# Exploring deep learning for the analysis of emotional reactions to terrorist events on Twitter

Jonathas G. D. Harb, Régis Ebeling, Karin Becker

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil
{jgdharb, tebeling, karin.becker}@inf.ufrgs.br

**Abstract.** Terrorist events have a substantial emotional impact on the population, and understanding these effects is very important to design effective assistance programs. However, investigating community-wide traumas is a complex and costly task, where most challenges are related to the data collection process. Social media has been used as a relevant source of data to investigate people's sentiments and ideas. In this article, we study the emotional reactions of Twitter users regarding two terrorist events that occurred in the United Kingdom. The contributions are twofold: a) we experiment two deep learning architectures to develop an emotion classifier, and b) we develop an analysis on tweets related to terrorist events to understand whether there is an emotional shift due to a terrorist attack and whether the emotional reactions are dependent on the event, or on the demographics of the users. Both models, based on convolutional and recurrent neural architectures, presented very similar performances. The analyses revealed an emotion shift due to the events and a difference in the reactions to each specific event, where gender is the most significant factor.

Categories and Subject Descriptors: Information Systems [**Data Mining**]: Data Streaming; Artificial Intelligence [**Machine Learning**]: Supervised Learning; Information Systems [**Web Applications**]: Social Network

Keywords: Sentiment Analysis, Deep learning architecture, Twitter, Terrorism

## 1. INTRODUCTION

Terrorism in all its forms remains a constant threat that continues to be present in the global agenda and raises questions concerning prevention and consequences. Hoffman (2013) defines terrorism as "the deliberate creation and exploitation of fear through violence or the threat of violence in the pursuit of political change". The goal of terrorism is to create instability by propagating fear, arousal, and uncertainty on a broader scale in comparison to targeting a single victim [Horgan 2014]. Social media and 24-hour news coverage of the attacks and their aftermath reach far beyond the affected communities, to the entire nation, and beyond [Lowe et al. 2015]. Reactions to terrorist events include, among others, losing the sense of safety, feeling helpless, experiencing anger and fear, and intolerance towards certain ethnic or religious groups. Understanding the emotional reactions of the population regarding these events is very important to design assistance programs that effectively help the population to deal with these issues [Crepeau-Hobson et al. 2012; Maguen et al. 2008; Cohen-Louck and Ben-David 2017].

The study of the emotional impact of community-wide trauma is a complex and costly task. Most challenges are related to the data collection process, including the inability to control the respondents' behavior prior to the event; the timing for collecting data; and difficulty of getting access to traumatized community members. Researchers are circumventing these challenges by exploring data extracted from social media as an alternative [Jones et al. 2016]. Twitter is a popular social media platform used for posting real-time discussions, thoughts, sentiments, and opinions with regard to several topics. Sentiment Analysis deals with the automatic extraction and interpretation of people's opinions, attitude, and emotions from documents [Liu 2012]. While opinion mining addresses the *polarity* of the sentiment towards a target (i.e. positive, negative), emotion mining focus on the identification of other affect states according to an *emotion* model (e.g. basic emotions model, such as joy, anger or fear) [Munezero et al. 2014].

Sentiment analysis was deployed in terrorism-related tweets to study post-event emotional contagion [Chong 2016] and information diffusion model [Burnap et al. 2014; Garg et al. 2017; Simon et al. 2014], as well as to identify sentiment towards terrorist organizations [Azizan and Aziz 2017; Mirani and Sasi 2016; Mansour 2018]. All of these works are restricted to *polarity* analysis.

In this article, we explore deep learning techniques to analyze the *emotions* people express on Twitter about terrorist events. More specifically, we: a) experiment with two deep learning architectures to develop emotion classifiers targeted at this type of event, namely convolutional and recurrent neural networks [Murphy 2012], and b) deploy an emotion classification model to analyze the emotional reactions based on the demographics of the tweeters, particularly gender, age and location. We collected and analyzed data on two terrorist events that occurred in England. Our analysis aims to answer the following research questions:

Q1: Is there an emotion shift due to terrorist events?

Q2: Do different terrorist events evoke the same emotional reaction?

Q3: Does the proximity to the event influence the emotional reaction?

This study is an extension of our previously presented work [Harb and Becker 2018]. We complement it by the development and comparison of emotion classification models according to two distinct deep learning architectures; improvements on the method used to analyzes tweets to answer the research questions, as well as by the update and deeper analysis of related work. A complementary work [Harb and Becker 2019] has deployed the original emotion classifier to analyze other terrorist events.

In comparison to related work, the main contributions of our research are: a) we focus on emotional reactions using the basic emotions model [Ekman and Friesen 1982], whereas related work address polarity [Burnap et al. 2014; Garg et al. 2017; Simon et al. 2014; Chong 2016; Azizan and Aziz 2017; Mirani and Sasi 2016]; b) we explore distinct deep learning architectures for emotion classification, in opposition to works that rely on feature engineering [Azizan and Aziz 2017; Mirani and Sasi 2016]; c) inspired by works such as [Sakaki et al. 2010; ElSherief et al. 2017; Walter and Becker 2018], we use information extracted from users' profiles to analyze emotions according to users' demographics (i.e. location, gender, and age), which is a novel aspect of sentiment analysis in the terrorism domain, and d) like [Jones et al. 2016], we also collected pre-event tweets in addition to post-event tweets to evaluate the emotional shift.

The remainder of this article is structured as follows. Section 2 summarizes the theoretical background, and Section 3 describes related work. Section 4 details the methods and materials used for providing answers to the defined research questions. Section 5 describes the experiments that compare the performance of the developed emotion classifiers. Section 6 presents the analysis performed over the data to answer the research questions. Finally, Section 7 presents the conclusions and opportunities for future works.

## 2. THEORETICAL BACKGROUND

### 2.1 Sentiment Analysis

Sentiment Analysis deals with the automatic extraction and interpretation of people's opinions, attitude, and emotions from documents [Liu 2012]. It is a process that encompasses different steps, including pre-processing the input texts, classification of the sentiment they convey, and aggregation of the sentiment contained in different documents. Although sentiment is frequently measured in terms of *polarity* (i.e. positive, negative, or neutral), it can also be evaluated according to an emotion model [Munezero et al. 2014]. Ekman's basic emotions [Ekman and Friesen 1982] (e.g. joy, anger, fear) is the most popular emotion model for emotion mining.

There are two basic approaches for sentiment classification: lexicon-based and machine learning. The former relies on the use of sentiment lexicons (e.g. SentiWordNet [Baccianella et al. 2010], NRC [Mohammad and Turney 2013a], in which each entry is associated with a sentiment measure. This approach tends not to perform well in tweets, as many expressions are not be contained in the dictionary due to the use of abbreviations, informal language, and emoticons [Liu 2012; Zimbra et al. 2018]. However, the performance of dictionaries is limited as the targeted domain usually is not taken into account, nor new vocabulary and internet-specific sentiment expressions [Zimbra et al. 2018].

Machine learning-based approaches apply supervised learning algorithms to derive associations between features extracted from documents and sentiments. Support Vector Machine (SVM) and Naive Bayes (NB) are popular classification algorithms to this end. The approach is combined with feature engineering to deal with the particular characteristics of tweets [Mohammad et al. 2013]. More recently, deep learning has been deployed for sentiment analysis [Zhang et al. 2018]. The performance of the supervised learning process depends on the quality and size of an annotated corpus, preferably domain-specific [Liu 2012; Zimbra et al. 2018]. The manual annotation of training sets is a costly activity, and therefore many works propose strategies for automatically labeling a corpus, such as the presence of emoticons [Go et al. 2009], emotion hashtags [Mohammad 2012; Wang et al. 2012], or sentiment words from a lexicon [Azizan and Aziz 2017; Mirani and Sasi 2016]. In the case of deep learning, the volume of training instances is even more critical.

### 2.2 Deep Learning

Deep learning has emerged as a powerful technique that allows computational models to learn representations of large sets of data using computing power. Deep learning architectures are different from traditional neural networks, as their structure is comprised of more layers and more units within a layer. In a nutshell, deep learning uses a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. The lower layers, closer to the data input, learn simpler features, while higher layers learn more complex features derived from lower layer ones. The architecture forms a hierarchical and powerful feature representation [Murphy 2012]. The result is a set of features that hierarchically grows in complexity.

Deep learning has become increasingly popular for text classification in general, and sentiment analysis in particular [Zhang et al. 2018]. Typically, deep learning models for textual data rely on *word embeddings* as input features. Word embeddings are low-dimensional dense vectors representations learned from data, such that words that frequently appear in similar contexts are close to each other. Word embeddings can be learned from the input corpora by an embedding layer in the neural network architecture, or produced independently using an unsupervised machine learning algorithm such as Word2vec [Mikolov et al. 2013]. Pretrained embeddings (e.g. GloVe[1]) can also be leveraged, as they are trained using a huge corpus that reflects general vocabulary usage. Pre-trained embeddings can

[1] https://nlp.stanford.edu/projects/glove/

compose static or non-static models. In the former, the layer referring to word embeddings is frozen, preventing their weights from being updated during training. In non-static models, word vectors are initialized according to the pre-trained embeddings, and these weights are updated through back-propagation during training.

A survey [Zhang et al. 2018] summarizes a number of studies that propose deep learning architectures for sentiment analysis. For short texts such as tweets and sentences, variations of LSTM (Long-Short Term Memory) and convolutional neural networks (CNN) are the most deployed architectures to learn the intrinsic semantic and syntactic relationships between words.

As a recurrent neural network, LSTM architectures are suitable for handling sequence elements since they maintain a state relative to what has been processed so far. LSTM architectures have been successfully applied to natural language processing applications, such as speech or hand writing recognition [Greff et al. 2017]. In sentiment analysis, they are successful under the premise that the order of words in documents are representative of the sentiment conveyed. Typically, the architecture is organized in terms of layers of LSTM units, where each unit is a cell composed of an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell.

Convolutional neural networks (CNNs) utilize layers with convolving filters. In a convolution, a filter slides (convolves) over the input space to find local patterns, and generate feature maps. Typically a convolutional layer applies different filters. Following a convolutional layer, a pooling layer is used to reduce the spatial size of the extracted representation progressively, and thus to reduce the number of features. Convolutional layers may be organized in a hierarchy, and finally, serve as input to a dense layer. CNNs were initially proposed for computer image problems, but have achieved particularly good results in traditional Natural Language Processing (NLP) tasks, including sentiment analysis [Kim 2014; Shen et al. 2014; Collobert et al. 2011]. CNNs have a particular spatially local correlation by enforcing a local connectivity pattern between neurons of adjacent layers. Such a characteristic is useful for classification in NLP, in which we expect to find meaningful local clues regarding class membership, such as the combination of specific terms/phrases, regardless where they appear in a document [Zhang et al. 2018].

One of the most popular CNN architecture for text analysis was proposed by Kim in for sentiment analysis at sentence level [Kim 2014]. Through his experiments, Kim concluded that even a CNN with one layer of convolution performs remarkably well when combined with pre-trained *word embeddings.* He observed that a simple model with static vectors (CNN-static, where *word embeddings* are pre-trained and fixed) provides good results, and that the fine-tuning the pre-trained vectors (CNN-non-static) yields further improvements.

In this work, we compare the performance of emotion classifiers based on CNN and LSTM architectures to predict emotions in terrorism-related tweets.

## 3. RELATED WORK

All works that leverage sentiment analysis techniques to understand terrorism develop their respective analysis on collected tweets using keywords (hashtags, terms, or URLs) that describe a terrorist event or terrorist organization (e.g. #prayForParis, ISIS).

Some studies focus on the analysis of sentiments after a terrorist event [Chong 2016; Burnap et al. 2014; Garg et al. 2017; Simon et al. 2014]. A study [Chong 2016] confirmed that the emotional contagion theory also applies to social media by analyzing tweets related to a series of coordinated terrorist attacks that occurred in Paris. The influence of the sentiment in the tweets' information flow model related to terrorist events was also investigated, revealing that emotive content is predictive of both size and survival of information flows [Burnap et al. 2014], and that negative tweets tend to survive more than the positive ones [Garg et al. 2017]. Another study investigated tweets related to a four-day siege after a terror attack in Kenya [Simon et al. 2014] to assess how social media could contribute to crisis management. In all of these works, the sentiment is analyzed in terms of polarity, and was determined using off-the-shelf tools (e.g. SentiStrengh[2], Alchemy[3]).

Other studies investigated the sentiment towards terrorism, also in terms of polarity. Using a sentiment lexicon [Liu et al. 2005] and geotagged ISIS-related tweets from different countries, a study [Mansour 2018] compared how people from Western and Eastern countries view ISIS, concluding that there are no significant differences. Other studies [Mirani and Sasi 2016; Azizan and Aziz 2017] combined feature engineering and machine learning methods, using datasets automatically labeled based on the presence of words found in sentiment lexicons. To assess sentiment classification models targeted at tweets about ISIS [Mirani and Sasi 2016], Mirani et al. trained five different algorithms (e.g. Naive Bayes, SVM) on a dataset labeled using Opinion lexicon [Liu et al. 2005]. An approach for detecting pro-terrorism behavior of Twitter users is described in [Azizan and Aziz 2017], in which a Naive Bayes classifier was trained on a dataset automatically labeled based on terms contained in the SentiWordNet lexicon.

A study [Jones et al. 2016] investigated whether or not Twitter could be a viable data source to study the emotional impact of mass violent events on communities. Using three violent college shooting events, they collected pre/post-event tweets from users located within impacted communities and identified sentiment emotions using LIWC[4]. To define a pre-event control group, they collected tweets from accounts likely to be followed by community residents only (e.g. radio station, official college accounts). Keywords manually selected as representative of the attack were used to filter the tweets explicitly related to the event. In a more recent study [Jones et al. 2019], the post-traumatic effects of a mass shooting occurred in San Bernardino (USA) have been investigated. Post-event tweets were collected using specific hashtags. They assumed that people informing the location of the shooting in their profile were residents, and used that information to crawl pre-event tweets.

Compared to related work, the distinctive features of our work are: a) we investigate emotions, instead of polarity; b) we explore deep learning techniques to avoid relying on feature engineering; c) we use information extracted from users' profiles [Sakaki et al. 2010; ElSherief et al. 2017; Walter and Becker 2018] to analyze whether the specific emotions of the users are related to gender, age or the proximity to the event (location), and d) in addition to post-event tweets, we also collected pre-event tweets by adapting the method proposed in [Jones et al. 2016].

## 4. MATERIALS AND METHODS

### 4.1 Target Events

We targeted two terrorist events that occurred in the United Kingdom (UK). This choice was motivated by two factors. First, we focused on the English language in order to benefit from the resources available for natural language processing. Second, both events occurred in the same country (UK), a few days apart. These events and their aftermath drew a lot of attention from the world media, and thus their impact is not restricted to their respective local communities and surroundings.

The first event was the Manchester Arena bombing[5], which took place in Manchester on May 22nd 2017, when people were leaving a concert of Ariana Grande. The second one was the London Bridge attack[6], which occurred in London on June 3rd 2017, where a van left the road and struck a number

[2] http://sentistrength.wlv.ac.uk/
[3] Currently, Alchemy is part of IBM Watson system.
[4] www.liwc.net
[5] https://en.wikipedia.org/wiki/Manchester_Arena_bombing
[6] https://en.wikipedia.org/wiki/June_2017_London_Bridge_attack

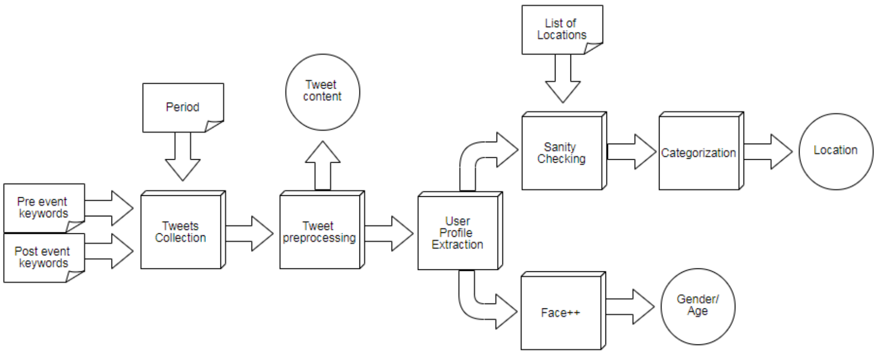

Fig. 1. Data collection and preparation process.

of passing by pedestrians.

## 4.2 Dataset Preparation

The method to create the datasets explored in this work is depicted in Figure 1. First, we collect tweets using the period and keywords that define the pre-event and the post-event tweets. Next, we pre-process these tweets with traditional cleaning actions (e.g. removal of hyperlinks and mentions, capitalization, spelling checking, etc.) in order to improve the quality of emotion classification. Finally, we extract from the users' profile demographic information, namely gender, age, and location. These steps are detailed in the remaining of this section. The resulting datasets, as well as the Gold Standard (Section 4.3), are available at a public repository[7].

4.2.1 *Data Collection and Pre-processing.* Our research involves both pre and post-event tweets, such that we are able to analyze the emotional reaction to an attack and compare it to the previous situation. Adapting the strategy defined in [Jones et al. 2016; Jones et al. 2019], we collected both pre/post-event tweets using search terms that characterized the community affected by the attack, and the event itself.

To compose the datasets of pre-event tweets, we queried tweets using the generic keywords referring to the cities where the events took place (i.e. London and Manchester). We observed that these keywords were commonly used to tweet about local citizen's thoughts on diverse topics such as soccer teams, universities, concerts, and daily news regarding these locations [Jones et al. 2019].

Post-event tweets were collected according to two strategies:

—keywords referring to the attacks: to identify keywords representative of the terrorist events, we manually inspected Twitter trending topics, raw data gathered from the web, as well as samples extracted using the official Twitter API on the respective event dates. We found recurrent hashtags for each one of the events, namely, #*prayformanchester* for the Manchester attack and #*londonbridge* for the London one. These hashtags were nearly 15 times more frequent than the second most frequent term used to denote an event. Thus, we assume these hashtags are representative due to their prevalence in tweets referring to these events.

—keywords referring to the locations (i.e. Manchester and London): In our original work [Harb and Becker 2018], we used only the above strategy to compose the post-event dataset. However, we realized that results regarding the analysis of emotional shifts could be biased if we considered only people who were clearly impacted by the event to the point they would tweet about it using specific event hashtags. Thus, we additionally adopted the same strategy used to crawl pre-event tweets. In

[7] https://github.com/regisebeling/Terrorism-Twitter-Dataset

Table I. Query Terms, Dates and Datasets per Event

| **Event Name** | **Query terms** | **Period** | **BEFORE** | **AFTER** | **AFTER** | **AFTER** |
|---|---|---|---|---|---|---|
| | | | (#tweets - city) | (#tweets - both) | (#tweets - #event) | (#tweets - city) |
| #prayformanchester | #prayformanchester, “Manchester” | 05-20-2017 to 05-24-2017 | BM (5,351) | AM (929,518) | 25,010 | 904,508 |
| #londonbridge | #londonBridge, “London” | 06-01-2017 to 06-05-2017 | BL (20,379) | AL (392,711) | 29,656 | 363,055 |

this way, we capture both users concerned about the attack, and those concerned with other events related to the location.

For each targeted event, we collected tweets two days before the event, the actual day it happened, and two days after the event. We used the GelOldTweets[8] API to collect past tweets. Table I shows the search terms and boundary dates used as parameters for the collection of tweets, the total number of pre-event tweets (BEFORE), and the total number of post-event tweets (AFTER - both). The latter is also detailed in terms of the number of tweets collected using the event hashtags (AFTER - #event) and using the city names (AFTER - city).

Data pre-processing involved traditional steps, such as the removal of hyperlinks, hashtags used to collect the tweets, mentions to users, special marks and symbols (&, , ;̂ _, etc). In addition, we applied an English dictionary[9] to filter out tweets with too many misspelled words and non-English ones.

These actions resulted in four datasets, as shown in Table I. The column BEFORE summarizes the number of pre-event tweets: *BM* (before Manchester) containing 5,351 tweets, *BL* (before London) containing 20,379 tweets. The column AFTER indicates the total number of post-event tweets: *AM* (after Manchester) containing 929,518 tweets, and *AL* (after London) containing 392,711 tweets. This column represents the total number of tweets collected using both the event hashtag and the name of the city, and the value is the sum of the numbers in the last two columns. The AM/AL datasets are subdivided into *AM/AL-event* and *AM/AL-city*, representing the tweets collected using, respectively, event hashtags only and city names only. The structure of all these datasets is identical and include, among others, the filtered tweet text and the tweet ID.

4.2.2 *Demographics and Location Extraction.* We aimed to perform our analysis using the location of the tweet, together with the gender and age of the Twitter user. To that purpose, for each collected tweet, we searched for the respective user profile, from which we extracted such information, using the Twython API[10]. This API provides the method *“show_status”* that, given a tweet ID as input, returns the complete tweet structure in *JSON* format.

The whole tweet structure includes a sub-structure that provides information on the user profile. However, the profile of a Twitter user contains neither the age nor gender. Therefore, we extracted the user profile image and used the Face++[11] tool to obtain this information, as in [ElSherief et al. 2017; Walter and Becker 2018]. Face++ takes as input images and outputs an estimation of face attributes, including age and gender. Experiments performed using this tool [Fan et al. 2014] report an accuracy of 85%.

With regard to location, we observed that less than 1% of the collected tweets were geo-referenced. Following the strategy proposed in [Sakaki et al. 2010], we adopted the location as informed in the user's profile. Our original plan was to analyze the sentiment in the local community, i.e. the city where the event happened, against the sentiment in other locations. However, the number of tweets for comparison in that granularity was very small; thus, we abstracted to the respective country. To

[8] https://github.com/Jefferson-Henrique/GetOldTweets-python

[9] https://github.com/dwyl/english-words

[10] https://twython.readthedocs.io/en/latest/

[11] https://www.faceplusplus.com/

Table II. Gold Standard: Number of labelled tweets per category

| **Emotion** | **Anger** | **Disgust** | Fear | **Sadness** | **Surprise** | **None** |
|---|---|---|---|---|---|---|
| # tweets | 82 | 116 | 85 | 179 | 71 | 74 |

this end, we compared each declared location against a list of cities in UK[12], and United States of America (USA)[13]. Locations not matching any city in these lists were classified as "other location". Considering the number of resulting tweets in each category, we performed our analysis using only tweets related to the UK and the USA.

### 4.3 Gold Standard

Our work focuses on five out of the six basic emotion categories defined by Ekman [Ekman and Friesen 1982]. We focused on negative emotions only, because we assume people are not likely to express positive emotions (such as happiness) in reaction to terrorist events[14]. In addition, we considered the emotion surprise, which is not necessarily negative. According to Ekman, "the function of surprise is to focus our attention so we can determine what is happening and whether we are in danger or not". Thus, in the terrorism context, we assumed people could be surprised by the event in a negative way. In summary, the emotion categories considered are *anger*, *fear*, *sadness*, *surprise* and *disgust.* Our approach considers that a given tweet is related to one and only one emotion category, which is the prevalent emotion.

We created a terrorism gold standard to analyze the performance of the emotion classifiers. Tweets were labeled according to each emotion category considered, plus an extra *None* category. This was accomplished using Amazon Mechanical Turk[15].

First, one of the authors annotated 967 tweets with the considered 5 emotion labels, based on the presence of emotion keywords and expressions. For example, the tweet "*Deeply saddened by the loss of 22 beautiful lives. we should not live like this. They did not deserve to die*" was labeled as sadness due to the expression "deeply saddened"; the tweet "*It's so scary to not feel safe in this World*" was labeled as fear due to the expression "It's so scary", and so on. We started from a randomly selected set of tweets, discarding the ones that did not contain an unquestionable emotion word/expressions, or labeling it otherwise. This task was performed until we reached a minimum of 100 tweets per emotion. This procedure resulted in relatively well-balanced sets. Afterward, we created a HIT (Human Intelligence Task) with these tweets, where annotators were asked to determine which emotion best described a tweet, given a set of categories as options (Anger, Fear, Disgust, Sadness, Surprise, None). We instructed annotators to choose the primary emotion if more than one emotion could be identified, and to choose *None* if no emotion could be clearly determined. We targeted the HIT to two master annotators so that we would have three annotators in total, considering one of the authors. According to Amazon, master annotators typically have a 90% or more accuracy rate. We filtered out tweets in which there was a disagreement between all the three annotators and retained those with at least two agreements. The results, composed of 607 tweets, are displayed in Table II, which we consider as our ground truth for validating the emotion prediction model.

### 4.4 Automatic Generation of Training Seeds

There is no publicly available corpus for the domain addressed in this work, and manual annotation is a costly and error-prone activity. Therefore, we tried different approaches to gather enough seeds for training emotion classifiers. All approaches are based on resources (datasets, lexicons) developed

[12] https://www.paulstenning.com/uk-towns-and-counties-list/

[13] https://github.com/grammakov/USA-cities-and-states

[14] https://www.paulekman.com/blog/our-emotional-reactions-terrorism/

[15] https://www.mturk.com/

Table III. Number of training instances per emotion and strategy

| Strategy | Anger | Disgust | Fear | Sadness | Surprise | Total |
|---|---|---|---|---|---|---|
| Distant Supervision | 733 | 818 | 113 | 501 | 402 | 2567 |
| Emotion Hashtags | 9 | 6 | 47 | 87 | 1 | 145 |
| Dictionary-based | 8665 | 6508 | 7773 | 8866 | 7678 | 39490 |
| Emotion Keywords | 672 | 300 | 854 | 2006 | 186 | 4018 |

Table IV. Emotion keywords used for filtering training seeds

| Emotion | Keywords |
|---|---|
| **anger** | anger, fuck, fucked, pissed, lmaof, damm |
| **disgust** | disgust, disgusted, disgusting |
| **fear** | worried, worry, scary, scaring, scared, fear |
| **sadness** | sad, sadness, saddened |
| **surprise** | surprised, surprising, surprise, shocked, shocking |

according to Ekman's emotion model. The number of training instances resulting from each strategy is summarized in Table III.

—*Distant supervision*: we trained an emotion classifier using a labeled dataset related to another domain, and then applied the resulting model to predict the label of instances in our terrorism dataset. This strategy was deployed in works such as [Purver and Battersby 2012; Suttles and Ide 2013]. We adopted a manually labeled dataset of tweets related to electoral debates [Mohammad et al. 2015].

—*Emotion hashtags*: tweets were automatically labeled according to the presence of emotion hashtags (e.g. #anger, #inlove), as in [Wang et al. 2012; Mohammad 2012]. We used the emotion hashtags collected in [Mohammad 2012].

—*Emotion lexicon entries*: we used the NRC lexicon [Mohammad and Turney 2013b], in which each entry is associated with one or more emotion labels. Tweets were automatically labeled according to the presence of emotion words contained in NRC (e.g. for anger, entries such as "mad" or "angry"). Only tweets with a single prevailing emotion were included in the training dataset. This strategy was previously adopted in [Mirani and Sasi 2016; Azizan and Aziz 2017].

—*Emotion keywords*: we assigned a label according to the presence of keywords contained in a list of keywords representative of the emotions in our target events. To define this list, we manually inspected random sets of tweets for each event and identified an initial set of keywords representative of each emotion in the context of terrorism-related tweets. Then, we sampled tweets containing such keywords and assessed whether they were likely to belong to their respective emotion categories. In this process, we refined the set of keywords. The final set of keywords is shown in Table IV.

In all of our experiments, training seeds for the *None* category were chosen by selecting tweets that did not contain any of the following terms: a) defined keywords (Table IV), b) emotion hashtags defined in [Mohammad 2012], and c) emotion entries in the NRC lexicon. The number of tweets labeled with *None* category was much higher in comparison with the number of tweets labeled with the other categories. To balance the volume of instances labeled with *None* with regard to the other classes, we randomly selected a subset of tweets of which the size is the averaged number of seeds for the other five emotion categories.

### 4.5 Emotion Classification

In this work, we compare two alternative deep learning architectures to build emotion classifiers: CNN and LSTM. The former was adopted in our original work [Harb and Becker 2018], and it corresponds to the CNN architecture defined in [Kim 2014]. This choice was motivated by the popularity of this architecture for sentiment classification in short texts, which is the case of tweets. In this article,

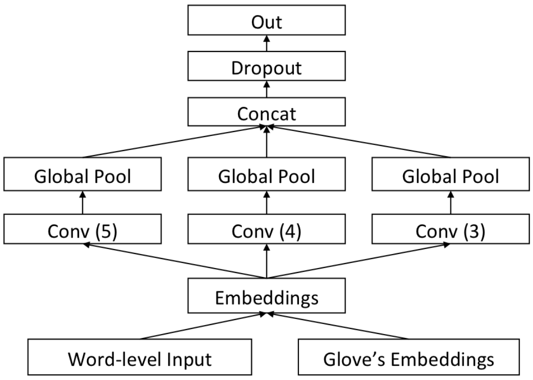

Fig. 2. Convolutional Neural Network Architecture.

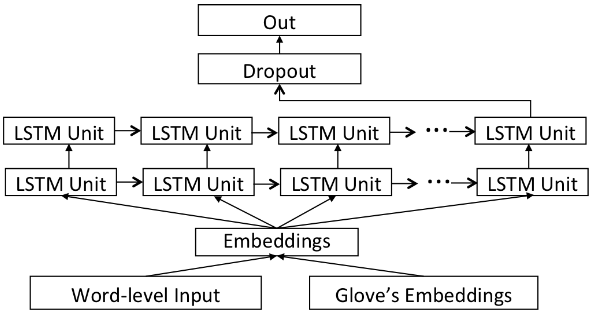

Fig. 3. Long Short Term Memory Architecture.

we also experiment with an LSTM architecture, which is the prevalent deep neural architecture for sentiment classification [Zhang et al. 2018]. Our models were implemented in Python and executed on top of TensorFlow[16]. The code corresponding to each emotion classifier implementation is publicly available[17].

The CNN architecture is depicted in Figure 2, and the LSTM architecture is displayed in Figure 3. Both architectures rely on two inputs: a word-level training set that was automatically generated according to one of the seed generation strategies described in Section 4.4, and a set of pre-trained word embeddings. We used the GloVe[18] embeddings set trained using a huge corpus of tweets, represented using 100 dimensions. We assume a tweet with a maximum of 70 words (size of the longest tweet) and zero-pad shorter inputs.

As depicted in Figure 2, the CNN architecture has four main layers. The first one corresponds to the non-static model of embeddings, i.e. in which the input words are mapped into pre-trained embeddings and then the input is fine-tuned by the network. The second layer is responsible for applying convolutions with multiple filter sizes over the word embeddings. For each filter size (3, 4, and 5), a convolution is performed, and a feature map describing the sentences is generated. The third layer is responsible for filtering the most important features into one feature vector through a max polling operation. Finally, the fourth layer is responsible for regularizing and normalizing the CNN, respectively. Regularization is achieved with the *dropout* function, which basically avoids overfitting and forces the network to learn features that are individually useful. The normalization is performed with Softmax, which normalizes the outputs into probability distributions over the predicted output

[16] https://www.tensorflow.org/
[17] https://github.com/regisebeling/EmotionsClassification
[18] https://nlp.stanford.edu/projects/glove/

classes. This implementation was an adaptation of a publicly available code [19] for Kim's CNN.

The implementation of the LSTM architecture, shown in Figure 3, assumes the same non-static model for mapping the input set into pre-trained embeddings, and fine-tunning them through the network. This input is processed by multiple layers of LSTM units. The model's task is to classify the text of a tweet into an emotion, so the architecture is a many-to-one application, where the prediction is only performed after reading the entire input sequence. The inputs for the first LSTM layer are the embeddings corresponding to the input words, which are mapped into pre-trainned embeddings. The next two layers are composed of LSTM units, in which the number of units was defined according to the tweet size. The outputs of the first LSTM layer are used as inputs to the respective units in the next LSTM layer, which in turn generate a single output: the emotion prediction. The prediction of the model is generated by the value of the highest probability among the predicted output classes. The last layer is responsible for regularization (dropout) to avoid model overfitting, as in the CNN architecture.

## 5. EXPERIMENTS

We performed experiments to determine the best seeding strategy, as well as to compare the performance of the CNN and LSTM models. For the CNN, we adopted the same hyperparameters defined in [Kim 2014], using the non-static model. We also experimented with distinct region sizes and pooling layers, but no improvements were observed. For the LSTM, we adopted two LSTM layers composed of 128 units each. We experimented with more LSTM layers, as well as variations on the number of units, which did not improve the results described below. For both models, dropout was set as 0.2. We experimented with different numbers of epochs for training the models, and the results reported here refer to 200 epochs. All results reported in this section refer to the average of 10 executions, where the model was trained using the same input and tested against the Gold Standard (Section 4.3). Table II displays the number of training instances for each seeding strategy. In the remaining of this section, we discuss the results of our experiments.

### 5.1 Evaluation of Seeding Strategies

Tables V and VI display the performance of the CNN and LSTM models according to each seeding strategy in terms of macro-averaged precision, recall, and f-measure metrics. The performance of each seeding strategy is very similar for both model architectures. The dataset composed of tweets filtered using the list of selected emotion keywords (Table IV), which are specific to this domain, significantly outperformed all other strategies (averaged *f-measure* = 63% for both models). The worst performance was observed for the emotion hashtags strategy (averaged f-measure around 17-19%), followed by the emotion lexicon and distance supervision strategies, which are quite similar. These results confirm our previous findings on the seeding strategies [Harb and Becker 2018], and reveal they are independent of the deep learning architecture.

The worst result (Emotion Hashtag strategy) is clearly related to the smallest number of training instances (145), but size alone does not explain these results. Indeed, the metrics for the models trained using the *Electoral Debate* dataset (2,567 instances) and the *Sentiment Lexicon-based* dataset (39,490 instances) are not that different (f-measure ranging between 24 and 32%). The former was annotated for a completely distinct domain, and the latter considers the generic usage of the emotion words.

We also analyzed the existence of common instances between the *Emotion Keywords* dataset and the ones generated using the other strategies. We observed only three instances in common with the dataset labeled using Emotion Hashtags. Compared to the *Electoral Debate* dataset, we found that

[19] https://github.com/cahya-wirawan/cnn-text-classification-tf

Table V. Results for the generated CNN prediction models

| Approach | Avg. Precision | Avg. Recall | Avg. F-measure |
|---|---|---|---|
| Distant Supervision | 0,3633 | 0,3833 | 0,3216 |
| Emotion Hashtags | 0,1433 | 0,272 | 0,173 |
| Emotion Lexicon | 0,2533 | 0,3233 | 0,245 |
| Emotion Keywords | **0,7533** | **0,6616** | **0,6316** |

Table VI. Results for the generated LSTM prediction models

| Approach | Avg. Precision | Avg. Recall | Avg. F-measure |
|---|---|---|---|
| Distant Supervision | 0,3027 | 0,3454 | 0,2912 |
| Emotion Hashtags | 0,2013 | 0,2398 | 0,1993 |
| Emotion Lexicon | 0,2722 | 0,3413 | 0,291 |
| Emotion Keywords | **0,7312** | **0,654** | **0,6350** |

Table VII. F-measure per emotion for the models generated using emotion keywords

| F-Measure | Anger | Disgust | Fear | Sadness | Surprise | None |
|---|---|---|---|---|---|---|
| CNN | 0,845 | 0,5162 | 0,6388 | 0,7283 | 0,5327 | 0,5286 |
| LSTM | 0,834 | 0,521 | 0,568 | 0,7352 | 0,5057 | 0,6461 |

8.16% of the *Emotion Keywords* dataset correspond to common instances. The greatest intersection was observed for the *Sentiment Lexicon-based* dataset (1402 common instances, representing 34.89% of the *Emotion Keyword* dataset). However, the Emotion Keywords dataset represents only 3.55% of the *Sentiment Lexicon-based* dataset. Thus, we conclude that the excess of terms that refer to emotions represent information that is out of context, therefore yielding a much inferior result. Based on these results, we adopt the dataset labeled using the Emotion Keywords strategy.

### 5.2 Evaluation of Emotion Classification Models

To compare the performance of the LSTM and CNN models trained using this dataset, we run a two-tailed paired t-test with a significance level of 0.05, and we did not find a significant difference between the two models ($p\text{-}value = 0.46$ for macro-averaged f-measure and recall, and 0.14 for macro-averaged precision).

Table VII details the f-measure scores per emotion. The main improvement of the LSTM model, compared to the CNN one, was in the classification of tweets with no (negative) emotion (class *None*), with an improvement of 12 percentage points (pp) in the f-measure. On the other hand, the LSTM performs poorly on the classification of the class *Fear* (decrease of 8 pp), which is a very important emotion in this context [Harb and Becker 2018]. A two-tailed paired t-test with a significance level of 0.05 was executed and pointed out that there are no significant differences for all other emotions (p-value= 0.14 for anger, 0.34 for disgust, 0.25 for sadness and 0.11 for surprise). A closer look reveals that the CNN performs better on the precision metric, with differences that range from 6.5 pp (*Fear*) to 14 pp (*Sadness*). Regarding precision, the LSTM model only outperformed the CNN model for the *None* class (14 pp). On the other hand, the LSTM performed better on the recall metric, particularly for *Sadness* (16 pp) and *Anger* (6 pp). Regarding recall, the CNN model presented a significantly superior performance only for the *None* class (difference of 5 pp).

In analyzing the confusion tables, we noticed some recurrent prediction issues, some of them illustrated in Table VIII. The prevalent error in both models is to predict tweets as belonging to the class *None* (absence of negative emotions) when actually they do evoke an emotion according to the Gold Standard. This problem happens particularly for tweets annotated with labels *Sadness*, *Fear* and *Surprise*, and it is more noticeable in the CNN model (*precision* = 28%, as compared to 42% for the LSTM model). The LSTM model tends to wrongly predict the label *Sadness* for some tweets,

Table VIII. Examples of wrong predictions.

| Gold Standard | CNN | LSTM | Tweet Text |
|---|---|---|---|
| FEAR | NONE | NONE | this is terrible but nowhere is safe at the moment |
| FEAR | NONE | ANGER | goodnight twitter hoping that everyone is safe |
| FEAR | ANGER | NONE | my nerves and emotions have been on overdrive all day why is such a beautiful planet such a dangerous world? |
| SADNESS | FEAR | NONE | i'm honestly heartbroken i feel like no where and nothing is safe anymore what is wrong with people? |
| SADNESS | NONE | FEAR | my heart breaks every second i scroll through my feed |
| SADNESS | NONE | DISGUST | i do n't wanna cry anymore ! it 's not right innocent people die in this atrocious way |
| DISGUST | FEAR | FEAR | people who can complain worry about their flights being threatened at the expense of manchester is actually pathetic |
| DISGUST | DISGUST | SADNESS | coming across jokes such as this on my feed how low can people get xx |
| ANGER | FEAR | FEAR | i can't even think about how ari feels rn im so done bout all that shit i ca n't stop crying i hate this world |
| ANGER | FEAR | ANGER | living in fear is the worst but do n't let the bastards win that 's how they want us to feel ! we are not scared of you |
| ANGER | DISGUST | ANGER | honestly do n't understand this messed up world we live in with absolutely messed up people it 's disgusting revolting |

particularly when they are annotated with *Surprise*, *Fear* and *Disgust*. This partially explains the low performance on the precision metric (71%, in comparison to 85% for the CNN model, where this problem is concentrated in tweets annotated with *Disgust*). Another error observed in both models is to incorrectly predict tweets annotated with *Disgust* as *Sadness* or *Fear*, an issue that is more serious in the CNN model (*recall* = 27%, as compared to 32% for the LSTM model). Both models make false predictions for tweets annotated with *Anger*, assigning them to the *Fear* class, an issue that is more significant in the CNN model (*recall* = 82%, in comparison to 87% for the LSTM model). Finally, we noticed that the CNN model is more sensitive to ambiguity when words characterizing distinct emotions are present in the same sentence (e.g. last two tweets of Table VIII, which mix words of anger and fear/disgust). This is probably due to the convolutional strategy of sliding windows for finding local patterns.

In conclusion, we consider both models reliable because their average precision and recall are approximately 70%, which are good results taking into account emotion classification results presented in well-known works such as [Suttles and Ide 2013; Purver and Battersby 2012; Mohammad et al. 2015]. Our original work has shown that sadness, anger, and fear are the three prevalent emotions in this context. Considering the specific precision and recall scores for these three emotions, we decided to adopt the CNN based model for the analysis developed in the next section.

## 6. ANALYSIS

### 6.1 Q1: Is there an emotion shift due to terrorist events?

To answer this question, we compared the distribution of emotions before the events (BM and BL), and after them (AM and AL, respectively). Figure 4 depicts this comparison, where the Y-axis represents the percentage with regard to the total number of tweets of the respective dataset. To deal with the fact that the post-event datasets collected using city names were much bigger then the ones gathered using the attack hashtags (e.g. 904,508 against 25,010 in the case of Manchester - Table I), we calculated these proportions for each type dataset, and then averaged them. For both events, we can notice that the number of tweets with negative emotions significantly increased after the terrorist events: 18 percentage points (pp) for the Manchester attack and 13 pp for the London Bridge event.

Figure 5 details these graphs per emotion. Three emotions were particularly evoked, namely anger, fear, and sadness. Disgust and surprise are not representative in the datasets, nor are the differences due to the events. When we consider each specific event, the Manchester attack evoked more fear and sadness, whereas tweeters expressed more anger regarding the London Bridge event. It is interesting to notice that the amount of anger for Manchester did not change significantly (0.8 pp increase). We examined the tweets to understand the reason, and we noticed that tweeters were much concerned about the participation of a soccer team (Manchester United) in the Final of the UEFA Europa League, which was to happen at the weekend following the concert attack. Examples of tweets expressing anger before the event are "*Fucking hell, turn it in on the do it for manchester shouts, 90% of the city fucking hate Man United*"or "*I know I should be supporting everything Manchester, but watching this makes me realize how much I hate Man U*". A possible explanation is that the population might have reduced

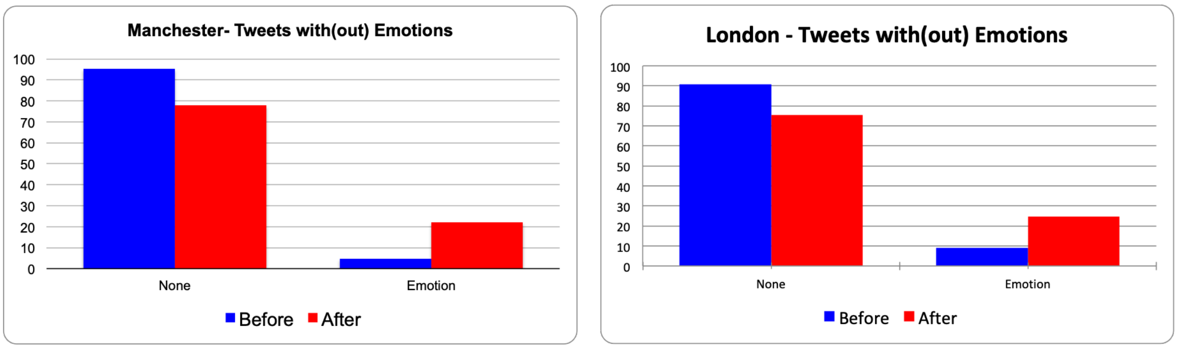

Fig. 4. Tweets distribution before and after the attacks.

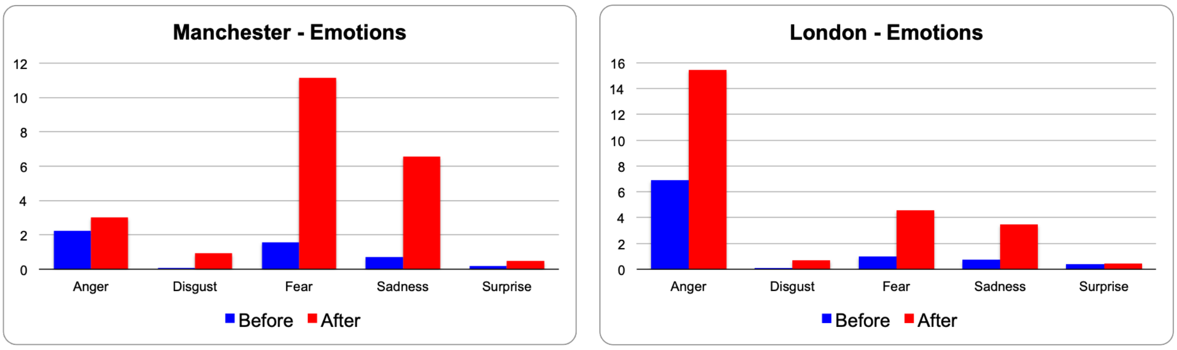

Fig. 5. Tweets distribution before and after the attacks.

this type of angry tweets due to the attack at Ariana Grande's concert, expressing fear and sadness instead, otherwise the number of angry tweets might tend to increase with the proximity of the match.

We also performed a comparison using different configurations of the post-event datasets: #event only (as in our original work [Harb and Becker 2018]) and city name only. In all cases, the trends were alike, but in different scales. For instance, if we consider the prevalent emotion of each event, the use of tweets collected using *#prayformanchester* and *#londonbridge* resulted in an increase of 18 pp in fear and 14 pp in anger, respectively. When these same emotions are compared using the datasets collected using the keywords *manchester* and *london*, the corresponding differences are 3 and 4 pp, respectively.

In conclusion, we can observe that there is indeed an emotional shift due to terrorist events. Significant changes were observed for three emotions, namely fear, sadness, and anger. We also noticed differences in the emotions expressed by tweeters depending on the specific event.

### 6.2 Q2: Do different terrorist events evoke the same emotional reaction?

To answer this question, we only used the post-event datasets collected using the event hashtags (AM#prayForManchester/AL#londonBridge). Figure 6 depicts the emotion distribution for both events, where the Y-axis represents the percentage with regard to the total number of tweets per event. Only tweets with emotions are considered. As already mentioned, the results reveal that there are differences between these two events. By far, Anger was the prevalent emotion in the London Bridge attack, whereas the Manchester event predominantly evoked fear, followed by sadness.

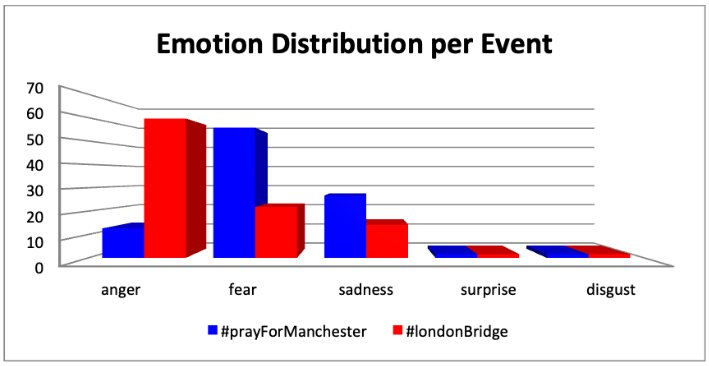

Fig. 6. Emotion distribution per terrorist event.

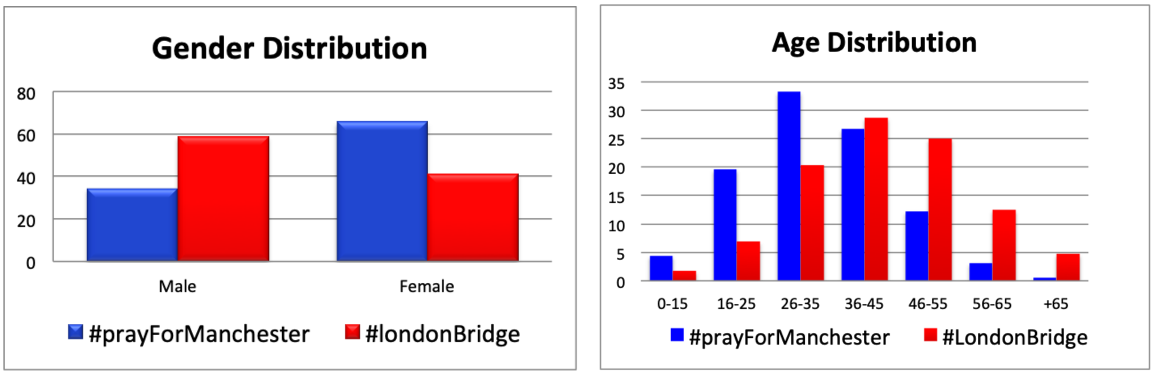

Fig. 7. Gender and Age Distribution per Event.

A stratified analysis per demographics helps us understanding the differences between these two events. Figure 7 shows gender and age distributions for both events. The Y-axis represents, for each class (Gender and/or Age), the percentage with regard to the total number of tweets with emotions. These distributions show that the users who expressed emotions regarding these events are quite different. Whereas for the Manchester event the majority of tweeters are female (65%), for the London event, they are predominantly male (60%). In addition, Manchester tweeters are also younger (the average age is 33 for the Manchester event, as compared to 42 for the London attack). This can be explained by the fact that Ariana Grande is very popular in this demographics. On the other hand, we believe that the London Bridge attack has affected the average citizen who could be potentially at the location of the attack. Thus, we can hypothesize that these differences are related to gender and/or age.

Figure 8 details for each event the emotions per gender. It is possible to see that the male gender is prevalent among those who express anger, whereas among those who express fear and sadness, the female gender is more frequent. Figure 9 displays the distribution of emotions per age. It shows that, as age increases, the feeling of anger increases. Fear, on the other hand, is a frequently expressed emotion for younger ages, and the percentage smoothly drops as age increases. A decline in sadness is observed for older users (45 years or more). Nevertheless, compared to gender, age does not seem to be a determinant factor that explains the differences in terms of emotional expression, since each event is characterized by a prevalent emotion, and the prevalence is consistent across all age ranges.

Thus, we conclude that each terrorist event may raise distinct predominant emotions. In the events analyzed, gender was the influential factor that explains the difference, where fear and sadness can often be more related to the female gender, and anger to the male gender.

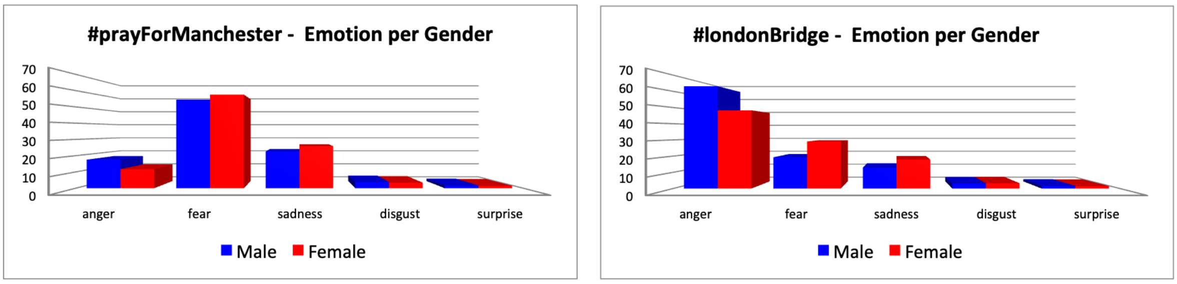

Fig. 8. Emotion Distribution per Gender.

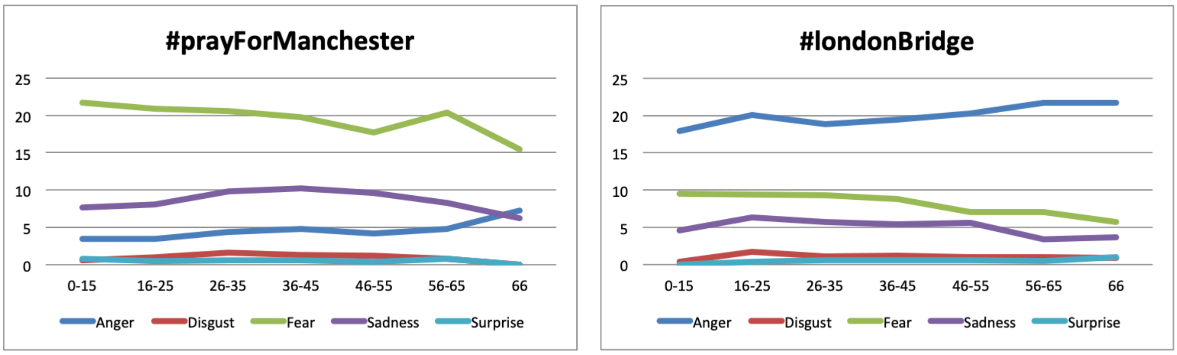

Fig. 9. Distribution of Emotions per Age.

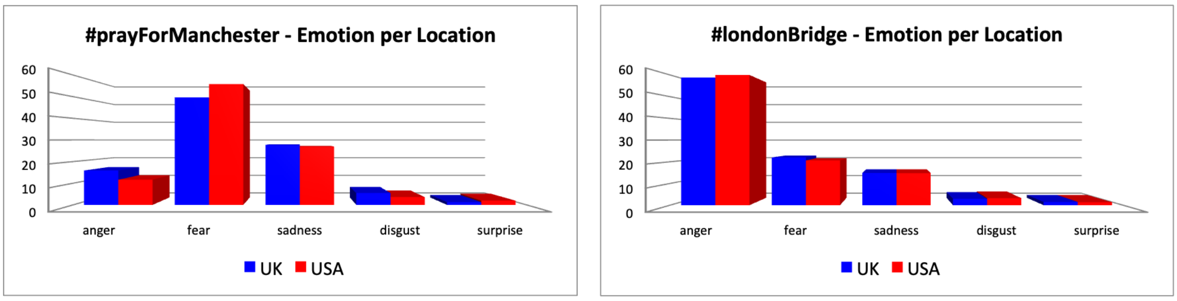

Fig. 10. Emotion distribution by Location.

### 6.3 Q3: Does the proximity to the event influence the emotional reaction?

To answer this question, we compared tweets from the UK and the USA. The distributions for each country are depicted in Figure 10, where the Y-axis represents, for each class (UK and USA), the percentage of its total number of tweets distributed in emotion categories. For both locations, the distribution of tweets into emotion categories for both events did not show any noticeable variation. These findings indicate that location may not be an important factor as much as age and gender are.

### 6.4 Threads of Validity

We conclude this section by discussing the threats that influence the validity of our analyses. One of the main threats is related to our reliance on automatic tools to infer the gender and the sex of Twitter users from their profile picture. Experiments on Face++ report high precision results, but users may adopt pictures that are not representative of their looks. We manually analyzed 100 pictures, and confirmed the accuracy of results produces. Face++ could not infer the age/gender of near 42% of the users, which were not included in the respective gender/age analysis. Another threat is the inference of location from users profiles, as the user may not have tweeted from the same location as declared in the profile. We used a widely deployed technique originally proposed in [Sakaki et al. 2010; Jones et al. 2019], and sanitized the self-declared locations using lists of cities in the UK and the USA to produce more reliable results. The adoption of these two techniques for demographics inference enabled us to generalize our findings by considering a bigger sample.

Another threat is related to the hypotheses underlying the collection of the pre/post-event tweets. The adoption of city names encompasses people concerned with all sort of events that might happen in a community, from entertainment to city maintenance issues (e.g. traffic, public services), and thus it is representative at some extent to the general state of mind of the citizens of that community [Jones et al. 2019]. It is also representative of outsiders who might express interest in and/or visiting these communities (e.g. tourism, entertainment). On the other hand, the inclusion of hashtags representing events of impact have been a behavior widely observed, and has been widely deployed in all works discussed in Section 3. Keeping in mind the pros and cons of these strategies, this methodological flexibility have enabled insights on how communities are impacted by a variety of collective traumas [Jones et al. 2016; Jones et al. 2019].

Finally, it should should be acknowledged that Twitter users are not representative of the general population. Nevertheless, our analysis have shown a diverse demographics of users for this particular events. The public forum Twitter provides for its users is also amenable to more sophisticated analyses of the ways in which the platform itself may be responsible for emotion contagion in a collective-trauma context, as pointed out by Jones et al. (2019).

## 7. CONCLUSIONS

In this work, we developed an analysis of the emotional reactions of Twitter users to two terrorist events that occurred in the UK. Demographic data of users, namely location, age, and gender, were extracted from users' profiles using automatic tools. We focused on identifying negative emotional reactions according to the Ekman's model, and developed emotion classifiers based on two deep learning architectures. The results of our analysis reveal that when terrorist events occur, a shift of emotion towards anger, sadness, and fear can be noticed. The prevalent emotions are specific to the event, according to the affected users. A detailed analysis based on age and gender revealed that the latter is the most influential factor in how users emotionally react to the event. Our data indicate that fear and sadness is a sentiment more prevalent among women, whereas anger is more prevalent among men. The proximity to the event did not provide any noticeable impact on the emotional reaction.

We compared two deep learning architectures to develop an emotion model, namely CNN and LSTM. The resulting models are not significantly distinct, with average f-measure of 63%. We performed a detailed analysis of the prediction results and noticed that the CNN model was stronger in terms of precision, whereas the LSTM model was better regarding the recall. An analysis of the prediction errors was performed. We also confirmed that the seeding strategy of filtering by emotion keywords was by far the best one, independently of the emotion classifier.

As a contribution, we created an emotion dataset and Gold Standard in the context of terrorism. The code of the developed classification models is also publicly available. The research questions answered in this article were a first step towards understanding the emotional reaction terrorist events raise on the general population. We applied the proposed method in other terrorism events [Harb and Becker 2019], as well as on data related to mass shooting incidents [Harb 2019].

Future work includes: a) further experiments with deep learning architectures to improve classification results; b) the extension of the emotion classification problem to a multilingual environment, thus reaching a larger number of events, countries and population; b) improvement of the techniques to extract the demographics from tweeters profiles; c) to employ topic analysis on tweets to further investigate the contents exchanged; among others.

## REFERENCES

Azizan, S. A. and Aziz, I. A. Terrorism Detection Based on Sentiment Analysis Using Machine Learning. *Journal of Engineering and Applied Sciences* 12 (3): 691–698, 2017.

Baccianella, S., Esuli, A., and Sebastiani, F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of the Intl. Conf. on Language Resources and Evaluation (LREC)*. Vol. 10. pp. 2200–2204, 2010.

Burnap, P., Williams, M. L., Sloan, L., and et alli. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining* 4 (1): 206, Jun, 2014.

Chong, M. Sentiment analysis and topic extraction of the twitter network of #prayforparis. *Proc. of the Association for Information Science and Technology* 53 (1): 1–4, 2016.

Cohen-Louck, K. and Ben-David, S. Coping with terrorism: Coping types and effectiveness. *International Journal of Stress Management* 24 (1): 1–17, 2017.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* vol. 12, pp. 2493–2537, 2011.

Crepeau-Hobson, F., Sievering, K. S., Armstrong, C., and Stonis, J. A coordinated mental health crisis response: Lessons learned from three colorado school shootings. *Journal of School Violence* 11 (3): 207–225, 2012.

Ekman, P. and Friesen, W. Emotion in the human face system. *Cambridge University Press, San Francisco, CA,*, 1982.

ElSherief, M., Belding, E. M., and Nguyen, D. # notokay: Understanding gender-based violence in social media. In *Proc. of the 11th Intl. Conference on Web and Social Media (ICWSM)*. pp. 52–61, 2017.

Fan, H., Cao, Z., Jiang, Y., Yin, Q., and Doudou, C. Learning deep face representation. *CoRR*, 2014.

Garg, P., Garg, H., and Ranga, V. Sentiment analysis of the Uri terror attack using twitter. In *Proc. of the Intl. Conf. on Computing, Communication and Automation (ICCCA)*. pp. 17–20, 2017.

Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision, 2009.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems* 28 (10): 2222–2232, Oct, 2017.

Harb, J. G. D. *Using a convolutional neural network to compare emotional reactions on Twitter to mass violent events*. M.S. thesis, Instituto de Informática, UFRGS. Porto Alegre - Brazil, 2019.

Harb, J. G. D. and Becker, K. Emotion analysis of reaction to terrorism on twitter. In *Proc. of the SBC Brazilian Symposium on Databases*. pp. 97–108, 2018.

Harb, J. G. D. and Becker, K. Comparing emotional reactions to terrorism events on twitter. In *Big Social Data and Urban Computing*, J. Oliveira, C. M. Farias, E. Pacitti, and G. Fortino (Eds.). Vol. 926. Springer, 7, 2019. *(To appear.)*.

Horgan, J. *The Psychology of Terrorism*. Taylor & Francis Group, 2014.

Jones, N. M., Brymer, M., and Silver, R. C. Using big data to study the impact of mass violence: Opportunities for the traumatic stress field. *Journal of Traumatic Stress* 32 (5): 653–663, 2019.

Jones, N. M., Wojcik, S. P., Sweeting, J., and Silver, R. C. Tweeting negative emotion: An investigation of twitter data in the aftermath of violence on college campuses. *Psychological Methods* vol. 21, pp. 526–541, 2016.

Kim, Y. Convolutional neural networks for sentence classification. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) - ACL*. pp. 1746–1751, 2014.

Liu, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5 (1): 1–167, 2012.

Liu, B., Hu, M., and Cheng, J. Opinion observer: Analyzing and comparing opinions on the web. In *Proc. of the 14th Intl. Conf. on World Wide Web (WWW)*. pp. 342–351, 2005.

Lowe, S. R., Blachman-Forshay, J., and Koenen, K. C. Trauma as a public health issue: Epidemiology of trauma and trauma-related disorders. In *Evidence Based Treatments for Trauma-Related Psychological Disorders: A Practical Guide for Clinicians*, U. Schnyder and M. Cloitre (Eds.). Springer, pp. 11–40, 2015.

Maguen, S., Papa, A., and Litz, B. T. Coping with the threat of terrorism: A review. *Anxiety, Stress, & Coping* 21 (1): 15–35, 2008.

Mansour, S. Social Media Analysis of User's Responses to Terrorism Using Sentiment Analysis and Text Mining. *Procedia Computer Science* vol. 140, pp. 95–103, 2018.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* pp. 3111–3119, 2013.

Mirani, T. B. and Sasi, S. Sentiment analysis of isis related tweets using absolute location. In *Proc. of the 2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. pp. 1140–1145, 2016.

Mohammad, S. #Emotional Tweets. In *Proc. of the 1rst Joint Conf. on Lexical and Computational Semantics.* pp. 246–255, 2012.

Mohammad, S., Kiritchenko, S., and Zhu, X. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proc. of the 7th Intl. Workshop on Semantic Evaluation (SEMEVAL).* pp. 321–327, 2013.

Mohammad, S. M. and Turney, P. D. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29 (3): 436–465, 2013a.

Mohammad, S. M. and Turney, P. D. Crowdsourcing a word-emotion association lexicon. 29 (3): 436–465, 2013b.

Mohammad, S. M., Zhu, X., Kiritchenko, S., and Martin, J. Sentiment, emotion, purpose, and style in electoral tweets. 51 (4): 480–499, 2015.

Munezero, M. D., Montero, C. S., Sutinen, E., and Pajunen, J. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing* 5 (2): 101–111, 2014.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

Purver, M. and Battersby, S. Experimenting with Distant Supervision for Emotion Classification. *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012.

Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proc. of the 19th Intl. Conf. on World Wide Web (WWW).* pp. 851–860, 2010.

Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web.* WWW '14 Companion. ACM, New York, NY, USA, pp. 373–374, 2014.

Simon, T., Goldberg, A., Aharonson-Daniel, L., Leykin, D., and Adini, B. Twitter in the cross fire—the use of social media in the westgate mall terror attack in kenya. *PloS One* vol. 9, Aug, 2014.

Suttles, J. and Ide, N. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 121–136, 2013.

Walter, R. and Becker, K. Caracterização e comparação das campanhas do outubro rosa e novembro azul no twitter. In *Proc. of the SBC Brazilian Symposium on Databases.* pp. 133–144, 2018.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. Harnessing twitter 'big data' for automatic emotion identification. In *Proc. of the 2012 ASE/IEEE International Conference on Social Computing.* pp. 587–592, 2012.

Zhang, L., Wang, S., and Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 8 (4), 2018.

Zimbra, D., Abbasi, A., Zeng, D., and Chen, H. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Trans. Management Inf. Syst.* 9 (2): 5:1–5:29, 2018.