Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises

Joel Brynielsson[1,2*], Fredrik Johansson[1], Carl Jonsson[3] and Anders Westling[4]

Abstract

One of the key factors influencing how people react to and behave during a crisis is their digital or non-digital social network, and the information they receive through this network. Publicly available online social media sites make it possible for crisis management organizations to use some of these experiences as input for their decision-making. 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 dataset 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 with about 60% accuracy on the multi-classification problem. This classifier has been used as a basis for constructing a decision support tool where emotional trends are visualized. To evaluate the tool, it has been successfully integrated with a pan-European alerting system, and demonstrated as part of a crisis management concept during a public event involving relevant stakeholders.

Keywords: Alert and communication; Social media; Affect analysis; Machine learning; Trend analysis; Information visualization

Introduction

During crises, enormous amounts of user generated content, including tweets, blog posts, and forum messages, are created, as documented in a number of recent publications [1-6]. Undoubtedly, large portions of this user generated content mainly consist of noise with 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 [7]. Unfortunately, the flood of information that is broadcast is infeasible for people to effectively extract information from, 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 [8]. Another example is the European Union security research project Alert4All, having as its aim to improve the authorities' effectiveness of alert and communication towards the population during crises [9-11]. In order to accomplish this, screening of social media is deemed important for becoming aware of how communicated alert messages are perceived by the citizens [12]. 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 being able to classify new tweets as belonging to any of the classes

*Correspondence: joel.brynielsson@foi.se
[1]FOI Swedish Defence Research Agency, Stockholm, Sweden
[2]KTH Royal Institute of Technology, Stockholm, Sweden
Full list of author information is available at the end of the article

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. Further, the optimum classifier has been integrated with, and constitutes an important part of, the Alert4All proof-of-concept alerting system. In the presence of relevant stakeholders representing politics, industry, end users, and research communities, this system was successfully demonstrated as a cohesive system during a public event. As part of this event, the classification of social media posts was used to visualize emotional trend statistics for the purpose of demonstrating the idea of using social media input for informing crisis management decisions. Overall, the concept was well received, considered novel, and makes it possible for crisis management organizations to use a new type of input for their decision-making.

The rest of this paper is outlined as follows. In the next section we give an overview of related work. A methodology section then follows, 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 constructed. After that, we present experimental results achieved for various classifiers and parameter settings. Details regarding the design and implementation of a decision support tool making use of the developed support vector machine classifier is then elaborated on in a separate section. The results and their implications are then discussed in more detail in a separate section before the paper is concluded in the last section.

Related work

The problem of sentiment analysis has attracted much research during 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 [13]. 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 [14]. A seminal work on the use of machine learning for sentiment analysis is the paper by Pang et al. [15], showing that good performance (approximately 80% accuracy for a well-balanced dataset) 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 [16]. 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 negative tweets is evaluated in [17,18], suggesting that classification accuracies of 60–80% can be obtained. 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 most of the previous work focus on sentiment analysis (classifying crisis tweets as positive or negative), whilst we focus on affect analysis or emotion recognition [19], 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 of 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 [20,21].

The work presented in this article is the result of a long-term research effort where related studies have been presented along the way. A first visionary paper [10] discusses and presents the concept of using social media monitoring for coming into dialogue with the population. The overall idea is for emergency management organizations to follow what people publish and adjust their information strategies in a way that matches the expectations and needs of the public. A systematic literature review and a parallel interview study were then undertaken [11], where the possibility to use social media analysis for informing crisis communication was deemed promising and important design issues to take into account were highlighted. Based on this insight, we outlined a more detailed design concept for how a screening tool could potentially be used for the purpose of increasing situational awareness during crises [12]. This paper identifies data acquisition and data analysis to be two important parts of such a tool. Then, in parallel to presenting the initial results with regard to tweet classification [22], crisis management stakeholders

were involved in a series of user-centered activities in order to understand the user needs and further inform the design of a social media screening tool to be used for crisis management [23]. It became clear that within crisis management it is more important to be able to distinguish between negative emotions such as fear and anger than to be able to differentiate between different positive emotions. Also, a further understanding of crisis management working procedures was obtained, which made it clear that a social media screening tool needs to be focused on trend analysis since, in crisis management, relevant actions are to be undertaken for the purpose of improving some kind of crisis state in order to bring the situation into a better state.

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 [12]. Through a series of user-centered activities involving crisis management stakeholders [23], 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.

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** to retrieve tweets related to the Sandy hurricane, hitting large parts of the Caribbean and the Mid-Atlantic and Northeastern United States during October 2012. The **tweetstream** package fetches 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, as exemplified in Table 1. An average tweet in the dataset contained 14.7 words in total and 0.0786 "emotional words" according to the lists of identified keywords as will be described in the next subsection.

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 dataset, 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*.

Table 1 Sample tweets obtained in late 2012 during the Sandy hurricane along with the resulting emotion class output from the developed emotion classifier

Tweet	Predicted class
the anticipation of when the power is going to go out! I NEED TO STUDY WHAT IS HAPPENING STOP SANDY	Anger
God damn it #sandy! There goes my cable...	Anger
Sandy just made landfall on the great State of New Jersey & NYC. Hang tight, you guys.	Anger
Sandy has denied me my jog. I'm crying as much as it's raining right now...	Anger
Shed in backyard was knocked over #sandy	Other
Lovely, there are fallen tree branches in my swimming pool. Eh, It could be worse... #413Sandy #MASandy #Sandy	Positive
So my childhood beach town is basically being destroyed. That's cool.. Stupid Sandy. :/	Anger
So much food in my house because my moms stocking up for Sandy. I'm cool with it.	Anger
Hurricane Sandy might not kill me but this boredom sure will. -__-	Anger
This storm sandy is so scary :0 #scarystuff #mothernature	Fear
Hurricane Sandy is powerful af!!! This wind is NO joke!!!	Other
Power back on. Not sure how much longer that will last. Damn you #sandy - get up off my #raw!	Anger
im like really scared.... stuff like this doesn't happen in Ohio ! #Sandy #Manhattan	Fear
NZ's embassy in Washington is closed as the city hunkers down ahead of #Sandy	Other
11 killed in #Cuba, #Sandy toll reaches 51 in #Haiti	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 use manual inspection to identify a small set of keywords which were likely to indicate emotional content belonging to any of the emotional classes *positive*, *fear*, or *anger*[a]. 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 [24]. 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 data files containing 1000 tweets each. 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. The experiments are further described in the next section.

Once the files containing tweets had been 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 noted that far from all the tweets in a file were tagged as belonging to the corresponding 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 2. 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 3. Since we wanted to have a training dataset with equally many samples for each class, we decided to balance the classes, resulting in 461 training samples for each class.

Creating a separate test dataset

While it is popular in the machine learning community to make use of *n*-fold cross validation to allow for training as well as testing on all the available data, we have decided to create a separate test set in this study. 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 well using the annotated data, but generalizes poorly 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 originally collected Twitter dataset until sufficiently many tweets had 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.

Experiments

There exists 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.

Table 2 Inter-annotator agreement for the various categories

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 3 Number of annotated tweets per class based on majority agreement

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

Classifiers

We have experimented with two standard machine learning algorithms for classification: Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers. Available in Weka [25], the multinomial NB classifier [26] was used for the NB experiments, and the sequential minimal optimization algorithm [27] was used for training a linear kernel SVM. Although many additional features such as part-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 $\rightarrow$ 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. Negation impact refers to the situation when a negation (such as "not") is detected, and 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 that should be affected by the negation (where 0 means that no negation is used). Finally, 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.

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 4. The results are also compared to two baseline algorithms: 1) a naïve algorithm that picks a class at random (since all the classes are equally likely in a balanced dataset, this corresponds to a simple majority classifier), 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 the

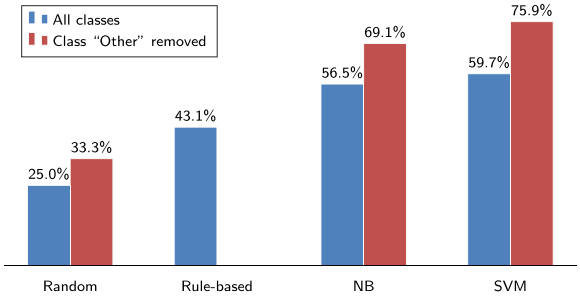

Figure 1 Achieved accuracy for the various classifiers. Blue color shows the results on the full dataset, red color shows the results when the *other* category is removed. The rules used within the rule-based classifier assume that all classes are present, wherefore no results have been obtained on the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

Table 4 Used parameter settings for the best performing classifiers

Parameter settings	NB	SVM
n-gram size	1 (unigram)	2 (unigram + bigram)
Stemming	Yes	Yes
Stop words	Yes	Yes
Min. no. of occurrences	4	4
Information gain	75%	75%
Negation impact	2	2
Threshold τ	0.6	0.7

SVM (59.7%) performs somewhat better than the NB classifier (56.5%). For a more detailed accuracy assessment, see Tables 5 and 6 where the confusion matrices show how the respective classifiers perform. The use of stemming, stop words, minimum number of occurrences, and information gain according to Table 4 have consistently been providing better results, while the best choices of n-gram size, negation impact, and threshold τ have varied more in the experiments.

For comparison, Table 7 contains the confusion matrix for the baseline classifier, i.e., the rule-based classifier which chooses its class based on possible emotion words found within a tweet. As can be seen in Table 7, the classifications of emotions (i.e., "anger", "fear", or "positive") are often correct, but a large amount of the tweets tend to erroneously fall into the *other* category. Now looking back at the machine learning confusion matrices according to Tables 5 and 6, we see that these classifiers do not exhibit the same behavior as the rule-based classifier with regard to the *other* category, but instead shows more evenly distributed errors. Hence, we can see that the machine learning classifiers have indeed learnt about emotional patterns that cannot be distinguished by simply applying rules based on a predefined list of emotion words.

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). This 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 in Figure 1, the SVM gets it right in three out of four classifications (75.9%) on this task, while the accuracy of the NB classifier reaches 69.1%. See Tables 8 and 9 for the corresponding confusion matrices.

Table 5 Confusion matrix for the optimized SVM classifier

	Predicted class			
Actual class	**Anger**	**Fear**	**Positive**	**Other**
Anger	38	5	6	5
Fear	4	37	3	10
Positive	8	4	29	13
Other	14	12	3	25

The matrix shows how the classifier predictions are distributed, and thereby how well the classifier has learnt to distinguish between the classes.

Table 6 Confusion matrix for the top-performing NB classifier

	Predicted class			
Actual class	**Anger**	**Fear**	**Positive**	**Other**
Anger	35	4	5	10
Fear	3	35	9	7
Positive	13	2	19	20
Other	12	6	3	33

Design and implementation of a tool for visualizing emotional trends

Based on a series of stakeholder workshops [23], the developed emotion classifier has been used as a basis for the design and implementation of a decision support system entitled the "screening of new media" tool where emotional trends are visualized. To evaluate the tool, it has been integrated with the Alert4All system, which is an implemented prototype of a future pan-European public alert concept. As shown during the final demonstration of the Alert4All system and through the collocated user-centered activities, the social media analysis component of Alert4All provides additional benefit for command and control personnel in terms of providing immediate feedback regarding the development of a crisis in general and regarding the reception of crisis alerts in particular.

Figure 2 shows the developed tool, which has been implemented using HTML5 and JavaScript components. The core component of the tool is the graph which is shown to the upper right in Figure 2 and on its own

Table 7 Confusion matrix for the rule-based baseline classifier which chooses class based on the occurrence of certain words

	Predicted class			
Actual class	**Anger**	**Fear**	**Positive**	**Other**
Anger	21	0	0	33
Fear	0	11	0	43
Positive	1	0	10	43
Other	3	0	0	51

As can be seen, many tweets end up in the *other* category.

Table 8 Confusion matrix for the SVM classifier when the task has been simplified so that the *other* class is not relevant

	Predicted class		
Actual class	**Anger**	**Fear**	**Positive**
Anger	41	5	8
Fear	5	45	4
Positive	13	4	37

in Figure 3. Here, a number of interactive chart components are used in order to visualize how the emotional content in the acquired dataset changes as a function of time. Through interacting with this graph, the user has the possibility to interact with the underlying dataset, and thereby obtain a further understanding of how the feelings expressed on social media vary as time passes.

At the bottom of the tool, the user has the possibility to drill down into the underlying dataset and see the actual posts in the database. From a command and control perspective, it is important to remember that these individual messages cannot and should not be used for inference regarding the whole dataset, but should be used solely for generating new hypotheses that need to be tested further by, e.g., experimenting with the filters in order to obtain sound statistical measures. Also to be noted, the posts are color coded so that it is easy to see which emotion a certain post has been classified as. However, the classification is not always correct, and therefore the user has the possibility to manually reclassify a post and, at a later stage, use the manually classified post as a basis for improving the classifier.

The GUI provides a number of ways to apply filters to the underlying dataset and thereby choose the social media posts to be visualized. The different visualizations are always kept consistent with these filters and with all other settings, i.e., the different parts of the graphical user interface provide different means to visualize one and the same dataset. As can be seen in Figure 2, there exists three main components for applying the filters: a time-line for filtering the time interval to be used, a tag cloud for filtering based on keywords, and the grey box located to the upper left that provides means to filter based on keywords, emotion classes, and data sources.

Table 9 NB classifier confusion matrix for the simplified problem

	Predicted class		
Actual class	**Anger**	**Fear**	**Positive**
Anger	41	5	8
Fear	3	43	8
Positive	19	7	28

An important part of the GUI, and a result of the earlier-mentioned design workshops, is the possibility to shift between the absolute probability distribution according to Figure 2 vis-à-vis the relative probability distribution as depicted in Figure 3. Most often, it will be important to visualize both the relative graph and the absolute graph since it will be easier to visualize the trend using the relative graph whilst the absolute graph is still needed in order to visualize, e.g., trends regarding how the total volume of posts vary.

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. The confusion matrices also indicate that even better accuracy could have been achieved using a simple ensemble combining the output from the rule-based and machine learning-based algorithms.

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 [28], better accuracy can most likely be achieved. More training data would probably also improve the accuracy, but the high cost in terms of manpower needed for the creation of even larger training datasets needs to be taken into account. Additionally, the learned classifiers ought to be evaluated on other datasets in order to test the generalizability of the obtained results.

Some of the classification errors were a result of the annotators receiving 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

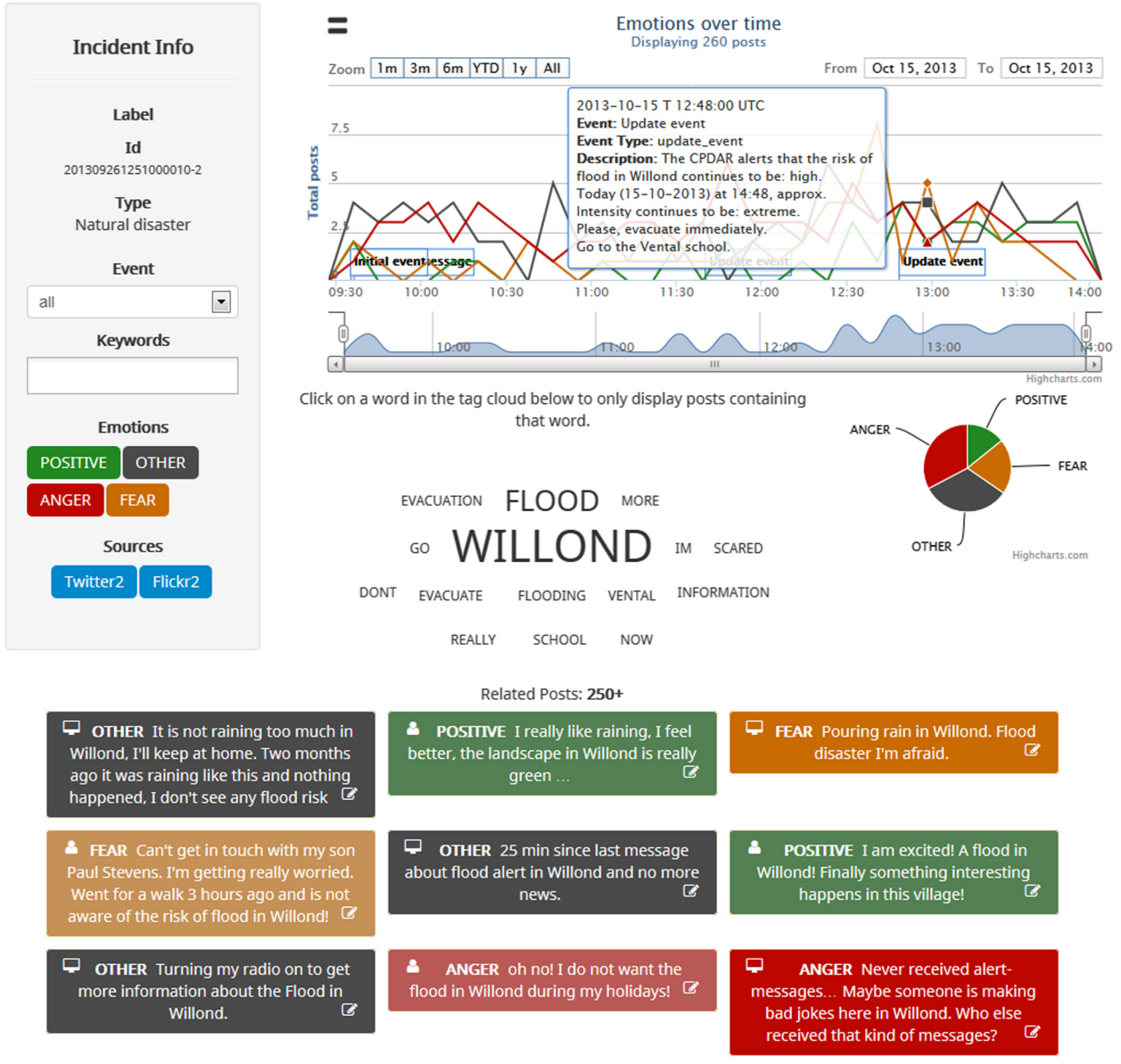

Figure 2 The figure shows the Alert4All "screening of new media" tool visualizing a fictive scenario which was used during the final demonstration in Munich during autumn 2013. As part of the graph, one can see the different messages that have been sent during the development of the crisis. Also, the content of one of these messages is shown due to the mouse pointer being positioned at this location.

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 domain 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. [20,21]). 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. [16,29]). Similarly, the tool and algorithms described in this article could also be adapted to be used for evaluating the effects of information campaigns or psychological operations during military missions [30].

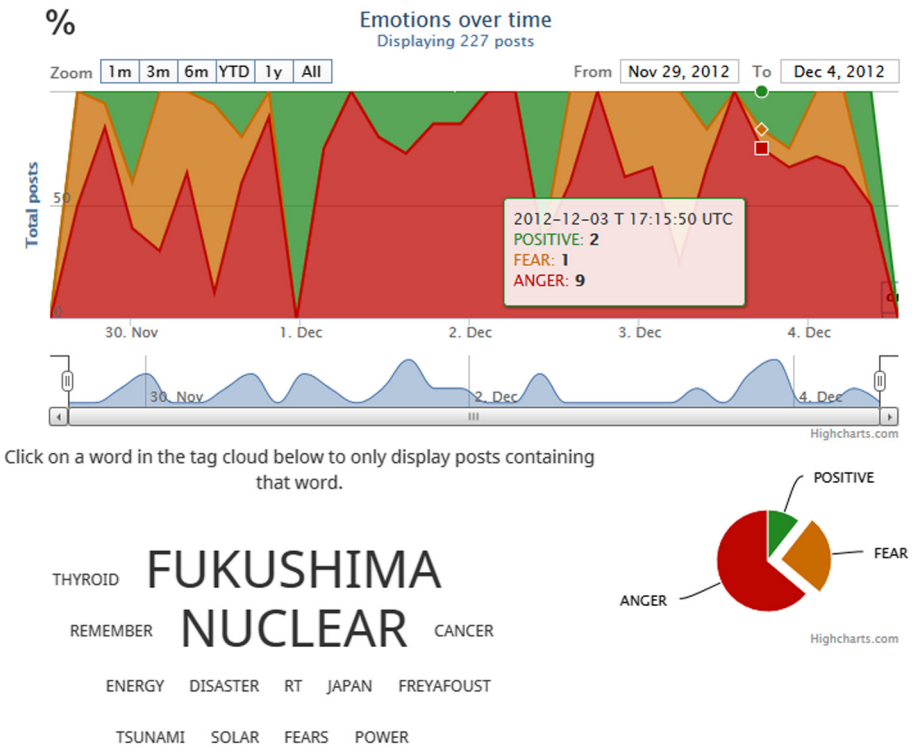

Figure 3 A snippet from the Alert4All "screening of new media" tool showing the relative probability distribution of emotions within a dataset gathered during the Fukushima disaster. The snippet shows a close-up of the graph component seen to the upper right in Figure 2, but now showing the relative distribution of emotions in a real scenario.

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 dataset 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 performs 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 aggregate level the citizens' emotions and attitudes toward 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 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. Investigating this requires that one retrieves and tags new datasets to test the classifiers on. Doing this for several different crisis types and then applying the same classifiers, should make it possible 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.

Endnote

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

Appendix: instructions given 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.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors drafted, read and approved the final manuscript.

Acknowledgments
We would like to thank Roberta Campo, Luísa Coelho, Patrick Drews, Montserrat Ferrer Julià, Sébastien Grazzini, Paul Hirst, Thomas Ladoire, Håkan Marcusson, Miguel Mendes, María Luisa Moreo, Javier Mulero Chaves, Cristina Párraga Niebla, Joaquín Ramírez, and Leopoldo Santos Santos for their effort during the tagging process.
This work has been supported by the European Union Seventh Framework Programme through the Alert4All research project (contract no 261732), and by the research and development program of the Swedish Armed Forces.

Author details
[1]FOI Swedish Defence Research Agency, Stockholm, Sweden. [2]KTH Royal Institute of Technology, Stockholm, Sweden. [3]Avanti Communications Group plc, London, UK. [4]Ericsson, Stockholm, Sweden.

Received: 10 February 2014 Accepted: 20 June 2014
Published online: 28 August 2014

References

1. A Zielinski, U Bugel, Multilingual analysis of Twitter news in support of mass emergency events, in *Proceedings of the Ninth International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)* (Vancouver, Canada, 2012)
2. S-Y Perng, M Buscher, R Halvorsrud, L Wood, M Stiso, L Ramirez, A Al-Akkad, Peripheral response: Microblogging during the 22/7/2011 Norway attacks, in *Proceedings of the Ninth International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)* (Vancouver, Canada, 2012)
3. R Thomson, N Ito, H Suda, F Lin, Y Liu, R Hayasaka, R Isochi, Z Wang, Trusting tweets: The Fukushima disaster and information source credibility on Twitter, in *Proceedings of the Ninth International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)* (Vancouver, Canada, 2012)
4. J Yin, A Lampert, MA Cameron, B Robinson, R Power, Using social media to enhance emergency situation awareness. IEEE Intell Syst. **27**(6), 52–59 (2012). doi:10.1109/MIS.2012.6
5. A Nagy, J Stamberger, Crowd sentiment detection during disasters and crises, in *Proceedings of the Ninth International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)* (Vancouver, Canada, 2012)
6. S Verma, S Vieweg, WJ Corvey, L Palen, JH Martin, M Palmer, A Schram, KM Anderson, Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency, in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (Barcelona, Spain, 2011), pp. 385–392
7. MR Endsley, Toward a theory of situation awareness in dynamic systems. Hum Factors. **37**(1), 32–64 (1995)
8. American Red Cross, The American Red Cross and Dell launch first-of-its-kind social media digital operations center for humanitarian relief. Press release 7 March 2012

9. C Párraga Niebla, T Weber, P Skoutaridis, P Hirst, J Ramírez, D Rego, G Gil, W Engelbach, J Brynielsson, H Wigro, S Grazzini, 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, 2011)
10. H Artman, J Brynielsson, BJE Johansson, 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, 2011)
11. S Nilsson, J Brynielsson, M Granåsen, C Hellgren, S Lindquist, M Lundin, M Narganes Quijano, 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, 2012)
12. F Johansson, J Brynielsson, M Narganes Quijano, Estimating citizen alertness in crises using social media monitoring and analysis, in *Proceedings of the 2012 European Intelligence and Security Informatics Conference (EISIC 2012)* (Odense, Denmark, 2012), pp. 189–196. doi:10.1109/EISIC.2012.23
13. B Liu, ed. by N Indurkhya, FJ Damerau, Sentiment analysis and subjectivity, in *Handbook of Natural Language Processing, Chapman & Hall/CRC Machine Learning & Pattern Recognition Series.* 2nd edition. Chap. 26 (Taylor & Francis Group, Boca Raton, Florida, 2010), pp. 627–666
14. B Pang, L Lee, Opinion mining and sentiment analysis. Foundations Trends Inf Retrieval. **2**(1–2), 1–135 (2008). doi:10.1561/1500000011
15. B Pang, L Lee, S Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in *Proceedings of the Seventh Conference on Empirical Methods in Natural Language Processing (EMNLP-02)* (Philadelphia, Pennsylvania, 2002), pp. 79–86. doi:10.3115/1118693.1118704
16. K Glass, R Colbaugh, Estimating the sentiment of social media content for security informatics applications. Secur Inform. **1**(3) (2012). doi:10.1186/2190-8532-1-3
17. A Pak, P Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, 2010), pp. 1320–1326
18. L Barbosa, J Feng, Robust sentiment detection on Twitter from biased and noisy data, in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)* (Beijing, China, 2010), pp. 36–44
19. C Strapparava, R Mihalcea, Learning to identify emotions in text, in *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC'08)* (Fortaleza, Brazil, 2008), pp. 1556–1560. doi:10.1145/1363686.1364052
20. A Abbasi, H Chen, S Thoms, T Fu, Affect analysis of web forums and blogs using correlation ensembles. IEEE Trans Knowl Data Eng. **20**(9), 1168–1180 (2008). doi:10.1109/TKDE.2008.51
21. A Abbasi, H Chen, Affect intensity analysis of dark web forums, in *Proceedings of the Fifth IEEE International Conference on Intelligence and Security Informatics (ISI 2007)* (New Brunswick, New Jersey, 2007), pp. 282–288. doi:10.1109/ISI.2007.379486
22. J Brynielsson, F Johansson, A Westling, Learning to classify emotional content in crisis-related tweets, in *Proceedings of the 11th IEEE International Conference on Intelligence and Security Informatics (ISI 2013)* (Seattle, Washington, 2013), pp. 33–38. doi:10.1109/ISI.2013.6578782
23. J Brynielsson, F Johansson, 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, 2013), pp. 559–568. doi:10.1007/978-3-642-39226-9_61
24. GA Miller, WordNet: A lexical database for English. Commun ACM. **38**(11), 39–41 (1995). doi:10.1145/219717.219748
25. M Hall, E Frank, G Holmes, B Pfahringer, P Reutemann, IH Witten, The WEKA data mining software: An update. ACM SIGKDD Explorations Newsl. **11**(1), 10–18 (2009). doi:10.1145/1656274.1656278
26. A McCallum, K Nigam, A comparison of event models for naive Bayes text classification, in *AAAI/ICML-98 Workshop on Learning for Text Categorization* (Madison, Wisconsin, 1998), pp. 41–48
27. JC Platt, ed. by B Schölkopf, CJC Burges, and AJ Smola, Fast training of support vector machines using sequential minimal optimization, in *Advances in Kernel Methods: Support Vector Learning.* Chap. 12 (MIT Press, Cambridge, Massachusetts, 1999), pp. 185–208
28. C Strapparava, A Valitutti, WordNet-Affect: an affective extension of WordNet, in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)* (Lisbon, Portugal, 2004), pp. 1083–1086
29. F Johansson, J Brynielsson, P Hörling, M Malm, C Mårtenson, S Truvé, M Rosell, Detecting emergent conflicts through web mining and visualization, in *Proceedings of the 2011 European Intelligence and Security Informatics Conference (EISIC 2011)* (Athens, Greece, 2011), pp. 346–353. doi:10.1109/EISIC.2011.21
30. J Brynielsson, S Nilsson, M Rosell, Feedback from social media during crisis management (in Swedish). Technical Report FOI-R--3756--SE, Swedish Defence Research Agency, Stockholm, Sweden, December 2013

doi:10.1186/s13388-014-0007-3
Cite this article as: Brynielsson *et al.*: **Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises.** *Security Informatics* 2014 **3**:7.