Comparing Overall and Targeted Sentiments in Social Media during Crises

Saúl Vargas, Richard McCreadie, Craig Macdonald, and Iadh Ounis
{firstname.lastname}@glasgow.ac.uk
School of Computing Science, University of Glasgow, G12 8QQ, Glasgow, UK

Abstract

The tracking of citizens' reactions in social media during crises has attracted an increasing level of interest in the research community. In particular, sentiment analysis over social media posts can be regarded as a particularly useful tool, enabling civil protection and law enforcement agencies to more effectively respond during this type of situation. Prior work on sentiment analysis in social media during crises has applied well-known techniques for overall sentiment detection in posts. However, we argue that sentiment analysis of the overall post might not always be suitable, as it may miss the presence of more targeted sentiments, e.g. about the people and organizations involved (which we refer to as sentiment targets). Through a crowdsourcing study, we show that there are marked differences between the overall tweet sentiment and the sentiment expressed towards the subjects mentioned in tweets related to three crises events.

1 Introduction

Social media platforms are a popular medium for posting real-time discussions about world events. The use of social media during crises and emergencies such as social unrest, human-induced mass incidents and natural disasters has attracted the interest of the research community in recent years (Imran et al. 2014; Sakaki et al. 2010). For instance, tracking citizens' messages in social media can improve *situational awareness* (Schulz et al. 2013; Verma et al. 2011) and can help other citizens, as well as emergency response and law enforcement agencies to make decisions during such situations (Brynielsson et al. 2014).

One of the areas where social media can be helpful is the tracking of sentiments expressed during an event. Indeed, for crisis events, government bodies are often interested in tracking the sentiments of interest related to particular named entities or subjects (which we refer to as sentiment targets), such as emergency response agencies, politicians and companies. There exists a large body of work on analysing the general/overall sentiments expressed in social media posts (Agarwal et al. 2011; Santos et al. 2012). In contrast, there has been comparatively little examination of sentiments expressed for particular entities or subjects, mainly focusing on approaches to either classify the sentiment towards the entities themselves (Moilanen and Pulman 2009) or to use entities to enhance the sentiment scoring process (Jiang et al. 2011).

As a result, a broader question has been left unanswered, namely: whether (and if so, how) sentiment differs between the overall post sentiment and the sentiment expressed about entities/subjects (sentiment targets) within that post. If it differs, then approaches that perform a more detailed sentiment analysis than the classical ones will be needed. For instance, consider the following tweet about the Aurora Shooting in 2012:

> *"14 Dead in #theatershooting - Somehow, Obama will simultaneously blame this on both George W. Bush and Mitt Romney."*

As we can see, the overall sentiment in the tweet is neutral, as the user simply states his opinion without using terms that reveal any subjectiveness. There is however a clear negative sentiment towards Barack Obama. For the purpose of someone tracking discussions about Barack Obama, this difference between overall and targeted sentiment in the tweet is important.

In this paper, we contribute a fine-grained analysis over three crisis-related datasets comprised of Twitter posts, with the aim of determining whether sentiment often differs when considering the post as a whole and the sentiment target of that post. In particular, through an analysis of sentiment labels generated via a crowdsourced experiment, we compare the sentiments identified both when performing a classical overall sentiment labelling and a more targeted sentiment labelling for named entities/subjects (targets). We aim to answer the following research question: Are there often differences between the overall sentiment expressed within a social media post and the targeted sentiment within that post?

Our results show that there are marked differences between overall and targeted sentiments, illustrating the importance of properly analysing the sentiment towards the entities/subjects involved in crises, rather than relying on overall sentiment analysis techniques.

Event	Country	Language	#Tweets	Start date	End date
Aurora Shooting	USA	English	151,046	20/07/12	30/07/12
Hurricane Isaac	USA	English	238,165	28/08/12	07/09/12
Ebro Flood	Spain	Spanish	123,872	26/02/15	09/03/15

Table 1: Crisis-related datasets and their statistics.

2 Related Work

Sentiment analysis and opinion mining (Pang and Lee 2008) are active research areas. Indeed, discovering public sentiments and opinions is of great value in various fields such as reputation monitoring, marketing, recommendation and emergency management. For these different fields, sentiment analysis has been applied to a variety of textual sources, such as blogs (He et al. 2008), hotel reviews (Marcheggiani et al. 2014) and, more relevant to our case, social media posts such as tweets (Agarwal et al. 2011; Barbosa and Feng 2010; Jiang et al. 2011).

Notably, while most of the reviewed work on sentiment analysis focuses on estimating the overall sentiment of a text, some prior works have proposed a more detailed analysis of the sentiments within subsets of that text. For instance, Moilanen and Pullman (2009) consider the sentiment towards named entities appearing in documents. Meanwhile, Marcheggiani et al. (2014) examined aspect-orientated opinion (sentiment) mining for pre-defined information aspects (e.g. cleanliness, location, food) within hotel reviews. Similarly, Jiang et al. (2011) considered an approach for enhancing this sentiment classification by considering features about named entities. However, none of these works address the question of whether within a social media post, there is often a difference between the overall sentiment of a post and the sentiment expressed about the targets (entities/subjects) of that post. In this paper, we analyse and quantify the differences between overall and targeted sentiments over the key subjects involved in three different crisis events as reflected on Twitter. Furthermore, we show through experimentation the extent to which supervised sentiment classification techniques such as those used in the aforementioned works can be tailored to targeted sentiment classification.

3 Experimental Setup

Datasets: To evaluate the differences between overall and targeted sentiments we use three tweet datasets crawled via the Twitter Streaming API.[1] Each of these three tweet datasets relate to different crisis-related events, namely the Aurora Shooting (2012), the Hurricane Isaac (2012) and the Ebro River Flood (2015). These datasets cover two different types of crises, human-induced and natural disasters, and two languages, English and Spanish. Table 1 provides salient statistics about the three datasets. For the two 2012 events, a random tweet sample (approximately 1% of the total the full Twitter stream) was collected during the time period of these events. These tweets were then filtered based on a set of keywords/phrases[2] to remove posts that do not discuss the two events. For example, terms/phrases such as "#Isaac", "Red Cross" and "Plaquemines Parish" were used to filter the Hurrican Issac dataset. For the 2015 dataset, similar types of keywords/phrases were used to collect a live tweet stream during the event, using the Twitter Streaming API. For all three datasets, the keywords used were defined by human annotators who were native speakers of the primary language of each event (English/Spanish).

Event	Sentiment Targets	#Tweets
Aurora Shooting	Aurora PD, Christian Bale, DC-Comics, FBI, James Holmes, Christopher Nolan, Barack Obama, Mitt Romney, Warner Bros	2,184
Hurricane Isaac	Army Corps of Engineers, B. Jindal, Nat. Guard, B. Obama, P. Bryant, Red Cross, R. Scott, R. Bentley, M. Romney	2,085
Ebro Flood	Aragonese Gov., Ebro Hydrographic Conf., Civil Guard, Civil Protection, Firemen, Police, M. Rajoy, Red Cross, L. F. Rudi, Spanish Gov., I. García Tejerina, Military Emergencies Unit	2,089

Table 2: Sentiment targets selected for each event and number of tweets mentioning them.

Subject selection and filtering: For these three datasets, we have manually selected lists of sentiment targets/subjects among the political figures, institutions, companies and celebrities that were involved in these events. These sentiment targets were selected based on the associated coverage of each event in news articles and on Wikipedia. We then automatically filtered the tweets from each event to include only those that made reference to the selected targets. To improve the accuracy of the filtering process, we considered multiple ways that a user might refer to each subject (e.g. Barack Obama, Obama, the president, POTUS) as well as Twitter handles (such as @BarackObama, @POTUS) to determine the presence of the sentiment targets within the tweets. The sentiment targets and the number of related tweets for each event are listed in Table 2.

Sentiment labelling: For the resulting tweets mentioning the selected subjects, we conducted two crowdsourced sentiment labelling tasks. In the first task, crowdsourced workers were asked to label a tweet according to the sentiment that the author of the tweet expresses: negative (anger, disgust, sadness, surprise, hatred, etc.), neutral (statements) or positive (happiness, gratitude, joy, love, pride, etc.). In the second task, a tweet and subject pair was given, and the worker is tasked with labelling the tweet as containing a negative, neutral or positive sentiment by the author of the tweet with respect to the given subject. For each of these tasks, three different crowd workers labelled each tweet or tweet and subject pair.

[1] `https://dev.twitter.com/streaming/overview`

[2] Hashtags, names of people, organisations or locations associated to each event.

Event	Sent.	#w	a ≥ 2	a $= 3$	Fleiss' κ
Aurora Shooting	overall	129	99.3%	73.1%	0.513
	targeted	53	99.7%	77.8%	0.519
Hurricane Isaac	overall	53	99.4%	69.3%	0.315
	targeted	56	98.4%	69.4%	0.330
Ebro Flood	overall	252	99.1%	63.4%	0.387
	targeted	191	99.2%	64.2%	0.377

Table 3: Statistics of the crowsourced experiment.

Crowdsourcing Configuration: For all crowdsourcing labelling tasks, we use the CrowdFlower crowdsourcing platform. The unit of assessment is a single page, which contains 5 tweets. For the US-based events, we restricted the geographical regions that could participate to only the United States, whereas for the Ebro Flood event, only Spanish-speaking workers were used. For both labelling tasks, we paid US $0.07 for each page of 5 tweets. The number of tweets a single worker could label was limited to 200. Worker quality was dynamically assessed against a gold standard set of 100 tweets per experiment labelled by the authors. Workers whose accuracy on these gold standard tweets dropped below 70% were ejected from the experiment.

Worker Agreement:The statistics of the workers and their agreement are shown in Table 3. From the table, we observe that the number of workers (#w) was high: above fifty over all tasks. Moreover, the percentage of tweets for which at least two (out of three) users agree (a ≥ 2), is above 98%, which indicates that there is little or no randomness in the answers given by the crowdsourced workers. We also see that the total agreement (a $= 3$) is reasonably high ($\geq 63\%$). Finally, the Fleiss' κ measure provides a statistical confirmation of the degree of agreement between several workers for each task, showing that we obtain fair (~ 0.3) to moderate (~ 0.5) agreement measurements. In general, these results indicate that the described crowdsourcing configuration produces good quality labels.

Reproducibility: The filtered dataset described above, as well as the associated crowdsourced labels used for evaluation are available as a free download:[3]

`http://dx.doi.org/10.5525/gla.researchdata.286`

Metrics: In order to report the differences between the sentiments identified from the two labelling tasks (overall sentiment labelling and subject targeted labelling) within a single metric, we report the conditional distribution of sentiment labels assigned to each tweet:

$$p(s|t) = L_{t,s}/L_t \quad (1)$$

where L_t is the number of labels assigned to the tweet t ($L_t \geq 3$) and $L_{t,s}$ the number of labels for tweet t that correspond to sentiment $s \in \{neg, neu, pos\}$. Differences between the two sentiment labelling tasks (overall vs. targeted) are then measured by means of the comparison between the individual distributions of sentiments in each experiment and an analysis of the joint distribution of sentiments in both tasks. The individual distribution of each sentiment space is calculated as the aggregate probability of each tweet being assigned to each sentiment class:

$$c(s) = \sum_{t \in T} p(s|t) \quad (2)$$

In turn, the joint distribution of sentiments is computed as the aggregate probability of each tweet being assigned to a pair of sentiments in the two tasks (e.g. the probability that a tweet had an overall sentiment of neutral but a targeted sentiment of positive):

$$c(s_t, s_o) = \sum_{t \in T} p(s_t|t)\ p(s_o|t) \quad (3)$$

s_o/s_t denote overall/targeted sentiments, respectively.

4 Experimental Evaluation

We now describe the results of our experiment that aim to answer whether there is a difference between overall and targeted sentiments in tweets. Table 4 reports the individual $c(s_o)$, $c(s_t)$ and joint $c(s_t, s_o)$ distributions of the overall s_o and targeted s_t sentiment spaces in the three collected datasets. The inner portions of the table for each dataset report the joint distribution $c(s_o, s_t)$ for the sentiment classes between the two tasks, i.e. to what extent the sentiment changes on a per class basis. The final rows and columns for each dataset report the individual distributions $c(s_o)$ and $c(s_t)$ of the sentiment classes of the two tasks, i.e. how often each sentiment class occurs in the dataset.

When considering the differences between overall and targeted individual sentiment distributions for each individual dataset (i.e. $c(s_o)$ and $c(s_t)$), we see important differences between overall and targeted sentiments within the tweets. For example, Table 4 shows that the overall sentiments are markedly more negative than the targeted ones (192.2 vs. 106.3) for the Aurora Shooting event. This already indicates a mismatch between the presence of overall sentiment in tweets and targeted sentiments towards the subjects in those tweets. That same pattern repeated in the other two datasets, although the magnitude of the difference between the overall and targeted sentiments is smaller. This result answers our research question, i.e. there is indeed a difference between overall and targeted sentiment within various types of crisis events.

However, it is also important to investigate where precisely sentiment tends to differ between the overall and targeted scenarios. To do so, we next examine the joint distribution scores $c(s_t, s_o)$ for the individual class pairs. First, when comparing the proportion of tweets that remain in the same sentiment class in both labelling tasks (the values in bold of Table 4), we see that these numbers are small for the negative and positive sentiments with respect to the total number of overall and targeted negative and positive sentiments.

[3]Only unique tweet identifier's are provided due to Twitter's ToS, however, tweet texts can be recovered for named ids using publicly available tools (McCreadie et al. 2012)

		Aurora Shooting				Hurricane Isaac				Ebro Flood			
		s_o				s_o				s_o			
$c(s_t, s_o)$		**neg**	**neu**	**pos**	$c(s_t)$	**neg**	**neu**	**pos**	$c(s_t)$	**neg**	**neu**	**pos**	$c(s_t)$
s_t	**neg**	**42.9**	61.3	2.2	106.3	**115.9**	156.6	3.9	276.3	**222.3**	196.4	7.0	425.7
	neu	141.4	**1,557.0**	89.0	1,787.3	176.5	**1,493.7**	44.0	1,714.2	219.5	**1,312.8**	44.0	1,576.3
	pos	7.9	78.7	**203.7**	290.3	7.1	65.0	**22.4**	94.5	6.3	40.4	**40.2**	87.0
	$c(s_o)$	192.2	1,696.9	294.9	2,184.0	299.5	1,715.3	70.3	2,085.0	447.1	1,549.7	91.3	2,089.0

Table 4: Individual $c(s_t), c(s_o)$ and joint $c(s_t, s_o)$ distributions of sentiments.

For instance, the negative-negative pair (tweets that were labelled as containing negative targeted and overall sentiment) in the Aurora Shooting dataset receives a score of 42.9. Contrast this score to the total overall and targeted negative sentiment scores for the event (192.2 and 106.3 respectively). We observe the same pattern when considering the positive class as well – the positive-positive pair received a score of 203.7, while the overall and targeted total scores are 294.9 and 290.3, respectively, which indicates that a large number of tweets were labelled differently under the overall and targeted sentiment labelling scenarios. Furthermore, we observe a recurrent pattern between tweets being labelled as having an overall neutral sentiment but also being labelled as having a targeted positive/negative sentiment. For instance, for the Aurora Shooting event, we observe that the neutral-negative pair has a score of 61.3, while the neutral-positive pair has a score of 78.7. Finally, we see that the crossover between negative and positive classes is rare, i.e. the scores for the positive-negative and negative-positive pairs are low. Indeed, we see the same pattern across all three of the datasets. These observations reveal that, frequently, tweets expressing a polarised sentiment do not target all the subjects in it and, on the contrary, tweets written in a somewhat neutral language may actually contain a negative or positive sentiment towards a particular subject.

5 Conclusions

In this paper, we have analysed the differences between the overall and targeted sentiment analysis of social media posts related to three crises events. Through an experiment over three tweet datasets pertaining to different crisis events, we show marked and relevant differences between sentiment labels when considering the overall and targeted sentiments as obtained via crowdsourcing, indicating that these are distinct tasks. These differences highlight the need for a deeper sentiment analysis in social media posts in order to obtain meaningful and valuable insights about public opinion related to disasters or other types of critical events.

6 Acknowledgements

This work has been carried out in the scope of the EC co-funded SUPER (FP7-606853) project.

References

[Agarwal et al. 2011] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis on Twitter data. In Proc. of LSM, 2011.

[Barbosa and Feng 2010] L. Barbosa and J. Feng Robust sentiment detection on Twitter from biased and noisy data. In Proc. of COLING, 2010.

[Brynielsson et al. 2014] J. Brynielsson, F. Johansson, C. Jonsson, and A. Westling. Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. *Security Informatics*, 3(1), 2014.

[He et al. 2008] B. He, C. Macdonald, J. He, and I. Ounis. An effective statistical approach to blog post opinion retrieval. In Proc. of CIKM, 2008.

[Imran et al. 2014] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *CoRR*, abs/1407.7071, 2014.

[Jiang et al. 2011] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent Twitter sentiment classification. In Proc. of HLT, 2011.

[Marcheggiani et al. 2014] D. Marcheggiani, O. Täckström, A. Esuli, and F. Sebastiani. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In Proc. of ECIR, 2014.

[McCreadie et al. 2012] R. McCreadie, I. Soboroff, J. Lin, C. Macdonald, I. Ounis and D. McCullough. On Building a Reusable Twitter Corpus. In Proc. of SIGIR, 2012.

[Moilanen and Pulman 2009] K. Moilanen and S. Pulman. Multi-entity sentiment scoring. In Proc. of RANLP, 2009.

[Pang and Lee 2008] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 2008.

[Sakaki et al. 2010] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proc. of WWW, 2010.

[Schulz et al. 2013] A. Schulz, T. D. Thanh, H. Paulheim, and I. Schweizer. A fine-grained sentiment analysis approach for detecting crisis related microposts. In Proc. of ISCRAM, 2013.

[Santos et al. 2012] R. L.T. Santos, C. Macdonald, R. McCreadie, I. Ounis and I. Soboroff. Information Retrieval on the Blogosphere. Foundations and Trends in Information Retrieval Journal, 6(1), 2012.

[Verma et al. 2011] S. Verma, S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson. Natural language processing to the rescue? Extracting "situational awarenes" tweets during mass emergency. In Proc. of ICWSM, 2011.