Analyzing Disproportionate Reaction via Comparative Multilingual Targeted Sentiment in Twitter

Karin Sim Smith, Richard McCreadie, Craig Macdonald, Iadh Ounis
University of Glasgow
Email: firstname.lastname@glasgow.ac.uk

***Abstract*—Global events such as terrorist attacks are commented upon in social media, such as Twitter, in different languages and from different parts of the world. Most prior studies have focused on monolingual sentiment analysis, and therefore excluded an extensive proportion of the Twitter userbase. In this paper, we perform a multilingual comparative sentiment analysis study on the terrorist attack in Paris, during November 2015. In particular, we look at targeted sentiment, investigating opinions on specific entities, not simply the general sentiment of each tweet. Given the potentially inflammatory and polarizing effect that these types of tweets may have on attitudes, we examine the sentiments expressed about different targets and explore whether disproportionate reaction was expressed about such targets across different languages. Specifically, we assess whether the sentiment for French speaking Twitter users during the Paris attack differs from English-speaking ones. We identify disproportionately negative attitudes in the English dataset over the French one towards some entities and, via a crowdsourcing experiment, illustrate that this also extends to forming an annotator bias.**

I. 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 that event. For this reason, social media is a valuable resource to help understand how events are being perceived by different social groups. However, most social media studies only analyse content in a single language (typically English) (Thelwall, Buckley, and Paltoglou, 2011; Vargas et al., 2016), and hence exclude a large proportion of the social media userbase.

In contrast, in this paper, we present a study of both English and French language tweets posted following the terrorist attack that took place in Paris on the 20th November 2015. English and French were the most frequent languages tweeted in following the attacks, with the largest amount of tweets being in English, followed by French. In particular, we analyse how sentiment expressed about the targets involved in the event on Twitter differs between users writing in different languages and explore the challenges in accurately identifying such sentiments. While there have been prior sentiment analysis studies that aim to detect varying opinion following the Paris attack (Magdy, Darwish, and Abokhodair, 2015), our work is different, as it both examines sentiment *about* particular targets of interest, and more importantly provides insights into how sentiment varies across geographical regions, as well as some implications of this variance.

More precisely, we analyse sentiment towards three different targets for the 80 hours after the attack, namely French President François Hollande, Europe and Muslims. These were chosen as being of significance for this event. We use crowdsourcing over English and French tweet samples for each target to track public sentiment about those targets. Interestingly, based on the labels produced, we find a markedly higher proportion of negative sentiment expressed towards the targets by users posting in English than users posting in French, even though the attack in question took place in Paris. Indeed, if we assume that sentiment expressed by people local to the event is a reasonable baseline against which reactions can be compared, we show that by contrast the reaction by English-speaking regions to the event was disproportionately negative. We also examine annotator bias of the crowdsourced workers by comparing annotations from workers in different regions. We show that there are marked differences in the annotations produced from users in different regions when labelling the same data.

The contributions of this work are two-fold: First, we show that 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; Second, we examine how annotator bias can affect the analysis of sentiment during an event, showing that regional bias also affects (crowdsourced) tweet labelling. 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 before defining our task (Section III) and experimental setup (Section IV). We then examine the Twitter user bias (Section V) and the annotator bias (Section VI), as well as discuss implications for building automatic classifiers (Section VII). We summarize our conclusions in Section VIII.

ASONAM '17, July 31 - August 03, 2017, Sydney, Australia
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4993-2/17/07 $15.00
http://dx.doi.org/10.1145/3110025.3110066

II. Related Work

Previous work on monolingual sentiment in Twitter has included analysis following important events (Thelwall, Buckley, and Paltoglou, 2011), finding negative sentiment generally exceeds positive sentiment, including for positive events. Again in a monolingual setting Agarwal et al. (2011) used rich linguistic features in a tree kernel to improve Twitter sentiment detection. Vargas et al. (2016) as well as Jiang et al. (2011) investigated targeted sentiment in monolingual settings, but not as comparative multilingual analysis. There has also been work on multilingual Twitter sentiment analysis (Narr, De Luca, and Albayrak, 2011; Tromp, 2012), although not targeted towards specific entities, and in the case of the latter, in a language-independent manner. In their multilingual study, Mozetič, Grčar, and Smailović (2016) compared human labelling and classification models, hypothesizing that 'the inter-annotator agreement approximates an upper bound for a classifier performance'. In deeper monolingual analysis on the public response in Twitter following this same attack, Magdy et al. (2016) predict stance, particularly towards Muslims, based on user profile. They use retweets and 'likes' as a benchmark in researching emotional reaction (Magdy, Darwish, and Abokhodair, 2015). In contrast, we are interested in the textual content, and the basis of our work is a multilingual approach, which is comparative and targeted in nature, as well as being focused on one particular but important event.

III. Task Definition

In this paper, we analyse how sentiment expressed about an event on Twitter differs between users writing in different languages and explore the challenges in accurately identifying such sentiments. More precisely, for a tweet post p that is part of a larger discussion about a sensitive event e and that also mentions a particular entity of interest (target) t, we analyse whether that post p expresses sentiment ($s \in \{negative, positive, neutral\}$) about its target t. To support this analysis, we use crowdsourced workers to label the sentiments expressed within tweet samples in two languages (French and English) for a major event e (the 2015 Paris attacks). We answer two main research questions:

RQ1: Do the sentiments expressed towards the targets differ among the French & English speaking Twitter users?

RQ2: Do the sentiment labels about the targets differ among the geographically diverse crowdsourcing workers?

IV. Experimental Setup

Dataset: The dataset we base our analysis on consists of Twitter tweets posted during Paris attack on 20th-23rd Nov 2015, containing '#Paris'. This crawl contains tweets in a wide variety of languages. We filter on the language using the 'lang' tag, which identifies the language via Twitter's own language classifier[1]. According to this classifier, the most common language was English (1,232,100 tweets) followed by French (402,914 tweets).

Sentiment Targets: Manually analysing millions of tweets is not feasible due to time/cost constraints. Hence, we choose a small number of entities (targets) of interest to analyse in detail. In particular, we select French President François Hollande, the European Union and Muslims as our targets. We filter the above 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.

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 (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[2] for ranking. We select the top 100 tweets from each hour to create the sample for each set. We then remove near-duplicate tweets from each sample by applying a cosine similarity threshold τ over that sample in a greedy time-ordered manner (τ= 0.7).

Crowdsourcing 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. As in earlier work on targeted sentiment labelling (Vargas et al., 2016), 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 the target given. For the three English tweet samples, only English-speaking users were allowed to participate in labelling those samples; similarly only French-speaking users could label the three French tweet samples. To avoid a few users 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 45 (French) or 48 (English) tweets, labelled by the authors, fluent in both languages. 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 I.

Reproducibility: The tweet samples and crowdsourced labels used for evaluation are available as a free download at:

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

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

TABLE I
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	neutral	negative	positive
Paris	All / French	1998	1521 (76%)	312 (16%)	165 (8%)
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	All / English	1997	1199 (60%)	681 (34%)	118 (6%)
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%)
Paris	Muslim / English / GeoRestricted	466	226 (48%)	194(42%)	46(10%)

http://dx.doi.org/10.5525/gla.researchdata.42

V. Twitter User Bias

Table I reports the number and proportion of tweets from each of six tweet samples that were labelled as containing either neutral, negative or positive sentiment. As we can see from Table I, there is a clear polarity over the various targets. For example, for the target 'Hollande', the polarity breakdown is similar across the two languages. There is a similar proportion of the French tweets that constitute negative sentiment (23.5%), as for the English (22%). The proportion of French tweets that are positive for this target is (11.7%). Whereas the English tweets analysed were less positive in their judgement of him, as indicated by the lower positive score (8%). However, what is particularly striking is the significant discrepancy between 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 14.2% 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% tweets labelled as negative by French annotators, and 32% were labelled as negative by English annotators. Hence, to answer **RQ1**, there are marked differences in the sentiments expressed by Twitter users in different geographical regions.

This result is unexpected, since those in Paris (and France more generally) 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.

VI. 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 given subject of 'Muslim', by the English speaking annotators:

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

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[3]

On manual examination, we identified that over 10% of these posts for the 'Muslim' target have wrongly 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 annotators. To explore this in more detail, we perform an additional labelling experiment. In particular, we restrict the geographical location of our annotators to prevent users from the UK, Australia, the USA and Canada from participating, and then re-label the Muslim/English sample using a new pool of crowd workers, which we refer to as Muslim / English / GeoRestricted.

The last row in Table I shows the distribution of sentiment labels from this additional annotation experiment. If we compare the sentiment distribution of these sentiment annotations to the original sentiment annotations (the row above), we observe that 72 (13%) fewer tweets were labelled as negative (again excluding items where annotator agreement was below 67%). This indicates that the workers from the UK, Australia, the USA and Canada, are more likely to label posts about Muslims as negative than workers in other regions. Of the 1599 individual labels, 1031 were USA-based workers, 88 Canada, and 480 UK. This is in line with findings of Darwish and Magdy (2015) on the source of anti-Muslim sentiment following the attack, where they found that the largest amount of anti-Muslim sentiment following this attack was in fact in the USA.

To answer **RQ2**, we do indeed observe marked differences between the sentiment labels produced by crowdsourced workers from different geographical regions. This is an important consideration for future crowdsourced annotation experiments,

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

TABLE II
CLASSIFICATION WITH CROSSFOLD VALIDATION (10-FOLD).
RELABELLED RESULTS INCLUDE REPLACED 'MUSLIM' SECTION ONLY.

language	tweets	precision	recall	F_1
French	2025	0.72	0.76	0.72
English	2033	0.62	0.65	0.63
Substitute relabelled *Muslim* set:				
English	2014	0.59	0.63	0.59

since otherwise any conclusions drawn from such labels would also be biased. Furthermore, there are implications when using such biased labels for classification, which we discuss below.

VII. 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 ($1 \leq n \leq 5$) to detect longer sequences which include the entity of the targeted sentiment. Table II reports the accuracy of a SVM (SGD) sentiment classifier trained using scikit-learn, in terms of precision, recall and F_1.

From Table II, we see that when classifying the French tweets, the SVM classifier achieves 0.72 F_1, which is a good performance for Twitter (Agarwal et al., 2011; Jiang et al., 2011). Interestingly, when classifying the English set, the performance is lower (0.63 F_1). The better scores for the French tweets are 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 trained a second classifier, where we replaced the Muslim/English sample in the original dataset with the re-annotated Muslim / English / GeoRestricted version. Interestingly, as we see from Table II the replacement of the labels for the Muslim target with the reannotated ones for this subset, results in a drop in F_1 to 0.59. This can be attributed to the fact that there is a more even label distribution, instead of a majority class. Also, there clearly is ambiguity in the English tweets, as the following tweet was labelled with 3 different labels:

> *A small antidote to political vitriol. "Muslim asks Parisians to hug him if they trust him. Many do". https://t.co/xXlzGglyJx*

The increased ambiguity is clear from the lack of annotator agreement, which led to us then having to disregard more tweets for that target (496 to 466).

However, while our results highlight the limitations of human annotators, it is worth noting that this is only one cause of classifier error. For instance, some of the misclassified negative tweets are quite subtle, or nuanced, or instances where the classifier cannot grasp the cynicism for instance:

> *"UK Muslims Feel Backlash After Paris Attacks.Alway moaning Oh look at how bad it is for us!!"*

There are also tweets which are overtly racist, but where the classifier would struggle to detect negativity, as it is too subtle:

> *"I've been locked in a cupboard since the Paris attacks and am starving to death. Anyone know a delivery service that doesn't employ muslims?"*

Indeed, detecting the negativity in these tweets may be simple for humans, but requires more sophisticated classifier features.

VIII. 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 geographical 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. Furthermore, we examined how bias in crowd annotators can affect the analysis of sentiment during an event. Our results indicated that regional bias can have a strong influence when crowdsourcing tweet sentiment labels. 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 as social media data, particularly when the resultant labels are used as training for supervised classifiers.

IX. ACKNOWLEDGEMENTS

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

REFERENCES

Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; and Passonneau, R. Sentiment analysis of Twitter data. In *Proceedings of LSM 2011*.

Darwish, K., and Magdy, W. Attitudes towards refugees in light of the Paris attacks. *Computing Research Repository Journal*.

Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; and Zhao, T. Target-dependent Twitter sentiment classification. In *Proceedings of ACL-HLT 2011*.

Kraaij, W. and Spitters, M. Language models for topic tracking. *Language Modeling for Information Retrieval*.

Magdy, W.; Darwish, K.; Abokhodair, N.; Rahimi, A.; and Baldwin, T. #isisisnotislam or #deportallmuslims?: Predicting unspoken views. In *Proceedings of WebSci 2016*.

Magdy, W.; Darwish, K.; and Abokhodair, N. Quantifying public response towards Islam on Twitter after Paris attacks. *Computing Research Repository Journal*.

Mozetič, I.; Grčar, M.; and Smailović, J. Multilingual Twitter sentiment classification: The role of human annotators. *Public Library of Science ONE Journal*.

Narr, S.; De Luca, E. W.; and Albayrak, S. Extracting semantic annotations from Twitter. In *Proceedings of ESAIR 2011*.

Ounis, I.; Amati, G.; Plachouras, V.; He, B.; Macdonald, C.; and Lioma, C. Terrier: A high performance and scalable Information Retrieval platform. In *Proceedings of OSIR 2006*.

Thelwall, M.; Buckley, K.; and Paltoglou, G. 2011. Sentiment in Twitter events. *American Society for Information Science and Technology Journal*.

Tromp, E. 2012. *Multilingual Sentiment Analysis on Social Media*. LAP Lambert Academic Publishing.

Vargas, S.; McCreadie, R.; Macdonald, C.; and Ounis, I. Comparing overall and targeted sentiments in social media during crises. In *Proceedings of ICWSM 2016*.