are only able to learn the rules used to select the tweets in the first place. We were aware of such a potential risk, but could not identify any other way to come up with enough tweets corresponding to the "positive," "anger," and "fear" tags. In order to check the generalizability of the resulting classifiers, we have in the experiments compared the results to a baseline, implemented as a rule-based algorithm based on the keywords used to select the appropriate tweets (this will be further described in Section IV).

Once the files containing tweets were constructed, each file was sent by e-mail to three independent annotators, i.e., all annotators were given one file (containing 1000 tweets) each. All annotators were previously familiar with the Alert4All project (either through active work within the project or through acting as advisory board members) and received the instructions which can be found in the Appendix. It should be

[1] We use *class* to refer to the class a tweet actually belong to (given the annotation), and "class" to refer to the class suggested by the used keywords.

Category	Majority agreement	Full agreement
“positive”	92.7%	47.8%
“anger”	92.6%	39.2%
“fear”	95.2%	44.4%
“other”	99.7%	82.3%

TABLE I
INTER-ANNOTATOR AGREEMENT FOR THE VARIOUS CATEGORIES

Emotion class	Number of tweets
$Positive$	622
$Anger$	461
$Fear$	470
$Other$	2249

TABLE II
NUMBER OF TWEETS PER CLASS (BASED ON MAJORITY AGREEMENT)

noted that far from all the tweets in a category were tagged as belonging to that emotion by the annotators. In fact, a majority of the tweets were tagged as $other$ also in the “anger,” “fear,” and “positive” files. In order to get a feeling for the inter-annotator agreement, we have calculated the percentages of tweets for which a majority of the annotators have classified a tweet in the same way (majority agreement) and where all agree (full agreement) as shown in Table I. As can be seen, the majority agreement is consistently reasonably high. On the other hand, it is seldom that all three annotators agree on the same classification. For a tweet to become part of the resulting training set, we require that there has been a majority agreement regarding how it should be tagged. Now, ignoring which class a tweet was “supposed” to end up in given the used keywords (i.e., the used categories) and instead looking at the emotion classes tweets actually ended up in after the annotation, we received the distribution shown in Table II. Since we wanted to have a training data set with equally many samples for each class, we decided to balance the classes, resulting in 461 training samples for each class.

C. Creating a Separate Test Dataset

While it is popular in the machine learning community to make use of n-fold cross validation to allow training as well as testing on all available data, we have decided to create a separate test set in this case. The reason for this is the way training data has been generated. If the used strategy to select tweets based on keywords would impact the annotated data and thereby also the learned classifiers too much, this could result in classifiers that perform very well when using the annotated data, but that generalizes badly to “real” data without the bias. Hence, our test data has been generated by letting a human annotator (not part of the first annotation phase) tag tweets from the original collected Twitter data set until sufficiently many tweets have been discovered for each emotion. Since it, as a rule of thumb, is common to use 90% of the available data for training and 10% for testing, we continued the tagging until we got 54 tweets in each class (after balancing the set), corresponding to approximately 10% of the total amount of data used for training and testing.

IV. EXPERIMENTS

There exist many parameters related to affect analysis that influence the feature set. This section describes the parameters that have been varied during the experiments, and discusses how the parameters affected the achieved experimental results.

A. Classifiers

We have experimented with two standard machine learning algorithms for classification: Naïve Bayes (NB) and a Support Vector Machine (SVM). In the experiments we have used the NB and SVM implementations available in Weka [24]. Although many additional features such as parts-of-speech could have been used, we have limited the experiments to a simple bag-of-words representation. Initial experimentation showed that feature presence gave better results than feature frequency, wherefore only feature presence has been utilized. Before the training data was used, the tweets were transformed into lower case. Many different parameters have been varied throughout the experiments:

- n-gram size: 1 (unigram) / 2 (unigram + bigram),
- stemming: yes / no,
- stop words: yes / no,
- minimum number of occurrences: 2 / 3 / 4,
- information gain (in %): 25 / 50 / 75 / 100,
- negation impact (number of words): 0 / 1 / 2,
- threshold τ: 0.5 / 0.6 / 0.7.

If a unigram representation is used, individual words are utilized as features, whereas if bigrams are used, pairs of words are utilized as features. Stemming refers to the process in which inflected or derived words are reduced to their base form (e.g., fishing → fish). As stop words we have used a list of commonly occurring function words, so if a word in the tweet matches such a stop word it is removed (and is hence not used as a feature). The minimum number of occurrences refers to how many times a term has to occur in the training data in order to be used as a feature. Information gain refers to a method used for feature selection, where the basic idea is to select features that reveal the most information about the classes. When, e.g., setting the information gain parameter to 50, the fifty percent “most informative features” are kept, reducing the size of the resulting model. Finally, if a negation (such as “not”) is detected, the used algorithm replaces the words following the negation by adding the prefix “NOT_” to them. The specified negation impact determines how many words after a negation to be affected by the negation (where 0 means that no negation is used). Lastly, the threshold τ has been used for discriminating between emotional content versus other content, as described below.

In the learning phase we used the tweets tagged as $positive$, $anger$, and $fear$ as training data, which resulted in classifiers that learned to discriminate between these three classes. For the actual classification of new tweets we then let the machine learning classifiers estimate the probabilities $P(anger|f_1,\ldots,f_n)$, $P(fear|f_1,\ldots,f_n)$, and $P(positive|f_1,\ldots,f_n)$, where $f_1,\ldots,f_n$ refers to the used feature vector extracted from the tweet we want to classify. If the estimated probability for the most probable class is greater than a pre-specified threshold τ, we return the label of the most probable class as the output from the classifier. Otherwise $other$ is returned as the output from the classifier. The rationale behind this is that the content of tweets to be classified as $other$ cannot be learned in advance (due to the spread of what this class should contain). Instead, we learn what is considered to be representative for the other classes and interpret low posterior probabilities for $anger$, $fear$, and $positive$ as $other$ being the most likely class.

B. Experimental Results

The best results achieved when evaluating the learned classifiers on the used test set are shown in Figure 1, with the used parameter settings shown in Table III. The results are also compared to two baseline algorithms: 1) a naïve algorithm that picks a class at random, and 2) a somewhat more complex rule-based classifier constructed from the heuristics (keywords) used when selecting the tweets to be annotated manually in the training data generation phase. The results suggest that both the NB and SVM classifiers outperform the baseline algorithms, and that SVM (59.7%) performs somewhat better than the NB classifier (56.5%). The use of stemming, stop words, and information gain have consistently been providing better results, while the best choices of n-gram size, negation impact, and the used threshold have varied more in the experiments.

In addition to evaluating the classifiers' accuracy on the original test set, we have also tested what happens if the task is simplified so that the classifiers only have to distinguish between the emotional classes $positive$, $fear$, and $anger$ (i.e., it is assumed that the $other$ class is not relevant). The latter task can be of interest in a system where a classifier distinguishing between emotional and non-emotional or subjective and non-subjective content has already been applied. As can be seen, the SVM gets it right in three out of four classifications (75.3%) on this task, while the accuracy of the NB classifier reaches 69.1%.

V. Discussion

The obtained results show that the machine learning classifiers perform significantly better than chance and the rule-based algorithm that has been used as a baseline. Especially, the comparison to the rule-based algorithm is of interest, since the difference in accuracy indicates that the NB and SVM algorithms have been able to learn something more than just the keywords used to select the tweets to include in the annotation phase. In other words, even though the use of keywords may bias what tweets to include in the training data, this bias is not large enough to stop the machine learning classifiers from learning useful patterns in the data. In this sense the obtained results are successful. Although the results are promising it can be questioned whether the obtained classification accuracy is good enough to be used in real-world social media analysis systems for crisis management. We believe that the results are good enough to be used on an aggregate level ("the citizens' fear levels are increasing after the last alert message"), but are not necessarily precise enough to be used to correctly assess the emotions in a specific tweet. Nevertheless, this is a first attempt to classify emotions in crisis-related tweets, and by improving the used feature set and combining the machine learning paradigm with more non-domain specific solutions such as the affective lexicon WordNet-Affect [25], better accuracy can most likely be achieved. Increased training data sets would probably also improve the accuracy, but a problem related to this is the relatively costly effort in terms of manpower that is needed for the creation of even larger training data sets. Additionally, the learned classifiers should also be evaluated on other datasets in order to test the generalizability of the obtained results.

Some of the classification errors were a result of that the annotators received instructions to classify tweets containing any of the emotions $fear$, $anger$, or $positive$ as $other$ if the tweets relate to a "historical" state or if the expressed emotion related to someone else than the author of the tweet. Such a distinction can be important if the used classifications should be part of a social media analysis system (since we do not want to take action on emotions that are not present anymore), but no features have been used to explicitly take care of spatio-temporal constraints in the current experiments. If such features were added (e.g., using part-of-speech tags and extraction of terms that contain temporal information), some of the classification errors could probably have been avoided.

Although we in this article have focused on crisis management, there are obviously other potential areas within the intelligence and security domains to which the suggested methodology and algorithms can be applied. As an example, it can be of interest to determine what kind of emotions that are expressed toward particular topics or groups in extremist discussion forums (cf. [19], [20]). In the same manner, it can be used to assess the emotions expressed by, e.g., bloggers, in order to try to identify signs of emergent conflicts before they actually take place (cf. [15], [26]).

VI. Conclusions and Future Work

We have described a methodology for collecting large amounts of crisis-related tweets and tagging relevant tweets using human annotators. The methodology has been used for annotating large quantities of tweets sent during the Sandy hurricane. The resulting data set has been utilized when constructing classifiers able to automatically distinguish between the emotional classes $positive$, $fear$, $anger$, and $other$. Evaluation results suggest that a SVM classifier perform better than a NB classifier and a simple rule-based system. The classification task is difficult as suggested by the quite low reported inter-annotator agreement results. Seen in this light and considering that it is a multi-classification problem, the obtained accuracy for the SVM classifier (59.7%) seems promising. The classifications are not good enough to be trusted on the level of individual postings, but on a more aggregated level the citizens' emotions and attitudes toward

Parameter settings	SVM	NB
n-gram size	2 (unigram + bigram)	1 (unigram)
Stemming	yes	yes
Stop words	yes	yes
Min. nr. of occurrences	4	4
Information gain	75%	75%
Negation impact	2	2
Threshold τ	0.7	0.6

TABLE III
USED PARAMETER SETTINGS FOR THE BEST PERFORMING CLASSIFIERS

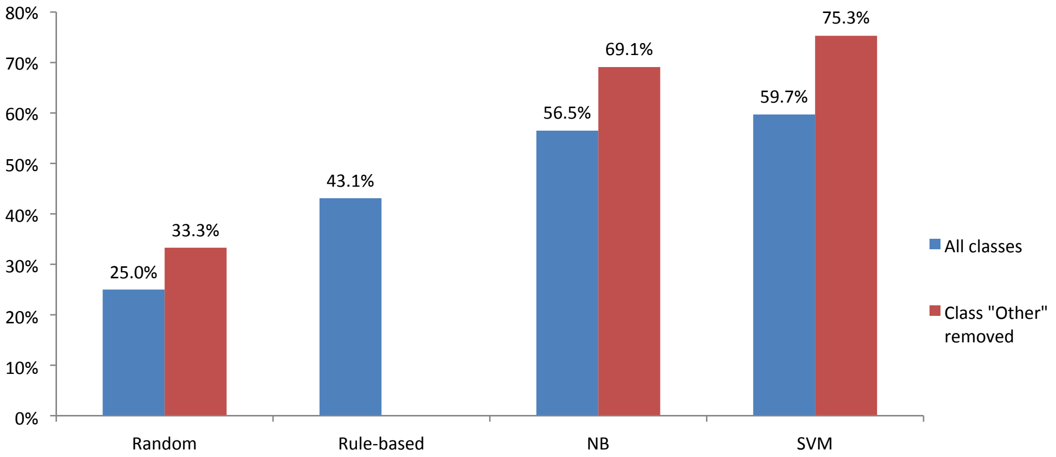

Fig. 1. Achieved accuracy for the various classifiers. Blue color shows the results on the full dataset, red color shows the results when the *other* category is removed. The rules used within the rule-based classifier assume that all classes are present, wherefore no results have been obtained on the simplified problem for this classifier.

the crisis can be estimated using the suggested algorithms. Results obtained when ignoring the non-specific category *other* (reaching accuracies over 75% for the SVM) also suggest that combining the learned classifiers with algorithms for subjectivity recognition can be a fruitful way forward.

As future work we see a need for combining machine learning classifiers learned from crisis domain data with more general affective lexicons. In this way we think that better classification performance can be achieved than if using the methods individually. Moreover, we suggest extending the used feature set with extracted part-of-speech tags since such information most likely will help determine if it is the author of a tweet who is having a certain emotion, or if it is someone else. Other areas to look into is how to deal with the use of sarcasm and slang in the user generated content.

From a crisis management perspective, it will also be necessary to investigate to what extent the used methodology and the developed classifiers are capable of coping with more generic situations. That is, we hope to have developed classifiers that to at least some significant extent classify based on hurricane and crises behavior in general, rather than solely being able to classifying Sandy-specific data. To investigate this, we will gather and manually tag new datasets to test our classifiers on. We will do this for several different crisis types, and then apply the same classifiers to be able to quantify how capable the developed classifiers are when it comes to classifying tweets from 1) other hurricanes, 2) other types of natural disasters, and 3) crises in general.

ACKNOWLEDGMENTS

We would like to thank the following annotators for their effort with the tagging process: Roberta Campo, Luisa Coelho, Patrick Drews, Montse Ferrer, Sebastien Grazzini, Paul Hirst, Thomas Ladoire, Håkan Marcusson, Miguel Mendes, Maria Luisa Moreo, Javier Mulero Chaves, Cristina Parraga Niebla, Joaquín Ramírez, and Leopoldo Santos.

This work has been supported by the Swedish Armed Forces Research and Development Programme and the Alert4All research project which is funded by the European Union Seventh Framework Programme under contract no 261732.

APPENDIX

Instructions to annotators:

You have been given 1000 tweets and a category. The tweets were written when hurricane Sandy hit the US in 2012. Hopefully most of the tweets you've been given are associated with your emotion. Your task is to go through these tweets, and for each tweet confirm whether this tweet is associated with the emotion you have been given, and if not, associate it with the correct emotion. To help make sure that the tagging is as consistent as possible between all annotators, you will be given some guidelines to make sure that everyone tags the tweets in a similar way:

- *"Fear" is the category containing tweets from people who are scared, afraid or worried.*
- *"Anger" contains tweets from people that are upset or angry. It's not always obvious whether someone is angry or sad, but if you think they are angry, tag it as "anger." It is acceptable if the person feels sadness as well.*
- *"Positive" contains tweets from people that are happy or at least feel positive.*
- *"Other" represents the tweets that don't belong to any of the other three categories. Tweets with none of the three emotions or mixed emotions where one of them isn't dominating belong to this category.*
- *The emotion should relate to the author of the tweet, not other people mentioned by the author. For example, the tweet "Maggie seems real concerned about Hurricane Sandy..." should not be tagged as "fear," since it's not the author of the tweet that is being concerned. Instead it should be tagged with "other."*
- *The tag should be based on the author's mood when the tweet was written. For example, the tweet "I was really scared yesterday!" should not be tagged as "fear," since it relates to past events, while we want to know how people were feeling when the tweets were posted. Exceptions can be made to events that happened very recently, for example: "I just fell because sandy scared me," which can be tagged as "fear."*
- *Obvious sarcasm and irony should be tagged as "Other." If you can't decide whether the author is being sarcastic or not, assume that he is not being sarcastic or ironic.*
- *A couple of the tweets might not be in English. Non-English tweets belong to "Other" regardless of content.*
- *A few of the tweets are not related to the hurricane. Treat them in the same way as the rest of the tweets.*
- *If a tweet contains conflicting emotions, and one of them doesn't clearly dominate the other, it belongs to "Other."*
- *Some of the tweets will be difficult to tag. Even so, don't leave a text untagged, please choose the alternative you believe is the most correct.*

References

[1] A. Zielinski and U. Bugel, "Multilingual analysis of twitter news in support of mass emergency events," in *Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)*, 2012.

[2] S.-Y. Perng, M. Buscher, R. Halvorsrud, L. Wood, M. Stiso, L. Ramirez, and A. Al-Akkad, "Peripheral response: Microblogging during the 22/7/2011 Norway attacks," in *Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)*, 2012.

[3] R. Thomson, N. Ito, H. Suda, F. Lin, Y. Liu, R. Hayasaka, R. Isochi, and Z. Wang, "Trusting tweets: The Fukushima disaster and information source credibility on Twitter," in *Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)*, 2012.

[4] J. Yin, A. Lampert, M. A. Cameron, B. Robinson, and R. Power, "Using social media to enhance emergency situation awareness," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 52–59, 2012.

[5] A. Nagy and J. Stamberger, "Crowd sentiment detection during disasters and crises," in *Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)*, 2012.

[6] S. Verma, S. Vieweg, W. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson, "Natural language processing to the rescue? extracting 'situational awareness' tweets during mass emergency," in *Proceedings of the 2011 International AAAI Conference on Weblogs and Social Media*, 2011.

[7] [Online]. Available: http://www.enterpriseefficiency.com/author.asp?section_id=1523&doc_id=240584

[8] C. Párraga Niebla, T. Weber, P. Skoutaridis, P. Hirst, J. Ramírez, D. Rego, G. Gil, W. Engelbach, J. Brynielsson, H. Wigro, S. Grazzini, and C. Dosch, "Alert4All: An integrated concept for effective population alerting in crisis situations," in *Proceedings of the Eighth International Conference on Information Systems for Crisis Response and Management (ISCRAM 2011)*, Lisbon, Portugal, May 2011.

[9] H. Artman, J. Brynielsson, B. J. E. Johansson, and J. Trnka, "Dialogical emergency management and strategic awareness in emergency communication," in *Proceedings of the Eighth International Conference on Information Systems for Crisis Response and Management (ISCRAM 2011)*, Lisbon, Portugal, May 2011.

[10] S. Nilsson, J. Brynielsson, M. Granåsen, C. Hellgren, S. Lindquist, M. Lundin, M. Narganes Quijano, and J. Trnka, "Making use of new media for pan-european crisis communication," in *Proceedings of the Ninth International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)*, Vancouver, Canada, Apr. 2012.

[11] F. Johansson, J. Brynielsson, and M. N. Quijano, "Estimating citizen alertness in crises using social media monitoring and analysis," in *Proceedings of the 2012 European Intelligence and Security Informatics Conference*, 2012, pp. 189–196.

[12] B. Liu, *Handbook of Natural Language Processing*, 2010, ch. Sentiment Analysis and Subjectivity.

[13] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1–135, 2008.

[14] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002.

[15] K. Glass and R. Colbaugh, "Estimating the sentiment of social media content for security informatics applications," *Security Informatics*, vol. 1, no. 3, 2012.

[16] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

[17] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 36–44.

[18] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM Symposium on Applied Computing*, 2008.

[19] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect analysis of web forums and blogs using correlation ensembles." *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1168–1180, 2008.

[20] A. Abbasi and H. Chen, "Affect intensity analysis of dark web forums," in *Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics*, 2007.

[21] J. Brynielsson, F. Johansson, and S. Lindquist, "Using video prototyping as a means to involving crisis communication personnel in the design process: Innovating crisis management by creating a social media awareness tool," in *Proceedings of the 15th International Conference on Human-Computer Interaction*, Las Vegas, Nevada, Jul. 2013.

[22] [Online]. Available: https://github.com/intridea/tweetstream

[23] G. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[25] C. Strapparava and A. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004.

[26] F. Johansson, J. Brynielsson, P. Hörling, M. Malm, C. Mårtenson, S. Truvé, and M. Rosell, "Detecting emergent conflicts through web mining and visualization," in *Proceedings of the 2011 European Intelligence and Security Informatics Conference*, 2011, pp. 346–353.