Sentiment Analysis of German Social Media Data for Natural Disasters

Gayane Shalunts
SAIL LABS Technology AG, Vienna, Austria
Gayane.Shalunts@sail-labs.com

Gerhard Backfried
SAIL LABS Technology AG, Vienna, Austria
Gerhard.Backfried@sail-labs.com

Katja Prinz
SAIL LABS Technology AG, Vienna, Austria
Katja.Prinz@sail-labs.com

ABSTRACT

Analysis of social media and traditional media provides significant information to first responders in times of natural disasters. Sentiment analysis, particularly of social media originating from the affected population, forms an integral part of multifaceted media analysis. The current paper extends an existing methodology to the domain of natural disasters, broadens the support of multiple languages and introduces a new manner of classification. The performance of the approach is evaluated on a recently collected dataset manually annotated by three human annotators as a reference. The experiments show a high agreement rate between the approach taken and the annotators. Furthermore, the paper presents the initial application of the resulting technology and models to sentiment analysis of social media data in German, covering data collected during the Central European floods of 2013.

Keywords

Sentiment analysis, natural disasters, social media

INTRODUCTION

Sentiment analysis is an active research area, aiming to identify public opinions and attitudes in certain contexts. Whereas the term seems to implicate the identification of a basic set of emotions (such as anger, disgust, fear, happiness, sadness, surprise), the overwhelming majority of sentiment analysis methods solves a simpler task by classifying texts as either being positive or negative. Despite the considerable amount of literature in the field, most publications address the domains of movie- and product reviews in English only.

Sentiment analysis methods can roughly be divided into two categories: machine-learning- and lexicon-based methods (Gonçalves, Araujo, Benevenuto and Cha, 2013). Machine learning methods are typically implemented as supervised binary (i.e. positive or negative) classification approaches, in which labeled data is employed to train classifiers (Gonçalves et al, 2013; Pang, Lee, and Vaithyanathan, 2002). However, this dependence on a labeled data-set also forms their major drawback, as labeling is usually costly, time-intensive and even prohibitive in some cases. In contrast, lexicon-based methods use a predefined set of words and patterns (often referred to as a *sentiment dictionary* or *lexicon*) associating each entry with a specific sentiment and do not require any labeled training data. The challenge, however, lies in obtaining or designing an appropriate sentiment lexicon. Lexicon-based methods are often tuned towards specific target domains, media-types and the respective style of language present (e.g. formal language on traditional media, rather colloquial language on social media).

The authors of (Gonçalves et al, 2013) compare eight state-of-the-art sentiment analysis methods. All experiments are carried out using two English datasets of Online Social Networks messages. The methods compared are SentiWordNet (Esuli and Sebastiani, 2006), SASA (Wang, Can, Kazemzadeh, Bar and Narayanan, 2012), PANAS-t (Gonçalves, Benevenuto and Cha, 2013), Emoticons, SentiStrength (Thelwall, Buckley, Paltoglou, Cai and Kappas, 2010), LIWC (Tausczik and Pennebaker, 2010), SenticNet (Cambria, Speer, Havasi and Hussain, 2010) and Happiness Index (Dodds and Danforth, 2009). They find that existing sentiment analysis methods have varying levels of applicability on *real-world* events. Regarding the predicted polarity, the different methods vary widely in their agreement, indicating that the same social media text is interpreted very differently depending on the choice of a sentiment analysis method. All methods examined differ substantially in their sentiment prediction of notable social events. For an airplane crash, half of the methods predicted the majority of relevant tweets to contain positive affect, which seems rather implausible for this kind of incident.

Only a few state-of-the-art sentiment analysis methods support German. The authors in (Remus, Quasthoff and Heyer, 2010) present a method called SentimentWortschatz (SentiWS) targeting the financial domain. Their sentiment lexicon is created based on the General Inquirer (GI) lexicon (Stone, Dunphy, Smith and Ogilvie, 1966) by semiautomatic translation into German using Google Translate[1] and subsequent manual revision. Another sentiment classification method in German, introduced by (Momtazi, 2012), targets the domain of German celebrities. Their approach utilizes SentiStrength (Thelwall et al, 2010) and permits the classification also of mixed sentiments (both positive and negative). As in the first case, an English opinion lexicon was first automatically translated into German and then manually revised. No work on sentiment-analysis within the scope of disasters has been carried out for German. Research on a similar domain, the demographic analysis of Twitter sentiment in English during hurricane Irene, is presented in (Mandel, Culotta, Boulahanis, Stark, Lewis and Rodrigue, 2012).

In our context, sentiment-analysis forms one component of a framework addressing cross-media, multimedia and multilingual communication during disasters (Backfried, Göllner, Quirchmayr, Rainer, Kienast, Thallinger, Schmidt and Peer, 2013). It forms an important part of a multifaceted media analysis approach and is integrated into the Sail Labs Media Mining System for Open Source Intelligence (Backfried, Schmidt, Pfeiffer, Quirchmayr, Glanzer and Rainer, 2012) with the aim to improve situational awareness during disasters. In particular, sentiment-analysis will be beneficial for the analysis process regarding the psycho-social situation and perceived state of the affected population.

The limitations of current sentiment-analysis systems regarding multilingual support and the domain of natural disasters form the starting point for our work. Analysis of state-of-the-art methods yielded SentiStrenth as the optimal framework to base our developments and extensions upon.

METHOD

SentiStrength is a lexicon-based method, for which the discriminative features are predefined lists of words and patterns including a sentiment lexicon with a positive or negative score associated with each entry. Additional features include boosters (intensifiers), negations, idioms and phrases, emoticons, slang expressions and abbreviations. The emotion emphasis expressed in social media by repeated letters, capitalization and exclamation marks is likewise detected. SentiStrength supports 14 languages; however, the support for languages other than English is somewhat limited.

SentiSAIL employs the same basic methodology as SentiStrength, but extends it by (1) the application to the domains of general news and natural disasters, (2) improving the support of multiple languages (German, Russian and to some extent also for English), (3) handling of data from both traditional as well as social media, (4) processing of short texts as well as full articles and (5) a four-class classification scheme. Unlike the majority of sentiment analysis approaches in the literature, our method solves a dual classification task by classifying a text into one of the following 4 classes: positive, negative, mixed (both positive and negative) or neutral (neither positive nor negative). SentiSAIL does not aim to perform opinion mining or to distinguish between objective and subjective aspects of the observed data.

One particular extension concerns the support for stemming, which is especially important for the processing of inflective languages such as Russian or German. We improved stemming of the trilingual data files (lexica, etc.) and broadened the lexica by context terms manually extracted from a multilingual dataset comprising articles from general as well as natural disaster related news (112 English, German and Russian articles). As a result, the SentiSAIL sentiment lexicon for English grew from an original set of 2546 entries to 2824 entries. For German and Russian the sets grew from 1983 to 2142 and from 1684 to 2001 entries respectively.

SentiSAIL calculates both positive and negative sentiment scores for each input document by accumulating all feature scores correspondingly. Positive and negative scores of the complete text are obtained by segmenting and tokenizing of the input document and averaging the respective scores of the individual segments. The overall sentiment of a text is produced by a double threshold: classification of the *positive* and *negative* classes

[1] http://translate.google.com

is straightforward. Segments passing both thresholds are classified into the *mixed* class. Those failing both thresholds are classified as *neutral*. Tuning on a held-out dataset resulted in a setting of the thresholds to the minimum sentiment-expressing scores in our framework (however, depending on the domain, the thresholds may have to be set to different values to detect stronger or weaker sentiment). Social media features are parameterized and may be enabled or disabled during traditional media processing.

EXPERIMENTS AND RESULTS

As part of a larger corpus collection effort on the topic of media-coverage of natural disasters some 182k Tweets covering the period from 05/20/2013 to 06/23/2013 were collected through Twitter's streaming-API (the collection was performed via a set of iteratively refined hash-tags and user-accounts). In order to evaluate the performance of SentiSAIL on German social media data covering natural disasters, 500 tweets were selected randomly from this set. Three annotators each classified the tweets into one of the 4 categories. Human opinions on a particular event or phenomenon may vary widely depending on factors such as cultural background or emotional state. Because of this subjective nature of classification, combining the annotations of multiple persons to create a joint reference is common practice in the field. The performance of a method is then evaluated against the average inter-annotator agreement rate.

For evaluation purposes, we apply the following scheme: since the 4 classes are correlated, the confusion matrix is first calculated and subsequently multiplied with the following matrix in Table 1 to take this correlation into account. E.g. classification of a "Positive" text as "Negative" is *100 % wrong* (coefficient 0), whereas classifying it as "Neutral" or "Mixed" is only *50% wrong* (coefficient 0.5):

	Positive	Negative	Mixed	Neutral
Positive	1	0	0.5	0.5
Negative	0	1	0.5	0.5
Mixed	0.5	0.5	1	0
Neutral	0.5	0.5	0	1

Table 1. Weights for correlation of error-types

Table 2 presents the agreement rates of human annotators and SentiSAIL. Whereas there is a 92.24% agreement rate between the annotators, SentiSAIL's performance yields an overall agreement rate of 91.12%. Our method's performance can thus be considered to be on par with human performance.

For comparison with SentiStrength, we evaluated SentiSAIL using the original SentiStrength features (sentiment lexicon, boosters, negations, etc.). As presented in the bottom line of Table 2, the average agreement rate using that specific setup is 45.11%. This low number can be explained by the fact that the German sentiment lexicon of SentiStrength includes entries with fixed endings only, not addressing German inflection properly. Furthermore, words from the domain of natural disasters are missing completely, e.g. Naturkatastrophe (natural disaster), Hochwasser (flood), etc.

	Annotator 1	Annotator 2	Annotator 3	Average
Annotator 1	-	92.71%	92.61%	92.24%
Annotator 2	-	-	91.40%	
SentiSAIL	94.53%	89.47%	89.37%	91.12%
SentiSAIL with SentiStrength features	44.94%	44.33%	46.05%	45.11%

Table 2. Inter-annotator and SentiSAIL-annotators agreement rate comparison

Figure 1 presents the results of applying SentiSAIL to the complete set of tweets collected during the period of May 20th to June 23rd 2013. For visualization, the tweets classified as *mixed* were added to both, the *negative* as well as the *positive* classes. Natural disasters and particularly floods are highly negative phenomena. Consequently negative sentiment dominates the dataset. Negative and positive tweets generally rise and fall in parallel with the overall volume of tweets, which in turns corresponds in activity to the most dramatic days of the disaster. The absolute maximum of the negative tweets coincides with the peak date of the flood for Southern Germany and Austria at the beginning of June 2013.

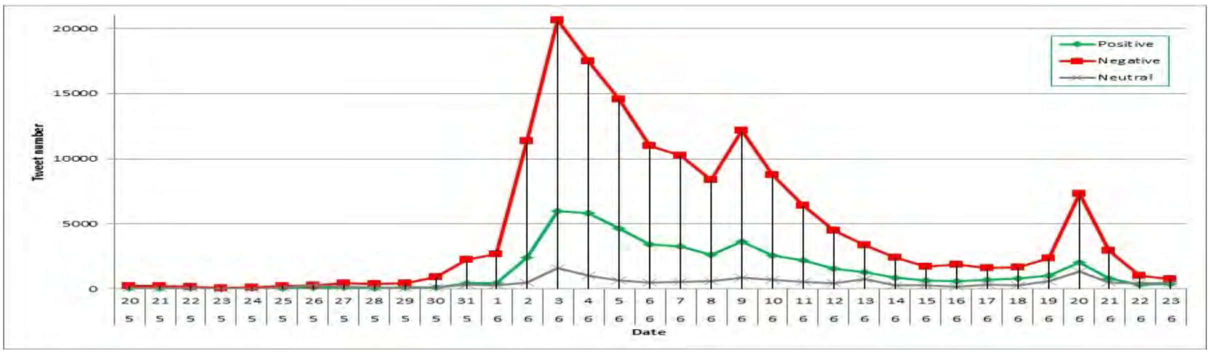

Figure 1. Temporal sentiment analysis of tweets in the period May 20th - June 23rd

The ratio of negative vs. positive tweets peaks on 05/30 just slightly before the peak of flooding in Southern Germany and Austria, possibly indicating anxiety of what was yet to come as opposed to negative sentiment of what had become a fact later on in time.

The prevailing sentiment among the 50 most frequent Twitter users in the dataset is depicted in Figure 2. As expected, negative sentiment dominates among all users, indicating the event explored to be subjectively negative. Only a few, selected accounts exhibit higher ratios of positive or neutral sentiment. Users not following the general trend, such as user *waspassierteam* in Figure 2, could merit closer inspection: in the particular case, the user turned out to be a bot tweeting historical weather data. This information in turn is used (among a series of other factors) for bot-detection in our framework.

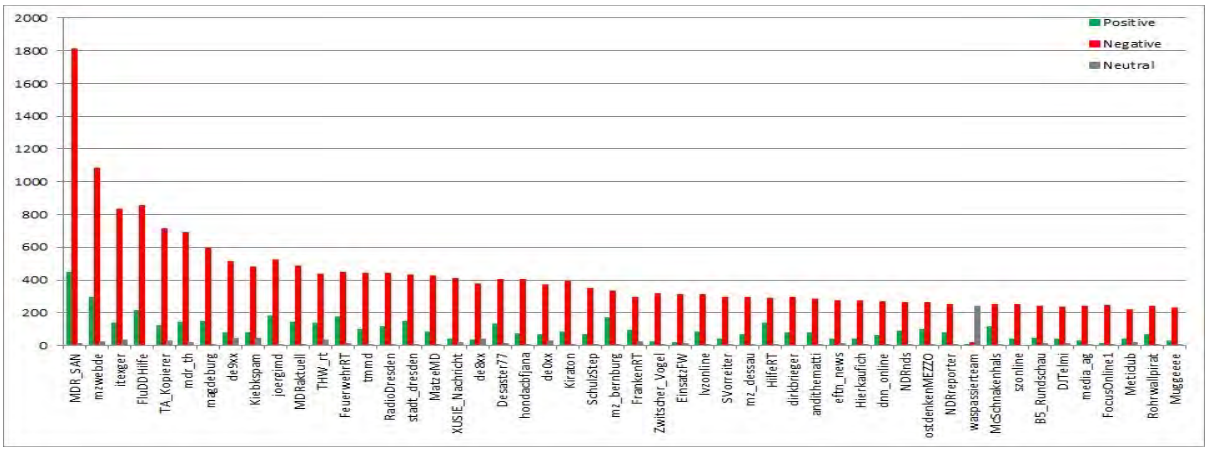

Figure 2. Sentiment classification of 50 most frequent Twitter users

CONCLUSION AND FUTURE WORK

The paper presents SentiSAIL, a tool based on SentiStrength, incorporating several extensions and modifications. It reports on the collection of a corpus covering the Central European floods of 2013 and applying sentiment analysis to social media in German related to this flood using SentiSAIL. The dominating sentiment in the data-set over the span of the disaster as well as across the most active twitter accounts is negative. Whereas this is to be expected, given the nature of the underlying event (a flood), there are some temporal patterns and specific users deviating from the general trends. Such deviation can be used e.g. in the detection of bots. Investigating the specific deviations, temporal variation and connections to the different disaster phases form areas of ongoing work. In particular, this work focuses on the roles of users based on their temporal behavior coupled with sentiment analysis. Within a larger context, we intend to establish links of communication between social media and traditional media taking into account sentiment analysis performed on all different media types. A further focus of investigation will be to link such patterns of communication to messages of resilience. Technical extensions, such as adding support for further languages, and continued data collection for natural disasters in Central Europe will be carried out to complement the above activities.

ACKNOWLEDGMENTS

This work is supported by the Austrian National Security Research Development Programme KIRAS.

REFERENCES

1. Backfried, G., Schmidt, C., Pfeiffer, M., Quirchmayr, G., Glanzer, M., Rainer, K. (2012) Open Source Intelligence in Disaster Management, EISIC 2012, Odense, Denmark
2. Backfried G., Göllner J., Quirchmayr G., Rainer K., Kienast G., Thallinger G., Schmidt C. and Peer A. (2013) Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project, *Proceedings of European Intelligence and Security Informatics Conference (EISIC)*, Uppsala, Sweden, 143-146.
3. Cambria, E., Speer, R., Havasi, C. and Hussain, A. (2010) SenticNet: A publicly available semantic resource for opinion mining, *AAAI Fall Symposium Series*,14-18.
4. Dodds, P. S. and Danforth, C. M. (2009) Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4), 441–456.
5. Esuli, Andrea and Sebastiani, Fabrizio (2006) SENTIWORDNET: A publicly available lexical resource for opinion mining, *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, 417–422.
6. Gonçalves, P., Araujo, M., Benevenuto, F. and Cha, M. (2013) Comparing and combining sentiment analysis methods, *Proceedings of the 1st ACM Conference on Online Social Networks (COSN)*, Boston, USA, ACM, 27-38.
7. Gonçalves, P., Benevenuto, F. and Cha, M. (2013) PANAS-t: A Pychometric Scale for Measuring Sentiments on Twitter. *CoRR*, abs/1308.1857.
8. Mandel Benjamin, Culotta Aron, Boulahanis John, Stark Danielle, Lewis Bonnie and Rodrigue Jeremy (2012) A Demographic Analysis of Online Sentiment During Hurricane Irene, *Proceedings of the 2nd Workshop on Language in Social Media*, Montreal, Canada, 27-36.
9. Momtazi, S. (2012) Fine-grained German sentiment analysis on social media, *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 1215-1220.
10. Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up?: sentiment classification using machine learning techniques, *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, 79-86.
11. Remus, R., Quasthoff, U. and Heyer, G. (2010) SentiWS – a publicly available German-language resource for sentiment analysis, *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC)*, Velletta, Malta, 1168-1171.
12. Stone, P. J., Dunphy, D.C., Smith, M.S. and Ogilvie, D.M. (1966) A computer approach to content analysis, MIT Press, Cambridge, MA.
13. Tausczik, Y. R. and Pennebaker, J. W. (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.
14. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, D. (2010) Sentiment strength detection in short informal text, *J. American Society for Information Science and Technology*, 61, 2544-2558.
15. Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S. (2012) A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. *ACL System Demonstrations*, 115–120.