A Fine-Grained Sentiment Analysis Approach for Detecting Crisis Related Microposts

Axel Schulz
Technische Universität Darmstadt
aschulz@tk.informatik.tu-darmstadt.de

Tung Dang Thanh
Technische Universität Darmstadt
thanhtung.nov@gmail.com

Heiko Paulheim
Universität Mannheim
heiko@informatik.uni-mannheim.de

Immanuel Schweizer
Technische Universität Darmstadt
schweizer@tk.informatik.tu-darmstadt.de

ABSTRACT

Real-time information from microposts like Twitter is useful for applications in the crisis management domain. Currently, that potentially valuable information remains mostly unused by the command staff, mainly because the sheer amount of information cannot be handled efficiently. Sentiment analysis has been shown as an effective tool to detect microposts (such as tweets) that contribute to situational awareness. However, current approaches only focus on two or three emotion classes. But using only tweets with negative emotions for crisis management is not always sufficient. The amount of remaining information is still not manageable or most of the tweets are irrelevant. Thus, a more fine-grained differentiation is needed to identify relevant microposts. In this paper, we show the systematic evaluation of an approach for sentiment analysis on microposts that allows detecting seven emotion classes. A preliminary evaluation of our approach in a crisis related scenario demonstrates the applicability and usefulness.

Keywords

Emergency Management, Sentiment Analysis, Microposts, Twitter, Machine Learning

INTRODUCTION

Twitter has become very popular during the last years. Research has shown that tweets provide valuable real-time information for decision-making during crisis situations (Vieweg, Hughes, Starbird, Palen, 2010). However, this information is not directly usable, because of the sheer amount of information. For example only one out of a thousand tweets is crisis related in a dataset we crawled during hurricane Sandy in 2012. Besides searching for relevant keywords or topic and event extraction, one option to make this information usable is to apply sentiment analysis (Pang & Lee, 2006) for differentiating important information from unimportant one. As sentiment analysis deals with expressions of emotions in texts, it could help to detect people in danger during crisis situations, like the advent of hurricane Sandy. In that case, selecting and analyzing tweets with negative emotions over tweets with positive emotions may help to detect microposts contributing to situational awareness (SA) more efficiently.

Detecting sentiments in tweets is difficult as tweets are unstructured and contain colloquial language. Furthermore, since tweet messages are restricted to 140 characters, users may have a propensity to use abbreviations, slang, or emoticons to shorten the text. This issue can lead to unusual text and makes it difficult to detect regularities. State-of-the-art approaches already address the problem of identifying sentiments on microposts. Recent approaches (Barbosa & Feng, 2010; Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Jiang, Yu, Zhou, Liu, & Zhao, 2011) propose advanced methods for classifying sentiments in microposts, taking typical features of microposts, like re-tweets, hashtags, replies, links, and emoticons, into account. In this case, Jiang et al. (Jiang, Yu, Zhou, Liu, & Zhao, 2011) achieve the highest accuracy with 68.3% in a three class classification problem. Furthermore, Nagy and Stamberger (2012) propose a simple classification method to detect sentiments in tweets created during crisis situations. The integration of this simple classification method with a machine learning model shows promising results on a three class classification problem.

Current approaches utilize machine learning algorithms to classify tweets in two or three classes: positive, negative, and neutral. During our evaluations we found out that these classes are not always sufficient, as the amount of information is still not manageable, if, e.g., only tweets with negative emotions are used. In this case, we propose a more detailed differentiation of human emotions into seven sentiment classes according to Ekman

(Ekman, 1992): anger, disgust, fear, happiness, sadness, surprise, and neutral. In our opinion this differentiation is necessary, because differentiating sadness ("Sick during holidays. I'm so unlucky") from fear ("The fire is approaching faster. So afraid") provides a better understanding of the situation at hand.

In this article, we contribute the first approach to identify seven emotion classes on microposts. It can be used to identify tweets contributing to situational awareness. Our approach is based on the comparison of different machine learning algorithms and an analysis of various features useful for detecting sentiments on microposts. A preliminary evaluation of our classifier on tweets gathered during the hurricane Sandy in October 2012 shows promising results towards detecting highly crisis relevant information.

This paper is organized as follows: Section II presents our approach, followed by an evaluation in Section III. In Section IV the preliminary evaluation on a crisis related set of microposts is described. Section V closes with the conclusion and a research outlook.

APPROACH

In Figure 1 our pipeline for detecting sentiments in tweets is shown. We first preprocess all tweets, before extracting relevant features. As a last step, we use those features to train different classifiers for sentiment analysis.

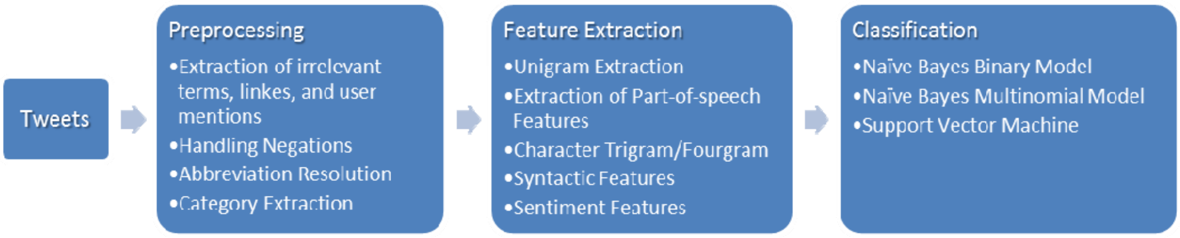

Figure 1: Reference pipeline for detecting sentiments in microposts.

First, every tweet is preprocessed. In the initial preprocessing step, irrelevant terms like links and user mentions (e.g. @Peter) are removed or replaced as we suppose they do not have any effect to the emotion expression in tweets. Words are normalized using the Porter-stemming[1] functions. All stop words are removed as they have only small influence on the machine learning algorithm. Furthermore, negations are detected, replaced with the term "NOT_", and appended before the following word. This is necessary, to lower the probability of a tweet like "No electricity is not funny jeez" to be misclassified as a positive opinion due to the presence of the word "funny".

After finishing the initial preprocessing steps, we extract general categories from the microposts. Named entities like "Barack Obama" or "Boston" are identified using the OpenCalais API[2] and replaced by their corresponding category ("Person" or "Place" in that case). At last, since tweet messages are restricted to a length of 140 characters, users tend to use abbreviations to shorten their messages. To deal with this problem, a dictionary of 5331 abbreviations was created using the data from noslang.com[3].

After preprocessing, several features are extracted:

- **Word unigram extraction:** A tweet is represented as a set of words. We use two approaches: a vector with the frequency of words, and a vector with the occurrence of words (as binary values).
- **Extraction of Part-of-speech features:** As words have different meanings depending on how they are used in tweets, we apply Part-of-speech Tagging (POS) using DKPro[4]. The POS-label is appended before the word E.g. "So happy" will become "RB_so JJ_happy".
- **Character trigram and fourgram extraction:** A string of three respective four consecutive characters in a tweet message is used as a feature. For example, if a tweet is: "Today is so hot. I feel tired" then

[1] http://tartarus.org/martin/PorterStemmer/

[2] http://www.opencalais.com/

[3] http://www.noslang.com

[4] http://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/

the following trigrams are extracted: “tod”, “oda”, “day”, “ay “, “y I” and so forth. To construct the trigram respective fourgram list, all the special characters, which are not a letter, space character, or number, are removed.

  - **Extraction of syntactic features:** Along with the features directly extracted from the tweet, several syntactic features that have not been used in related work before are expected to improve the performance of our approach. E.g. people might tend to use long tweets while expressing emotions, the repeated use of punctuations like “!”, or the use of multiple capitalized words may provide a strong indication for emotion. In this case, we extract the following features: the number of words in a tweet, the number of “!” and “?” in a tweet, and the number of capitalized characters.

- **Extraction of sentiment features:** To express opinions, people tend to use special words, phrases, or smileys. In this case, we extract positive and negative weights for words obtained from the AFINN word list (Nielsen, A., 2011) and SentiWordNet (Esuli & Sebastiani, 2006). Furthermore, the number of smileys that belong to either a positive or a negative category is calculated based on an emoticon library, we created manually based on 63 smileys extracted from Wikipedia[5] and labeled according to our classes.

The different features are combined and evaluated using three classifiers. For classification, the machine learning library Weka (Witten & Frank, 2005) is used. We compare a Naïve Bayes Binary Model (NBB), the Naïve Bayes Multinomial Model (NBM), and a classifier based on a Support Vector Machine (SVM).

EVALUATION

For evaluating an appropriate set of methods to detect multiple classes of sentiments in tweets, we use three data sets. Compared to related approaches, which have been trained on specific incident types, we train a classifier based on everyday tweets to avoid a bias towards a specific incident type.

- **SET1**: The first set consists of English tweets collected in Seattle at March 6, 2012. 200 randomly selected tweets were labeled into seven classes by voluntary participants in an online survey: “Anger”, “Disgust”, “Fear”, “Happiness”, “Sadness”, “Surprise”, and “None of those”. For the final baseline, we chose only tweets that were labeled identically by more than 50% of the users, resulting in a set with 114 English tweets. In this set, each tweet was labeled by at least eight persons.

- **SET2**: A second data set with a total of 2,000 randomly selected English tweets was created based on tweets crawled at March 06, 2012 in Seattle. In this case, nine participants participated in the study using Mechanical Turk[6] and each tweet was labeled by one person. The approval rates of these users are at least 95%. In comparison to the first data set, we added the “Cannot decide” category, which prevents the user from choosing a wrong answer. Furthermore, we divided “Surprise” into two sub-categories: “Surprise with positive meaning” and “Surprise with negative meaning”. The differentiation is necessary, because when a user is surprised, it could be a positive or negative emotion. This differentiation is important for creating SET2_GP. After removing tweets in the “Cannot decide” aggregating the two surprise categories to one “Surprise” category, a total of 1951 tweet messages remained.

- **SET2_GP**: To use SET2 also for a 3-class problem, the categories from SET2 were reorganized into three classes: positive, negative, neutral by grouping “Disgust”, “Fear”, “Sadness”, “Surprise with negative meaning” into the negative class, “Happiness”, “Surprise with positive meaning” into the positive class, and “None” into the neutral class. After grouping the tweets as describing above we obtain 872 positive tweets, 598 negative tweets, and 481 neutral tweets.

To measure the performance of the classification approaches, we report the following metrics:

  - Accuracy: Number of the correctly classified tweets divided by total number of tweets.
  - Averaged Precision: Is calculated based on the Precision of each class (how many of our predictions for a class are correct).
  - Averaged Recall: Is calculated based on the Recall of each class (how many tweets of a class are correctly classified as this class).

The results are calculated using stratified 10-fold cross validation.

[5] http://en.wikipedia.org/wiki/List_of_emoticons

[6] https://www.mturk.com/mturk/help?helpPage=overview

Results

For creating the results, we evaluated all combinations of classification methods and features. In Table 1 the best results for each classification method and feature combination are provided for the 7-classes problem as well as the 3-classes problem. The results show the optimal combinations of features and methods that provide the highest accuracy. We are able to identify seven classes of sentiments with accuracy of 65.79% and three classes with accuracy of 64.07%. The results on SET2 are 56.40%, which could be because each tweet was labeled by only one person.

	7-classes						3-classes		
	SET1			SET2			SET2_GP		
Accuracy	0.658	0.605	0.657	0.564	0.503	0.535	0.641	0.566	0.626
Avg. Precision	0.615	0.519	0.597	0.482	0.45	0.489	0.645	0.565	0.625
Avg. Recall	0.658	0.605	0.658	0.564	0.504	0.535	0.641	0.566	0.625
F-Measure	0.61	0.525	0.598	0.492	0.394	0.505	0.64	0.564	0.624
Class. Method	NBB	NBM	SVM	NBM	NBB	SVM	NBM	NBB	SVM
Unigram	x		x	x	x	x	x	x	x
Syntactic Features	x		x		x				
Sentiment Features		x				x		x	x
POS Tagging				x		x	x	x	x
Character tri-gram		x							

Table 1: Overview of evaluation results for 7-classes problem and 3-classes problem

The NBB model outperforms the other classifiers on SET1. On the other datasets, NBM provides better results, because of the increasing vocabulary size. As SET1 has only 114 instances with 477 different unigrams, the occurrence of unigram in this set is sparse, thus the binary model seems to be suitable in the case of SET1. The result on SET1 points out that adding the syntactic features improves the accuracy of the method.

The results also show that applying a POS tagger on SET1 is not suitable. As this dataset is relatively small, the results are strongly dependent on the quality of the POS tagger. Due to the identified tags of the POS tagger, the vocabulary would become diverse and the result based on the new feature set may deteriorate sharply. For example, the feature set contains four different forms of the verb "have": vh_have (original verb), vhp_have (have to do sth), vhg_have (verb –ing form), and vhz_have (has). In contrast, without using a POS tagger, those forms are normalized to "have" after stemming.

Using NBM on SET2 and SET2_GP together with POS-tagging provides the best results. However, the syntactic and sentiment features are not valuable in this case. This could be explained, because these additionally added features depend linearly on the existing features, which has a big influence on the computation of the class probabilities using NBM. E.g., the number of positive or negative words is dependent on the sum of the number of several words in a tweet, which results in a bias towards these words. In all the cases with unigram (w/o POS tagging), adding those syntactic and semantic features slightly reduced the accuracy rate. Those results lead to the conclusion, that the method using NBB model is a suitable choice for the corpus with the small vocabulary size. With a larger corpora using NBM model results in a better performance.

PRELIMINARY EVALUATION IN A CRISIS RELATED SCENARIO

To demonstrate how our approach can be applied to practical scenarios, we crawled a dataset of 150,000 tweets during hurricane Sandy in October 2012 using the Twitter Search API. The dataset consists of tweets created at the East Coast of the United States. From this set, we selected 60 tweets contributing to situational awareness and 140 random, non-contributing tweets. We compared the accuracy of detecting tweets contributing to situational awareness using the classifier for seven sentiment classes trained on SET1 and the classifier for three sentiment classes trained on SET2_GP.

		Detected	Contributing to SA	Accuracy	Recall
7-classes	Fear	96	38	0.395	0.633
	Disgust	41	10	0.243	0.166
	Fear & Disgust	137	48	0.35	0.80
3-classes	Negative	41	12	0.292	0.20

Table 2: Accuracy of detecting tweets contributing to SA of 3-class and 7-class classifier.

The results in Table 2 show that both the accuracy and the recall of an approach using tweets labeled with "Fear" with a 7-class sentiment classifier are better than simply using tweets with negative sentiment using a 3-class sentiment classifier. Note that in the sample used for the evaluation, the accuracy of a random baseline would only be 30% (and in a real-world distribution, which contains more irrelevant tweets, the baseline would be even lower). The following example tweets are found with our approach, but not with a 3-class sentiment classifier:

- "Day 3. No power. Limited Food. Limited shelter. Must survive. #Sandy" [7-classes: Fear, 3-classes: Neutral]
- "Family members in burned-out neighborhood hard to reach after Sandy: Volunteer firefighter .http://t.co/k5DviFc7 #stamford #ct #topix" [7-classes: Fear, 3-classes: Neutral]
- "Coney island area has no electric, no street lamps or traffic signals." [7-classes: Fear, 3-classes: Neutral]

Additionally, tweets with the expression "Sadness", e.g., "Not having school is ruining this for me!" or "This is my winter song, December never felt so wrong", could easily be filtered out. Furthermore, tweets labeled with "Disgust" could also provide valuable information, though only fewer tweets in this class could be detected. Summarized, this underlines our initial hypothesis that a more fine-grained classification of microposts according to human emotions can contribute to information filtering in crisis management. While the evaluation is only preliminary, it shows promising results towards good estimations for detecting valuable microposts.

CONCLUSION

We presented a novel sentiment analysis approach for detecting seven sentiment classes on microposts. The preliminary evaluation of our classifier in a crisis related scenario has shown how a classifier could be used to detect relevant tweets. Thus, we contribute a mean for decreasing potential information overload of user-generated content. Based on our approach it is possible to detect more tweets that are relevant for decision-making in emergency management than using traditional sentiment analysis methods.

Future work needs to be done to enhance the current machine learning models. In this case, first, a much larger labeled training set is needed. Second, post-processing the classified microposts can further reduce the set. E.g., specific words for crisis related tweets like "hurricane", "earthquake", or "fire" can be used to extract relevant tweets. Furthermore, other means for separating relevant information from irrelevant ones have to be found and combined with the presented approach.

ACKNOWLEDGMENTS

This work has been partly funded by the German Federal Ministry for Education and Research (BMBF, 13N10712, 01jS12024) and by the German Science Foundation (DFG, FU 580/2).

REFERENCES

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. Proceedings of the Workshop on Languages in Social Media. Portland, Oregon.
2. Barbosa, L., & Feng, J. (2010). Robust sentiment detection on Twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China.
3. Ekman, P. (1992) An argument for basic emotions. Cognition & Emotion, 6, 3-4, 169-200.
4. Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation. Genova, IT.

5. Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter Sentiment Classification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon.

6. Nagy, A. and Stamberger, J (2012) Proceedings of the 9th International ISCRAM Conference, Vancouver, CA.

7. Nielsen, A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microposts. Journal of the International Linguistic Association, 93-98.

8. Pang, B., & Lee, L. (2006). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 91-231.

9. Vieweg, S., Hughes, A. L., Starbird, K., Palen, L. (2010) Micropostging during two natural hazards events. Proceedings of the 28th international conference on Human factors in computing systems, Atlanta, GA.

10. Witten, I. H. and Frank, E. (2005). Data Mining: Practice Machine Learning Tools and Techniques, 2nd Edition, San Francisco, Morgan Kaufmann Publishers.