Information Systems Frontiers : Authors' Pre-Print
http://dx.doi.org/10.1007/s10796-018-9827-x

Regional Sentiment Bias in Social Media Reporting During Crises

Karin Sim Smith · Richard McCreadie · Craig Macdonald · Iadh Ounis

Accepted: 18/01/2018

Acknowledgements This work was supported by the EC co-funded SUPER (FP7-606853) project.

Author Note: This article is an expansion of a previously published paper entitled 'Analyzing Disproportionate Reaction via Comparative Multilingual Targeted Sentiment in Twitter' at the international conference on Advances in Social Network Analysis and Mining (ASONAM 2017).

Abstract Crisis events such as terrorist attacks are extensively commented upon on social media platforms such as Twitter. For this reason, social media content posted during emergency events is increasingly being used by news media and in social studies to characterize the public's reaction to those events. This is typically achieved by having journalists select 'representative' tweets to show, or a classifier trained on prior human-annotated tweets is used to provide a sentiment/emotion breakdown for the event. However, social media users, journalists and annotators do not exist in isolation, they each have their own context and world view. In this paper, we ask the question, 'to what extent do local and international biases affect the sentiments expressed on social media and the way that social media content is interpreted by annotators'. In particular, we perform a multi-lingual study spanning two events and three languages. We show that there are marked disparities between the emotions expressed by users in different languages for an event. For instance, during the 2016 Paris attack, there was 16% more negative comments written in the English than written in French, even though the event originated on French soil. Furthermore, we observed that sentiment biases also affect annotators from those regions, which can negatively impact the accuracy of social media labelling efforts. This highlights the need to consider the sentiment biases

School of Computing Science
University of Glasgow
Glasgow, G12 8QQ
E-mail: karinsim@gmail.com OR {firstname.lastname}@glasgow.ac.uk

of users in different countries, both when analysing events through the lens of social media, but also when using social media as a data source, and for training automatic classification models.

1 Introduction

When significant events occur, social media is often used as an outlet for people in different parts of the world to express their opinions, sentiments, as well as comment on those events. For this reason, social media is potentially a valuable resource to help understand how events are being perceived by different social groups (De Choudhury et al, 2012; Stieglitz and Dang-Xuan, 2013). To achieve this, a significant amount of time, effort and resources have been used to support the development of datasets and automatic systems for analysing social media (Tang et al, 2016b). Furthermore, there are a number of commercial companies, such as IBM Watson's Alchemy API[1], Brandwatch[2] and Crimson Hexagon [3], which currently sell automatic social media analytics as a service.

Sentiment analysis is a common functionality provided by social media analytics platforms. In particular, this involves the categorisation of content into different sentiment categories. Two class (positive or negative) (Agarwal et al, 2011; Jiang et al, 2011) and three class (positive, negative or neutral) (Ounis et al, 2008; Vargas et al, 2016) sentiment classification are the most common type of sentiment analysis approaches deployed in practice. Some works also examine more granular emotional categories, such as anger or despair (Purver and Battersby, 2012). Sentiment analysis is a functionality valued by companies as a means to track how their brand is being perceived by the general public. Furthermore, news media and social scientists use sentiment analysis to characterize the public's reaction to those events. For instance, sentiment statistics for an event can be used to gauge whether particular discussion topics are gaining traction during events such as elections (Wang et al, 2012).

However, the application of sentiment analysis over events with larger scope introduces new complexities. In particular, (social media) users will participate with the discussion from different geographical regions and in different languages. Moreover, these users come from very different social and cultural backgrounds, with their own biases. To-date, there have been a wide range of works that examine the practical aspects of building sentiment classifiers for events that span multiple languages (Narr et al, 2011; Tromp, 2012), such as which classifiers to use. On the other hand, no prior works have examined national vs. international biases on social media and their impact on sentiment analysis of events.

In contrast, in this paper, we present a multi-lingual study of two terrorist attacks based on analysis of social media content posted in three languages,

[1] `https://www.ibm.com/watson/alchemy-api.html`

[2] `https://www.brandwatch.com/`

[3] `https://www.crimsonhexagon.com/`

with the aim of determining to what extent national and/or international biases result in significant variation in sentiment expressed in different geographical regions, and the implications of this variation when developing sentiment analysis services. In particular, we study both the terrorist attack which took place in different parts of Paris on the 20th of November 2015, and the terrorist attack on a Christmas market in Berlin on the 19th December 2016. By doing so, we aim to answer two research questions:

- **RQ1:** Are there significant differences between the sentiments expressed on social media by users from different geographical regions?
- **RQ2:** Are there significant differences in sentiments identified by annotators based in different geographical regions?

The contributions of this work are two-fold. First, we show that a multilingual comparison of tweets allows for a more informative analysis of wider global opinion for a major event than a classical monolingual analysis. Indeed, our results highlight how external reactions to a disaster can be significantly more negative than local reactions. This has notable implications for developers of current sentiment analysis systems trained on mono-lingual datasets (labelled by users at a national level), as it highlights how such systems may provide a misleading view of public reactions to an event. Second, we examine how annotator bias can affect the analysis of sentiment during an event, showing that regional bias also affects crowdsourced tweet labelling, in addition to negatively impacting annotator agreement between workers in different regions. Such bias is an important factor to consider when using geographically-dispersed workers to label social media data.

In the next section, we survey related work (Section 2) before describing the dataset we use (Section 3) and how we label that dataset (Section 4). We then examine the Twitter user bias (Section 5) and the annotator bias (Section 6), as well as discuss implications for building automatic classifiers (Section 7). We summarize our conclusions in Section 8.

2 Related Work

When following real world events, people often comment and express their opinions via social media. Sentiment analysis on the various forms of social media can help political scientists, politicians or those in civil society understand how these events are perceived by the general public. In this section we summarize related works in the fields of sentiment analysis on Twitter in particular (as this is the social media platform we use in our study), as well as sub-domains of sentiment analysis that are relevant to our study, such as multi-lingual sentiment analysis and target-dependent sentiment analysis.

Challenges of sentiment analysis in Twitter: Sentiment analysis on Twitter is particularly challenging, due to abbreviations, the terseness of the tweets, lack of context and the ambiguity which results. This has already been documented (Maynard and Bontcheva, 2016; Maynard and Hare, 2015), with others

developing specially adapted NLP tools for the domain (Bontcheva et al, 2013). In addition, there is the difficulty inherent in detecting the sarcasm present in many tweets, as investigated by Maynard and Greenwood (2014). Sentiment analysis therefore has had to adapt to the Twitter genre. Tang et al (2014) developed a deep-learning system for Twitter sentiment classification, using sentiment word embedding features in addition to hand-crafted features such as emoticons, sentiment lexicons, negation, among others. In their work on sentiment-specific word embeddings, Tang et al (2016b) experimented with taking the sentiment context of the words into account, in contrast to classical approaches that use simple word features. They found that this impacts performance in Twitter sentiment classification, and posted significant improvements as a result, outperforming other embedding-based approaches. Rather than focusing on how to develop more effective algorithms for sentiment analysis like the above works, in this paper we examine how biases in sentiment training datasets can occur and their effect on learned sentiment analysis systems.

Crisis-related: A range of prior works have examined sentiment expressed during and after important events. For instance, Thelwall et al (2011) found that negative sentiment generally exceeds positive sentiment, including for cases examining positive events. In a similar setting to ours, i.e. post-crisis, although for purposeful dissemination of information, Kwak et al (2010) performed an extensive quantitative study on Twitter and information diffusion through it. Others have examined Twitter (on a monolingual level) as a source for organising and disseminating information (Hermida, 2010), as well as to enhance awareness during crises (for alerting and disseminating information). Schulz et al (2013) and Verma et al (2011) also investigated Twitter as a means of information dissemination during crises. Schulz et al (2013) proposed a fine grained sentiment analysis, experimenting with seven classes to better capture range of emotions expressed during realtime information crisis management. Classifier performance was generally quite low (excepting a dataset that consisted of only 114 tweets). Again in a similar domain to ours, Nagy and Stamberger (2012) performed crowd sentiment detection during crises. They showed that sentiment changed over time (varied over 0-12 and 12-24 hours), and that information-based tweets unsurprisingly had more nouns than adjectives. This analysis was crowdsourced, but not targeted or multilingual. We focus on Twitter post-crisis scenarios, and while much of this previous work on sentiment analysis in Twitter has been monolingual, we take a compararive multilingual approach. Our analysis is on targeted sentiment, investigating sentiment towards particular entities in a post-crisis scenario.

Multilingual sentiment analysis: There has also been work on multilingual sentiment analysis in social media, including Twitter (Narr et al, 2011; Tromp, 2012). In the case of the Narr et al (2011), they did so superficially, in a language-independent manner, for a short study using emoticons. Whereas Tromp (2012) examined three different types of social media data in an in-depth multilingual sentiment analysis study, covering six languages. That sys-

tem included components for language identification, uses POS tags for subjectivity detection as well as polarity detection, and works on extracted pattern-based rules. His work did not consider targeted sentiment, nor the aftermath of crisis events.

Balahur and Turchi (2012) explored using machine translation on multilingual data before performing sentiment analysis. They claim that the technology is mature enough, although the scores reported are low, with the classifier assigning all to majority class in the case of all 3 tested machine translation systems for German and French. For Spanish, only the Google system scores above 50%. It would seem that machine translation at that time was insufficient for the task. Meanwhile, in their multilingual study, Mozetič et al (2016) compared human labelling and classification models, interestingly hypothesizing that 'the inter-annotator agreement approximates an upper bound for a classifier performance', indicating that agreement is a good estimate of task difficulty.

Target-dependent sentiment analysis: In terms of target-dependent approaches, Dong et al (2014) used dependency parsing to establish the target of the sentiment for their target-dependent twitter sentiment classification. They used the features in SVM and RNN models, and found an Adaptive Recursive Neural Network (AdaRNN) outperformed the others. Also on target-dependent sentiment classification, Tang et al (2016a) developed a target-dependent long short-term memory (LSTM) which takes the sentiment polarity of the target word context into account. Testing on a dataset for target-dependent twitter sentiment classification, they show that by incorporating context words for the target into an LSTM they improve over all baselines. The challenges of entity-based opinion mining, analysing the sentiment of a tweet towards a particular entity contained within it, has already been studied by Maynard and Hare (2015). In addition to addressing the extraction of opinions on crucial events in society for the purposes of archiving, they also expanded their research to cover the integration of multimedia through the extraction of sentiment evidence from images. They developed a framework of sub-components, covering hashtag decomposition, negation, identifying factual versus opinionated statements, sarcasm and irony. Outside the Twitter domain, Marcheggiani et al (2014) explored aspect-oriented opinion mining, predicting sentence level sentiment towards a particular aspect of an entity, using conditional random fields. They used the CRFs to jointly model overall and aspect-based sentiment, and found that this lead to a slight improvement in scores. Again outside the Twitter domain, Moilanen and Pulman (2009) explored deeper multi-entity sentiment analysis, looking at all base nouns in a text and more linguistically sophisticated patterns via compositional sentiment parsing. In the Twitter domain, Jiang et al (2011) have also investigated targeted sentiment, similarly looking at a particular aspect of the tweet, not simply the overall sentiment. Meanwhile, Vargas et al (2016) established that there is a difference between the overall sentiment of a tweet, and sentiment expressed towards a particular target in the tweet. These studies have been

in different monolingual settings, the latter including multilingual tweets, but not as comparative multilingual analysis. In a deeper monolingual analysis on the public response in Twitter following this same attack in Paris, Magdy et al (2016) predicted stance, particularly towards Muslims, based on user profile. They used retweets and 'likes' as a benchmark in researching emotional reaction (Magdy et al, 2015) following the attack, which is similar to our work in terms of context, but in our case we are interested in the textual content.

Most recently, SemEval 2017 included a specific task on target oriented Sentiment Analysis in Twitter[4], comparing targeted sentiment in English and Arabic over a range of topics from people to products. In contrast, the basis of our work is post-crisis sentiment analysis in Twitter, adopting a multilingual approach, which is comparative and targeted in nature. It also encapsulates two different, but similar events in three languages (English, French and German) and focuses on how sentiment biases differ between national and international observers of a crisis event.

Task: As discussed above, there have been a wide range of sentiment analysis approaches examined in the literature. In this paper, we analyse how local vs. international bias can affect sentiment expressed about an event on Twitter. More precisely, for a tweet post p that is part of a larger discussion about a sensitive event e that also mentions a particular entity of interest (target) t, we analyse how sentiment identified differs when post p is sentiment labelled ($s \in \{negative, positive, neutral\}$) by local and international crowd workers.

3 Dataset Construction

To examine regional bias in social media, we require one or more suitable datasets. There have been a wide variety of studies into sentiment on social media platforms such as Twitter in previous works. However, most of these datasets are not publicly available, due to limits in the terms of service of the social media platforms themselves. Twitter is an exception, in that datsets can be shared in the form of tweet identifier lists, however, those datasets require retrospective crawling via API, which can be problematic as deleted tweets or tweets by deleted accounts are rendered non-accessible. Hence, for this work we crawl our own datasets for this study, as described below.

Dataset: To investigate the sentiment biases of social media users in different geographical regions, we use two different datasets containing social media posts collected during two terrorist attacks. We choose these events as they are large enough to attract international discussion. Furthermore, by choosing to focus on terrorist attacks, we can contrast local reactions to international reactions. More precisely, our first dataset consists of Twitter tweets posted during the Paris attack on 20th-23rd Nov 2015. Tweets were collected using the Twitter Streaming API using the hashtag '#Paris' as a filter. The second dataset

[4] http://alt.qcri.org/semeval2017/task4/

we use contains tweets collected during and after the attack on a Christmas market in Berlin on 19th of December 2016, and containing '#Berlin'. Both crawls contain tweets in a wide variety of languages.

Language Selection: Our aim is to analyse regional biases through the lens of social media. However, most social media content posted lacks any geographical identifier. For instance, research by Dredze et al (????) indicates that only around 1% of tweets have a place identifier associated to them. Hence, as an alternative, we use the language of the tweets as a proxy for location. In particular, for the Paris Attack dataset, we use tweets posted in French to represent users posting from France and posts in English to represent users from the US, UK and Canada. Meanwhile, for the Berlin Attack dataset, we use German tweets to represent local users from Germany and English tweets to similarly represent users from the US, UK and Canada. While using a tweet's language for geolocation is not exact, it should be sufficient for the purposes of contrasting local vs. international reactions to the events.

Language Filtering: We filter on the language using the 'lang' tag of each tweet, which identifies the language via Twitter's own language classifier.[5] According to this classifier, for the Paris Attacks dataset, the most common language was English (1,232,100 tweets) followed by French (402,914 tweets). Meanwhile, for the Berlin Attacks dataset, the most common language for that tag in Twitter at the specified time was English (232,469 tweets), followed by French (152,820 tweets), then German (136,012 tweets).

Sentiment Targets: Manually analysing millions of tweets is not feasible due to time/cost constraints. Hence, inspired by previous works that examine targeted sentiment (Vargas et al, 2016), we choose a small number of entities (targets) of interest to analyse in detail. In particular, we select targets that were central discussion topics during each event. In particular, for the Paris Attacks dataset, we select French President **François Hollande**, **Europe** and **Muslims** as our targets. These targets are chosen as they occur frequently and are likely to have emotion expressed about them. We filter the Paris Attacks dataset to only include posts that mention these targets using separate [keywords] for each: François Hollande:[hollande]; European Union:[europe]; and Muslims:[muslim OR musulman]. We then divide this filtered set into six subsets based on the target and language: Hollande/English; Hollande/French; Europe/English; Europe/French; Muslim/English; Muslim/French. Similarly, for the Berlin Attacks dataset, we choose German Chancellor **Angela Merkel**, **Muslims** and **Police** as our targets. We chose these targets as the German Chancellor Angela Merkel has been criticised by some as having been overly open to refugees, the Muslim community in some areas experienced an irrational backlash after the attack and the police role was a subject of discussion after the event. We filter the Berlin Attacks dataset to only include posts that mention these targets, again using separate [keywords] for each: Angela Merkel:[merkel]; Police:[police]; and Muslims:[muslim OR Muslime OR

[5] Rather than the user's self-defined language, which is less accurate.

Moslem OR Muselman]. We then divide this filtered set into six subsets based on the target and language: Merkel/English; Merkel/German; Police/English; Police/German; Muslim/English; Muslim/German.

Sampling: Furthermore, to provide a detailed analysis, it is desirable to have a diverse set of tweets to examine, both in terms of textual content and in terms of time (i.e. when during the event each post was made). As such, we apply the following sampling strategy to the six tweet sets to create a diverse tweet sample for each. First, we divide the tweets from each set into hour batches based on their publication timestamps and index each hour using the Terrier open source IR platform (Ounis et al, 2006). Per hour, we rank the tweets using the keywords for the associated target as the query. Inspired by Kraaij and Spitters (2003), we use a Gaussian function configured to promote sentences that are of approximately the length of a normal English sentence (in words[6]) for ranking. We select the top 100 tweets from each hour to create the dataset sample. We then remove near-duplicate tweets from each dataset sample by applying a cosine similarity threshold τ over that sample in a greedy time-ordered manner (τ= 0.7). A summary of the statistics of the dataset samples produced are provided in Table 1.

4 Sentiment Labelling:

Now that we have tweet sets in different languages for each of the two events, we next need to determine the sentiments expressed within each of these sets. To do so, we have human assessors manually label sentiments expressed. As we have tweets in three languages (English, French and German), we use the medium of crowdsourcing to obtain annotators who understand each language. It is of note that in practice we are measuring sentiment perceived by a third party for each tweet. For this reason, it is expected that in some cases workers may not be able to distinguish 'true' sentiment as meant by the original author, e.g. because of the use of localized language/slang. Indeed, sentiment labelling of events is generally seen as a 'difficult' labelling task, with expected Fleiss Kappa inter-worker agreement of only around 0.3 (fair agreement) (Hsueh et al, 2009). We describe the configuration of our crowdsoucing jobs below.

Labelling Task: To analyse how sentiment varies across tweets in different languages, we need to generate sentiment labels for the tweets in our six samples. To achieve this, we had crowdsourced workers manually annotate the tweets, using the Crowdflower platform.[7] Following earlier work on sentiment labelling (Kiritchenko and Mohammad, 2016) that indicated labelling accuracy does not significantly increase beyond 2-3 workers, each tweet-target pair is given to three different workers. Each worker is asked to label the sentiment (negative, positive or neutral) expressed by the author of the tweet towards

[6] Mean/expectation was set to 25 and the standard deviation was set to 20.

[7] `http://www.crowdflower.com`

Dataset/Event	Target	Language	# Sampled Tweets	#QA Tweets
Paris	Hollande	English	725	48
	Europe		800	
	Muslim		496	
	Hollande	French	718	45
	Europe		778	
	Muslim		513	
Berlin	Merkel	English	1011	66
	Police		1089	
	Muslim		897	
	Merkel	German	838	68
	Police		1009	
	Muslim		238	

Table 1: Summary of tweet samples labelled by crowdsourced workers from the Paris Attacks (Paris) and Berlin Attacks (Berlin) datasets.

the subject given. For the six English tweet samples (three for the Paris Attack and three for Berlin Attack), only English-speaking users were allowed to participate in labelling those samples. Similarly, only French-speaking users could label the three French tweet samples for the Paris Attack and only German-speaking users could label the three German tweet samples for the Berlin Attack.

Worker Instructions: Workers were provided task completion instructions and clarifications before accepting the job. An example of the instructions provided to the workers for the Berlin Attacks English samples is shown below. The instructions were translated by an expert into the other target languages (French or German) when submitting the labelling tasks for non-English samples. An example of the labelling interface is shown in Figure 1.[8]

```
Overview

Welcome to our tweet sentiment classification task. You will be given a tweet related
to the  terrorist attack in Berlin in December 2016 and a subject (some person or
institution related to the event). The task consists of labelling the sentiment that
the AUTHOR OF THE TWEET expresses towards THE GIVEN SUBJECT as negative, neutral or
positive.

Additional clarification for labelling:
> 'Negative': the tweet constitutes a negative feeling towards, or criticism of the
  subject (eg Muslims or Merkel), perhaps blaming them for the terrorist attacks.

> 'Positive': the tweet is sympathetic to the given subject (eg Muslims or Merkel)
  and expresses a positive sentiment towards them, perhaps commending them for
  condemning the attacks.

> 'Neutral': the tweet constitutes a statement of fact and not an opinion on the subject
  (e.g. Muslims or Merkel). It may be reporting facts of the event, but not making a
  judgement.

If the tweet is not in English, is not readable or does not really mention the given
subject, please label it as neutral.
```

[8] Minor changes were made to the instructions for labelling the Berlin dataset to clarify a small number of situations that arose when labelling the Paris dataset.

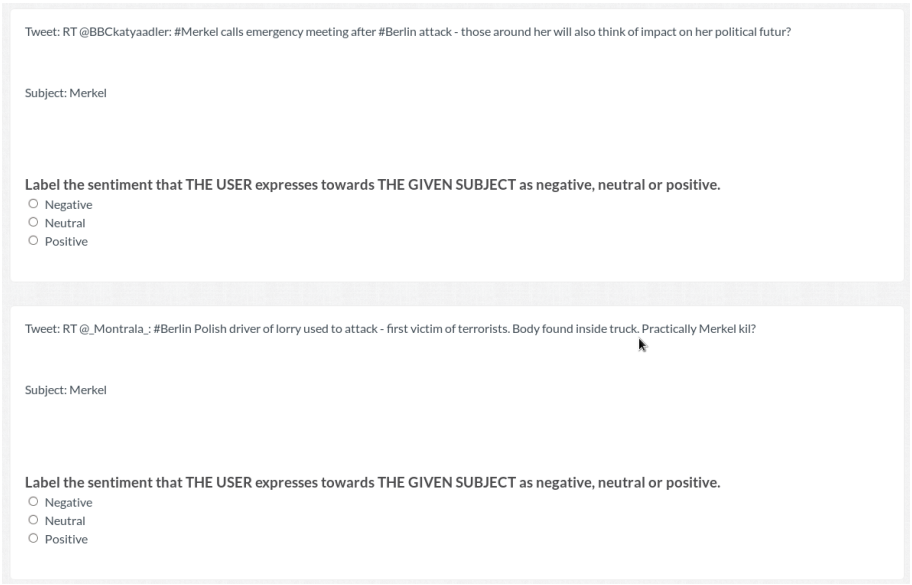

Fig. 1: An example of the Crowdflower sentiment labelling interface.

Quality Assurance: Following best practices in crowdsourcing (McCreadie et al, 2013), we apply a series of quality assurance techniques to avoid poor-quality work. First, to avoid a few workers dominating the labelling process, the number of tweets a single worker could label was limited to 200. Furthermore, to increase accuracy, worker quality was dynamically assessed against a gold standard set of tweets that were previously annotated by the authors (see the '#QA Tweets' column in Table 1). For the Paris Attacks dataset this was comprised of 45 (French) and 48 (English) tweets. For the Berlin Attacks this consisted of 68 (German) or 66 (English) tweets. We disregarded the tweets from workers whose accuracy dropped below 70%. To produce a single label for each tweet, we take the majority vote across the three labels produced. We discard any tweets where there was not majority agreement. The statistics of the six tweet samples after labelling and discarding are provided in Table 2.

Worker Agreement: Table 2 also reports the number of tweets for which there was a majority vote across the three workers that labelled each tweet, as well as the agreement level in terms of Fleiss Kappa for each event, target and language. First, from Table 2 we see that for a small number of tweets 3% and 5% no majority vote could be obtained (all three workers selected different labels). For the Berlin dataset, there was no majority among the annotators for 91 (3%) of the labels for the English tweets and 110 (5%) of the German ones. For the Paris dataset there was no majority for 102 (5%) of the French labels, and no majority for 98 (5%) of the English ones. The labels for these tweets were disregarded, as no majority vote could be reached.

Source	Tweet Sample	# tweets	# no majority	# unanimity	Fleiss' Kappa
Paris	All / French	1998	102	907	0.19
Paris	Hollande / French	769	51	263	0.21
Paris	Europe / French	782	15	454	0.17
Paris	Muslim / French	549	36	190	0.10
Paris	All / English	1997	98	695	0.18
Paris	Hollande / English	711	37	299	0.21
Paris	Europe / English	800	14	251	0.10
Paris	Muslim / English	487	46	147	0.13
Berlin	All / German	2085	110	1013	0.39
Berlin	Merkel / German	838	42	451	0.24
Berlin	Police / German	1009	52	432	0.24
Berlin	Muslim / German	238	16	130	0.44
Berlin	All / English	2997	91	1631	0.41
Berlin	Merkel / English	1011	36	433	0.25
Berlin	Police / English	1089	6	760	0.25
Berlin	Muslim / English	897	49	396	0.21

Table 2: Fleiss' Kappa scores from both Paris and Berlin experiment.

On the other hand, from Table 2 we also observe an unexpectedly large amount of discrepancy among the annotators in terms of agreement for some of the targets in comparison to the event datasets as a whole. When we consider the annotator agreement in terms of Fleiss' Kappa scores, for the German data (all targets combined) the Fleiss' kappa is 0.39. For the English data (all targets combined) Fleiss' Kappa is 0.41. The scores are similar, constitute 'fair agreement' and are generally in-line with prior works on sentiment analysis labelling (Hsueh et al, 2009; Vargas et al, 2016). However, the agreement over the individual targets is markedly lower. For example, while the German section of the Berlin dataset has a Fleiss' Kappa score of 0.39, the score for the 'Police' target is only 0.24. To explain this, it is important to understand that Fleiss' Kappa scores are affected by the majority class proportions. To illustrate, the proportion of labels subject to unanimous agreement is 49% for whole set, but 54% for the 'Merkel' target, i.e. higher unanimity for the latter. The difference here is the fact that the prevalent group for the 'Merkel' subset is negative (see later in Table 3), so the probability of random agreement is much higher. Meanwhile, for the set as a whole, the breakdown is more evenly spread, and therefore is less affected by the probability of random agreement. Another example of this effect is the Muslim subsection of the Paris dataset, with 0.13 agreement for the English and just 0.10 agreement for the French. For English there was unanimity for 147 of the labels, so just 30%. For the French tweets, there was unanimous agreement for 190 cases, i.e. 37%. Again here, the French score is affected by the majority class, which was 75.4% neutral, meaning that random probability for that category was higher. For this reason, we should generally avoid drawing conclusions from low per-target Fleiss' Kappa scores.

Overall, our settings are very similar to those chosen by Vargas et al (2016) in their crowdsourcing experiment, who found that 'these results indicate that the described crowdsourcing configuration produces good quality labels'. When considering event datasets as a whole, we obtained similar lev-

els of agreement and so conclude that the labels produced are usable for our subsequent analysis, although as others have previously observed (Moilanen and Pulman, 2009) sentiment labelling is a difficult and subjective task.

Reproducibility: The tweet samples described above, as well as the associated crowdsourced labels used for evaluation are available for download at:

– `http://dx.doi.org/10.5525/gla.researchdata.584`

In the following experiments we investigate our two research questions, each in a separate section:

– **RQ1:** Are there significant differences between the sentiments expressed on social media by users from different geographical regions? (Section 5)
– **RQ2:** Are there significant differences in sentiments identified by annotators based in different geographical regions? (Section 6)

5 RQ1: Twitter User Bias

Having labelled sentiments expressed by people in different regions (represented by languages) towards different targets for each of two events, we first examine whether there exists any observable regional bias in the collected tweet samples. To do so, we compare the sentiment distributions per target for each event across the two language pairs (representing local and international discussion). If there is no regional bias, then we would expect that the relative proportion of tweets belonging to each sentiment class (neutral, negative and positive) would be similar when comparing the samples for each target in the two languages. For instance, for the Paris event, we would expect that the proportion of negative sentiments would be roughly equivalent between the Hollande/English and Hollande/French samples.

The final three columns of Table 3 report the number and proportion of tweets from each of the tweet samples that were labelled as either neutral, negative or positive, across the two events. From Table 3, we make two main observations. First, we see that sentiment about the different targets tends to be polarised, i.e. the sentiments expressed about a target tend to be dominated by a single sentiment class. For instance, for the Paris event, for both languages, the Hollande and Europe targets are dominated by the neutral class, while the Merkel and Muslim targets are dominated by the negative class. This is to be expected, as discussion about a particular target tends to be focused on a single issue, such as immigration policy in the case of the Merkel target within the Berlin dataset. This results in the predominant sentiment about that issue biasing the discussion toward that sentiment. However, the second observation we can make from Table 3 is that the sentiment proportion expressed by users in different languages differ markedly. For example, for the Paris attacks, the English tweets analysed about the Holland target were less positive in their judgement of him, as indicated by the lower positive score (8%). However, what is particularly striking is the discrepancy between

Source	Tweet Sample	tweets	neutral	negative	positive
Paris	Hollande / French	718	465(64.8%)	169 (23.5%)	84(11.7%)
Paris	Europe / French	778	680 (87%)	70 (9%)	28 (4%)
Paris	Muslim / French	513	387(75.4%)	73 (14.2%)	53(10.3%)
Paris	Hollande / English	725	504 (70%)	163 (22%)	58 (8%)
Paris	Europe / English	800	520 (65%)	257(32%)	23 (3%)
Paris	Muslim / English	496	186 (37%)	273(55%)	38(8%)
Berlin	Merkel / German	838	153(18.3%)	650(77.5%)	35(4.2%)
Berlin	Police / German	1009	739(73%)	169(17%)	101(10%)
Berlin	Muslim / German	238	29 (12.2%)	124(52.1%)	85(35.7%)
Berlin	Merkel / English	1011	290(29%)	687 (68%)	34(3%)
Berlin	Police / English	1089	966(88.7%)	115 (10.5%)	8 (7.3%)
Berlin	Muslim / English	897	223 (25%)	624 (69.5%)	50 (5.5%)

Table 3: Results for Multilingual Targeted Sentiment Labelling on Twitter samples for '#Paris' between the 20th to the 23rd of November 2015. (excluding where no majority agreement)

the amount of tweets labelled negative by the English speaking annotators for targets 'Europe' and 'Muslim', compared to the French counterparts. For instance, the French annotators labelled 15% of the tweets with target 'musulman' ('muslim') as negative, compared to 55% of the English annotators. The results for target 'Europe' show a similar trend, with 9% of tweets labelled as negative by French annotators, while 32% were labelled as negative by English annotators.

Moreover, if we compare these observations from the Paris attacks dataset to the Berlin attacks dataset, we observe similar trends. In particular, for the target 'Muslim' there is again a larger amount of negative sentiment expressed by English tweets than in German tweets, as was the case for the Paris dataset. It would seem that target 'Muslim' provokes a disproportionate reaction among those expressing themselves via English tweets in this instance too, presumably now reflecting a wider societal trend. This reaction is also apparent when we consider that the same filtering and sampling process yielded 897 tweets with target 'Muslim' in the English dataset (after discarding those tweets where there was a lack of annotator agreement), yet only 238 for the German dataset. Hence, there was far less commenting on this target in the German dataset in the first place. The amount of positive tweets for the target 'Muslim' in the German subset is also noteworthy (37.5%), when compared to the English subset (6%). This indicates a great deal more positivity among German tweets as compared to English ones. Indeed, investigating this positivity in more detail we discover that for the English subset of the Berlin attacks dataset: 5 (out of a total number of 50 positive tweets) are from *within* the Muslim community, i.e. 10%. These first person comments are of the format "I" or "We". Meanwhile, for the German subset, 15 (out of a total number of 85 positive tweets) are from within the Muslim community, i.e. 18%. While not a huge discrepancy, there are more positive tweets coming from within the Muslim community for the German dataset. This is perhaps not surprising, as they are directly

Language	Subject	Text
EN	1st Person	*"RT @name: I am a Muslim and being Muslim I condemn the tragic incident in #Berlin"*
EN	3rd Person	*RT @name: Reminder: The attack in #Berlin is absolutely NOTHING to do with #Islam or #Muslims. Muslims aren't terrorists. #Terrori?*
DE	1st Person	*Ich bin ein Muslime und verurteile den Anschlag. #Berlin #Breitscheidplatz* [Translated as: *I am a Muslim and I condemn the attack*]
DE	1st Person Plural	*RT @name: Wir Ahmadi Muslime verurteilen den Anschlag in #Berlin. Unser Mitgefühl ist mit den Hinterbliebenen der Todesopfer. #MuslimeG?* [Translated as *RT @name: We Ahmadi Muslims condemn the attack in Berlin. Our sympathy is with the victims' loved ones*]

Table 4: Examples of positive Tweets for target *Muslim* from '#Berlin' dataset between the 19th to the 22nd of December 2016.

affected, potentially feeling under attack as illustrated by the example tweets in Table 4.

In summary, to answer **RQ1**, there are marked differences in the sentiments expressed by Twitter users posting in different languages, and hence in different geographical regions (local vs. international in this case). Indeed, we observe this behaviour across both the Paris attacks and Berlin attacks datasets. This result is unexpected, since those in Paris (and France more generally)[9] are the ones more directly affected by the attack. Indeed, if we consider the French reaction to be a reasonable baseline reaction to the terrorist attack, then by contrast it makes the English (predominantly USA, UK and Canadian) response disproportionately negative. We also see a similar picture when examining the Berlin attacks, where for the Muslim target, there are both more negative sentiments and fewer positive sentiments expressed. From a broader information systems development perspective, this finding is notable as it raises the question of whether monolingual sentiment analysis systems trained on user data inherit local or international biases, which may be undesirable (we examine this question more closely later in Section 7).

6 Annotator bias

In the previous section, we showed that there was a large difference between the proportion of English and French tweets that were labelled as positive and negative by crowd workers. However, the workers themselves come from particular geographical regions. Hence, an interesting question is whether the crowd workers are also a source of bias. To examine this, we first manually analyse a small subset of tweets. From this analysis, we observe a pattern, where tweets were wrongly labelled as negative for one of the targets. For instance, the following tweet was labelled negative for the 'Muslim' target:

"Italian Muslims march to denounce Paris attacks: Muslims marched through the streets of Rome to condemn religi... https://t.co/2Wl8sVvo0i"

[9] Given that the attack in question took place in Paris, we make the presumption that the reaction amongst French-speaking Twitter users is representative of reactions from that region. We recognize that using language as a geographic indicator does not strictly hold. However, pinpointing locations in Twitter is problematic (Magdy et al, 2016) and at best partial.

Source	Tweet Sample	tweets	neutral	negative	positive
Paris	Muslim / English	496	186 (37%)	273(55%)	38(8%)
Paris	Muslim / English / GeoRestricted	466	226 (48%)	194(42%)	46(10%)
Berlin	Muslim / English	897	223 (25%)	624 (69.5%)	50 (5.5%)
Berlin	Muslim / English / GeoRestricted	871	207(24%)	600 (66.89%)	64(7.3%)

Table 5: Results for Multilingual Targeted Sentiment Labelling on Twitter samples for '#Paris' between the 20th to the 23rd of November 2015. (excluding where no majority agreement)

Source	Tweet Sample	# tweets	# no majority	# unanimity	Fleiss' Kappa
Berlin	Muslim / English	897	49	396	0.21
Berlin	Muslim / English / GeoRestricted	871	75	339	0.16
Paris	Muslim / English	487	46	147	0.13
Paris	Muslim / English / GeoRestricted	466	67	139	0.13

Table 6: Fleiss' Kappa scores from both Paris and Berlin events.

However, it can be considered positive (given that the instructions were to label the sentiment of the author towards the subject) or at least neutral, if considered as a statement of fact. Comparing with the French tweets, we find the following similar example, which was labelled as positive:

RT @rtlinfo: La communauté musulmane condamne les attentats de Paris.#RTLinfo19h https://t.co/uA7MyohZ9H[10]

On manual examination, we have identified that over 10% of these posts for the 'Muslim' target have wrongly (in our opinion) been labelled as negative, when they should have been either neutral or even positive. The fact that they are labelled negative raises questions about the biases of the crowd sourced annotators as a source of labelling error.

To explore this in more detail, we perform an additional labelling experiment in an attempt to isolate this bias. In particular, we first select two of the English tweet subsets that were markedly more negative than their French/German counterparts. In particular, we select the Muslim target for both the Paris attacks and Berlin attacks datasets. These subsets were originally labelled by English speaking users, predominantly from the UK, USA and Canada. We have these two subsets re-labelled by workers excluding those residing in these regions. We refer to the re-labelled datasets as the GeoRestricted datasets. If annotator bias is not an issue, we would expect that the labels produced by our original and GeoRestricted workers to be similar, i.e. the distribution of sentiments across the three classes would be the same in the original and GeoRestricted subsets.

Table 5 reports the distribution of sentiment labels when comparing the original subset (Muslim / English) and the new GeoRestricted version. From Table 5, comparing the distribution of these GeoRestricted sentiment annotations to the original sentiment annotations (the row above) for the Paris attacks dataset, we observe that 72 (13%) fewer tweets were labelled as negative

[10] Manual Translation: *The Muslim community condemn the attacks in Paris.*

(again excluding items where annotator agreement was below 67%). Hence, for the Paris attacks, we can conclude that workers from the UK, the USA and Canada, are more likely to label posts as about Muslims as negative than workers in other regions. This is in line with findings of Darwish and Magdy (2015) on the source of anti-Muslim sentiment following the attack. On the other hand, examining the re-labelling of the Berlin attacks subset in Table 5, we do not see significant differences in the scores after reannotation. Indeed, from for this event it appears that annotator bias was not as pronounced when examining the resultant sentiment distribution.

However, examining the sentiment distribution is a results-orientated view of the process, i.e. it is a view of the final labels produced. On the other hand, during our crowdsourced labelling task, we have three different workers label each tweet, and then take the majority vote (if one exists). It is possible that the majority voting process is masking annotator biases of a subset of the workers. Hence, it is important to examine the level of agreement between the different workers for the original and GeoRestricted subsets.

Table 6 reports the number of tweets where a majority sentiment was identified, and the agreement in terms of Fleiss Kappa across the three workers for the original and GeoRestricted subsets. From Table 6, we observe that the overall level of agreement between the workers across the two datasets is small (0.21 to 0.13 Fleiss Kappa), particularly for the Paris dataset. However, the overall low agreement is expected, as we previously observed similar behaviour for the original labelling experiment when calculating per-target agreement (that, as discussed earlier, is due to Fleiss Kappa attempting to correct for the heavily imbalanced class distribution for this target). However, we do observe a notable increase in the number of tweets for which no majority was reached, e.g. for the Berlin dataset 49 tweets for the original labels vs. 75 tweets for the GeoRestricted labels. This provides some weak evidence to suggest that as we broaden the geographical regions that the workers are recruited from, more disagreements arise. However, these results are inconclusive due to the small sample size.

To answer **RQ2**, we do indeed observe marked differences between the sentiment labels produced by crowd workers from different geographical regions for one of our two events (the Paris Attacks). On the other hand this finding did not generalize to the smaller second event (Berlin Attack), indicating that geographical annotator bias may be limited to events that received significant international attention. This is an important consideration for designers of future crowdsourced annotation experiments to account for, since otherwise any conclusions drawn from such crowdsourced labels would also be biased. Furthermore, there are important implications when using such biased labels for classification, which we discuss in more detail below.

a	b	c	– classified as
317	381	1	a = negative
195	1000	20	b = neutral
21	93	5	c = positive

Table 7: Confusion matrix for original English Paris dataset

7 Implications for Supervised Classification

A common use for crowdsourced sentiment labels is as training for supervised classification approaches. Hence, in this section, we examine how classification accuracy is affected by the annotator bias we observed in the above section. For this experiment, we aggregate all tweets from each language into a single set and then train using a 10-fold-cross validation. We extract n-gram features, using up to 5-grams to detect longer sequences which include the entity of the targeted sentiment. Table 9 reports the performance of a SVM sentiment classifier trained using scikit-learn[11], in terms of precision, recall and F_1.

From Table 9, we see that when classifying the French tweets, the SVM classifier achieves 0.72 F_1, which is good performance for Twitter (Agarwal et al, 2011; Jiang et al, 2011). In contrast, when classifying the English set, the performance is lower (0.63 F_1). The better scores for the French tweets are potentially biased by the stronger majority class. However, relating these results to our discussion on annotator bias in the previous section, one reason for the markedly lower performance over the English tweets might be that annotator bias from a sub-set of the crowd workers has resulted in inconsistent training labels. To test this, we train a second classifier, where we replace the Muslim / English sample with the re-annotated Muslim / English / GeoRestricted version. Interestingly, as can be observed from Table 9, the replacement of the labels for the Muslim target with the reannotated ones for this subset alone, classification performance drops further to 0.59 F_1. From the confusion matrices in Tables 7 and 8, we observe that the classifier has now wrongly classifies a higher number of negative tweets as neutral, and classifies fewer negative tweets correctly as negative. There are two observations that can be made from this result. First, it would indicate that by moving to a more geographically diverse set of workers that are less likely to exhibit negative bias when labelling, the resultant classifier similarly reflects this by labelling fewer tweets as negative. Second, lessening the biases in the training data does not make the classifier more effective overall (user bias is a signal after all). Indeed, removing user biases in training datasets puts more pressure on sentiment classification systems to learn sentiment patterns, rather than falling-back majority class evidence in scenarios where biases have created very imbalanced classes.

[11] www.scikit-learn.org

a	b	c	– classified as
217	420	2	a = negative
185	1060	18	b = neutral
24	104	3	c = positive

Table 8: Confusion matrix for English Paris dataset with relabelled 'Muslim' subset

Source	Language	Classifier	Tweets	Precision	Recall	F_1
Paris	French	SVM, 10-fold cross validation	2025	0.72	0.76	0.72
Paris	English	SVM, 10-fold cross validation	2033	0.62	0.65	0.63
Paris	English / GeoRestricted	SVM, 10-fold cross validation	2014	0.59	0.63	0.59
Berlin	German	SVM, 10-fold cross validation	2085	0.75	0.74	0.72
Berlin	English	SVM, 10-fold cross validation	2995	0.69	0.70	0.69
Berlin	English / GeoRestricted	SVM, 10-fold cross validation	2178	0.62	0.68	0.62

Table 9: Classification on Berlin dataset with crossfold validation (10-fold).

8 Conclusions

In this paper we illustrated the value of comparative multilingual sentiment analysis as a tool to understand how sentiment about an event varies across national and international regions. Through a crowdsourced user study, we showed that the amount of negativity in the English tweets (34.39%), following the Paris attacks of 2015 far exceeds that of the French (15.09%), despite the fact that the attack was on French soil. This finding is notable as it raises the question of whether monolingual sentiment analysis systems trained on user data may inherit local or international biases that might be undesirable. Furthermore, we examined how bias in crowd annotators can affect the analysis of sentiment during an event. Our results indicate that regional bias can also affect crowdsourced tweet sentiment labelling. Indeed, we observed a 14% reduction in the number of tweets that were labelled as negative for the target 'Muslims' when we excluded workers from the USA, UK and Canada. This regional bias is an important factor to consider when using geographically-dispersed workers to label social media data, particularly when the resultant labels are used as training for supervised classifiers. Finally, we examined the effect that lessening regional biases has on supervised classification approaches, showing that the resultant classifiers also exhibit less bias, but are not more effective overall.

References

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, Stroudsburg, PA, USA, LSM '11, pp 30–38, URL `http://dl.acm.org/citation.cfm?id=2021109.2021114`

Balahur A, Turchi M (2012) Multilingual sentiment analysis using machine translation? In: Proceedings of the 3rd Workshop in Computational Ap-

proaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, Stroudsburg, PA, USA, WASSA '12, pp 52–60, URL http://dl.acm.org/citation.cfm?id=2392963.2392976

Bontcheva K, Derczynski L, Funk A, Greenwood MA, Maynard D, Aswani N (2013) Twitie: An open-source information extraction pipeline for microblog text. In: Angelova G, Bontcheva K, Mitkov R (eds) Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria, RANLP 2011 Organising Committee / ACL, pp 83–90, DOI http://aclweb.org/anthology/R/R13/R13-1011.pdf

Darwish K, Magdy W (2015) Attitudes towards refugees in light of the paris attacks. CoRR abs/1512.04310, URL http://arxiv.org/abs/1512.04310

De Choudhury M, Diakopoulos N, Naaman M (2012) Unfolding the event landscape on twitter: classification and exploration of user categories. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, ACM, pp 241–244

Dong L, Wei F, Tan C, Tang D, Zhou M, Xu K (2014) Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers, pp 49–54, URL http://aclweb.org/anthology/P/P14/P14-2009.pdf

Dredze M, Paul MJ, Bergsma S, Tran H (????) Carmen: A twitter geolocation system with applications to public health. In: AAAI workshop on expanding the boundaries of health informatics using AI (HIAI), pp 20–24

Hermida A (2010) Twittering the news: The emergence of ambient journalism. Journalism practice 4(3):297–308

Hsueh PY, Melville P, Sindhwani V (2009) Data quality from crowdsourcing: a study of annotation selection criteria. In: Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing, Association for Computational Linguistics, pp 27–35

Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, pp 151–160

Kiritchenko S, Mohammad SM (2016) Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In: American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL), pp 811–817

Kraaij W, Spitters M (2003) Language models for topic tracking. In: Language Modeling for Information Retrieval, Springer, pp 95–123

Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, ACM, pp 591–600

Magdy W, Darwish K, Abokhodair N (2015) Quantifying public response towards islam on twitter after paris attacks. CoRR abs/1512.04570, URL http://arxiv.org/abs/1512.04570

Magdy W, Darwish K, Abokhodair N, Rahimi A, Baldwin T (2016) #isisisnotislam or #deportallmuslims?: Predicting unspoken views. In: Proceedings of the 8th ACM Conference on Web Science, ACM, New York, NY, USA, WebSci '16, pp 95–106, DOI 10.1145/2908131.2908150, URL http://doi.acm.org/10.1145/2908131.2908150

Marcheggiani D, Täckström O, Esuli A, Sebastiani F (2014) Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In: Advances in Information Retrieval, Springer, pp 273–285

Maynard D, Bontcheva K (2016) Challenges of evaluating sentiment analysis tools on social media. In: Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016., European Language Resources Association (ELRA), URL http://www.lrec-conf.org/proceedings/lrec2016/summaries/188.html

Maynard D, Greenwood MA (2014) Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In: Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014., European Language Resources Association (ELRA), pp 4238–4243, URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/67.html

Maynard D, Hare JS (2015) Entity-based opinion mining from text and multimedia. In: Gaber MM, Cocea M, Wiratunga N, Göker A (eds) Advances in Social Media Analysis, Studies in Computational Intelligence, vol 602, Springer, pp 65–86, DOI 10.1007/978-3-319-18458-6_4, URL http://dx.doi.org/10.1007/978-3-319-18458-6_4

McCreadie R, Macdonald C, Ounis I (2013) Identifying top news using crowdsourcing. Information Retrieval 16(2):179–209

Moilanen K, Pulman S (2009) Multi-entity sentiment scoring. In: RANLP, pp 258–263

Mozetič I, Grčar M, Smailović J (2016) Multilingual twitter sentiment classification: The role of human annotators. PLoS ONE 11(5):1–26, DOI 10.1371/journal.pone.0155036, URL http://dx.doi.org/10.1371%2Fjournal.pone.0155036

Nagy A, Stamberger J (2012) Crowd sentiment detection during disasters and crises. In: Proceedings of the 9th International ISCRAM Conference, pp 1–9

Narr S, De Luca EW, Albayrak S (2011) Extracting semantic annotations from twitter. In: Proceedings of the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval, ACM, New York, NY, USA, ESAIR '11, pp 15–16, DOI 10.1145/2064713.2064723, URL http://doi.acm.org/10.1145/2064713.2064723

Ounis I, Amati G, Plachouras V, He B, Macdonald C, Lioma C (2006) Terrier: A high performance and scalable information retrieval platform. In: Proceedings of the OSIR Workshop, pp 18–25

Ounis I, Macdonald C, Soboroff I (2008) Overview of the trec-2008 blog track. Tech. rep., GLASGOW UNIV (UNITED KINGDOM)

Purver M, Battersby S (2012) Experimenting with distant supervision for emotion classification. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp 482–491

Schulz A, Thanh T, Paulheim H, Schweizer I (2013) A fine-grained sentiment analysis approach for detecting crisis related microposts. Conference on Information Systems for Crisis Response and Management (ISCRAM)

Stieglitz S, Dang-Xuan L (2013) Emotions and information diffusion in social mediasentiment of microblogs and sharing behavior. Journal of Management Information Systems 29(4):217–248

Tang D, Wei F, Qin B, Liu T, Zhou M (2014) Coooolll: A deep learning system for twitter sentiment classification. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pp 208–212, URL http://www.aclweb.org/anthology/S14-2033

Tang D, Qin B, Feng X, Liu T (2016a) Effective lstms for target-dependent sentiment classification. In: Calzolari N, Matsumoto Y, Prasad R (eds) COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, ACL, pp 3298–3307, URL http://aclweb.org/anthology/C/C16/C16-1311.pdf

Tang D, Wei F, Qin B, Yang N, Liu T, Zhou M (2016b) Sentiment embeddings with applications to sentiment analysis. Knowledge and Data Engineering, IEEE Transactions on 28(2):496–509, DOI 10.1109/TKDE.2015.2489653

Thelwall M, Buckley K, Paltoglou G (2011) Sentiment in twitter events. J Am Soc Inf Sci Technol 62(2):406–418, DOI 10.1002/asi.21462, URL http://dx.doi.org/10.1002/asi.21462

Tromp E (2012) Multilingual Sentiment Analysis on Social Media: An Extensive Study on Multilingual Sentiment Analaysis Performed on Three Different Social Media. LAP Lambert Academic Publishing, URL http://books.google.nl/books?id=ut4yLgEACAAJ

Vargas S, McCreadie R, Macdonald C, Ounis I (2016) Comparing overall and targeted sentiments in social media during crises. In: Tenth International AAAI Conference on Web and Social Media

Verma S, Vieweg S, Corvey WJ, Palen L, Martin JH, Palmer M, Schram A, Anderson KM (2011) Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In: International AAAI Conference on Web and Social Media (ICWSM)

Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S (2012) A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics, pp 115–120