Measuring and Understanding Overall News Media Sympathy on Twitter in the Aftermath of Crisis Events

1st Author Name
Affiliation
City, Country
e-mail address

2nd Author Name
Affiliation
City, Country
e-mail address

3rd Author Name
Affiliation
City, Country
e-mail address

ABSTRACT

Did Western media on Twitter exhibit a bias in coverage of the November 2015 Beirut and Paris attacks? Drawing on two Twitter datasets covering each attack, we use text-mining and crowdsourcing to investigate how Western and Arab media differed in coverage bias, sympathy bias, and resulting information propagation. By crowdsourcing labels across four languages (English, Arabic, French, German), we derived the Overall News Sympathy (ONS) score, a measure that factors in religious reference. We found both attacks were disproportionately covered, that Western media was overall less sympathetic when covering the Beirut attacks, and that sympathetic tweets did not spread further. We further trained a deep neural network to predict ONS scores from unlabeled data, and found each Twitter media (Western, Arab) to be more sympathetic to attacks in their respective regions (Paris and Beirut, respectively). We discuss our results in light of global news flow and their public perception impact.

Author Keywords

Twitter, sympathy, news media bias, crisis informatics, cross-cultural, crowdsourcing, NLP

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Web-based interaction

1. INTRODUCTION

During the days after the 13 November, 2015 attacks in Paris, people took social media by storm in response to the events. From outcries of sympathy and solidarity with Paris, to outcries against or support for Islam [23], to proclamations that mainstream coverage of the Beirut attacks, which happened a day earlier, have been sparse and uncaring. In such press reports, we find allegations suggesting reports about attacks, bombings and other crisis events that Western media does not sympathize with attacks in the Arab world as much as they do for attacks in the Western world. Even though many studies have already shown that news providers are inherently biased [2, 17], and further exemplified in bipartisan politics [4], the question arises if this would also applies to news media on social media platforms such as Twitter. This has far reaching implications given the media's power (even on Twitter) in shaping public discourse and perception of global events [32], which raises the following questions: To what extent is the news media biased in coverage of global unexpected, human-induced crisis events (such as terror attacks) on social media platforms like Twitter? And what makes a country newsworthy?

During November 2015, news stories across the web included headlines such as: *"Beirut, Also the Site of Deadly Attacks, Feels Forgotten"*[1] by The New York Times, *"Paris, Beirut, and the Language Used to Describe Terrorism"*[2] by The Atlantic, or *"Did the media ignore the Beirut bombings? Or did readers?"*[3] by Vox. While there was no doubt that coverage of the attacks was disproportionate, discussion as to why this was so was polarized. On one end, public and journalistic response blamed the media and its volume and style of coverage (cf., The Atlantic), and on the other end (cf., Vox) there were claims that the media in fact did its part in adequate news reporting, but since Western readers did not care, coverage was drastically reduced.

Bias can be viewed as a partial perspective on facts [37], which can be further broken down into three aspects [8]: selection bias (gatekeeping), or which stories are selected; coverage bias, or how much attention is given to a story; and statement bias, or how a story is reported. In this paper, we adopt research on media bias to study Twitter news (cf., [42, 15]) and use Twitter as a proxy. Here, we used these concepts and zoomed in on the Beirut and Paris attacks, to examine in detail whether their coverage on Twitter differed. Here, we focus on coverage bias, statement bias (specifically on characterizing overall sympathy), and whether sympathetic messages propagate further on Twitter. Since we are only concerned with the differences in reporting of these two events, we do not consider selection bias.

We believe these two events provide an interesting use-case to study for the following reasons: (a) it allows us to examine in

[1] http://www.nytimes.com/2015/11/16/world/middleeast/beirut-lebanon-attacks-paris.html; last retrieved: 25.04.2017

[2] http://www.theatlantic.com/international/archive/2015/11/paris-beirut-media-coverage/416457/; last retrieved: 25.04.2017

[3] http://www.vox.com/2015/11/16/9744640/paris-beirut-media; last retrieved: 25.04.2017

detail whether the Western and Arab media exhibit differences in reporting global crisis events, and if so, how such differences manifest (b) since the events were a day apart, it allows us to gain a deeper understanding of how news organizations on Twitter provide news under such reporting pressure[4] (c) it provides us with an opportunity to test how far our current computational techniques allow us to uncover such biases using publicly available social media data (in our case, Twitter), and in turn how this can aid journalistic social media practice in ensuring transparency and quality in produced content. For this study and all subsequent analyses, we use tweets as a proxy for media reporting. Twitter data was deemed suitable for three reasons: (a) the large quantity of tweets allows us to gain insight into the temporal aspect of news coverage at a finer grained level than with news articles (b) the uncensored nature of tweets that are collected with the Twitter Streaming API (c) the costs involved in labeling tweets for overall sympathy are lower than lengthier pieces of news coverage.

2. RESEARCH QUESTIONS & CONTRIBUTIONS

Given the foregoing motivation, we aim to better understand media coverage differences on Twitter through computationally capturing news sympathy during such unexpected, human-induced crisis events. We believe this has implications for how news organizations can better manage public opinion during the immediate and long-term aftermath of a political and/or religious crisis event (e.g., terror attack). Importantly, exposure to biases has been shown to have the capability to foster intolerance and create ideological segregation in major political and social issues [14], and this may be amplified across Western and Arab cultures. Given this, it is important to minimize such bias, even if only on social networks like Twitter. Therefore, we posit the following questions:

- **RQ1 - Coverage bias:** Was there a difference (in terms of normalized tweet volume) between Western and Arab media coverage of the Beirut and Paris attacks, and if so, to what extent?

- **RQ2 - Overall news sympathy:** Was there a difference between Western and Arab media in 'how' they covered the two attacks? Specifically, was there a difference in how sympathetic the tweets were in reporting the events?

- **RQ3 - Information propagation:** Do more sympathetic tweets propagate further throughout the Twitter network (i.e., receive more retweets)?

For coverage bias, we hypothesized that the Beirut attacks would receive less coverage from Western media, but not from Arab news media accounts on Twitter, with the inverse for coverage of Paris. This is in line with the news flow theory [35], which states that the prominence of a foreign country in the news is attributed to three groups of variables: (a) national traits (e.g. size and power of the foreign country), (b) relatedness (e.g., proximity to a foreign country in terms of geography, demography, etc.) and (c) events (e.g., wars, disasters, protests) [35, 43]. In this case, Paris is both geographically and culturally closer to Western countries, and given the timeline of both attacks, Paris would likely attract more coverage.

With respect to overall news sympathy, given news statements (e.g., NYTime's article[5]) on differential coverage and Diakopoulos's [11] work on Twitter newsworthiness, we expected that tweets from Western media covering the Beirut attacks would overall exhibit less sympathy than coverage of the Paris attacks, in contrast to Arab Twitter news media which would be impartial to both. In this paper, our objective is to explore these questions using a combination of NLP techniques and crowdsourcing on Twitter datasets. Finally, we look at information diffusion [39] during the two attacks, where we expect that tweets which are overall more sympathetic are more likely to spread throughout the Twitter network, by resulting in more retweets.

In this paper, we make the following contributions to CSCW and Social Computing research:

1. We show how NLP and machine learning techniques can be applied to answer whether there were differences in coverage by news media on Twitter between Western and Arab media, where we introduce a metric for measuring the overall sympathy of a news tweet that factors in religious reference.

2. We provide a public annotated multi-language (English, Arabic, French, German) dataset that can be used to train learning algorithms to predict overall sympathy during future unexpected, human-induced crisis events (see Supplementary Material A)).

3. BACKGROUND & RELATED WORK

3.1 November 2015 Attacks and Social Media Response

On 12 November, 2015, the city of Beirut (Lebanon) witnessed two bombings[6] at approximately 18:00 Eastern European Time (EET) / UTC+02:00, coordinated by two suicide bombers that detonated explosives in Bourj el-Barajneh, a southern suburb of Beirut. This suburb is largely inhabited by Shia Muslims, and reports estimate the number of deaths to be anywhere between 37 to 43, with over 200 injured. These bombings constituted the worst terrorist attack in Beirut since the end of the Lebanese Civil War in 1990. Shortly after the attacks, the Islamic State of Iraq and the Levant (ISIL) claimed responsibility for the attacks.

A day later, on 13 November, 2015 beginning at 21:20 Central European Time (CET) / UTC+01:00, three suicide bombers carried out a series of coordinated terrorist attacks in Paris[7] on its northern suburb, Saint-Denis. They struck near the Stade de France in Saint-Denis, followed by suicide bombings and mass shootings at cafés, restaurants and a music venue in central Paris. The attacks resulted in the deaths of 130 people, and injury of another 368 people. These attacks were purported to

[4] Any coverage bias here does not mean the two events are equivalent, only they are close in proximity which competes for attention.

[5] http://www.nytimes.com/2015/11/16/world/middleeast/beirut-lebanon-attacks-paris.html; last retrieved: 25.04.2017

[6] https://en.wikipedia.org/wiki/2015_Beirut_bombings; last retrieved: 25.04.2017

[7] https://en.wikipedia.org/wiki/November_2015_Paris_attacks; last retrieved: 25.04.2017

be the deadliest on France since World War II. Shortly after the attacks, ISIL also claimed responsibility for the attacks.

3.2 Defining and Modeling Overall News Sympathy

To quantify sympathy, we needed firstly to define sympathy. Merriam Webster's definition of sympathy is "the feeling that you care about and are sorry about someone else's trouble, grief, misfortune, etc. ; a feeling of support for something ; a state in which different people share the same interests, opinions, goals, etc."[8]. While sentiment can be captured through tokenization [28], the subjective nature of sympathy makes it harder to capture computationally. Furthermore, this is amplified in the context of coverage of crisis situations that involve political and religious individuals, entities, or organizations (as is the case in the Beirut and Paris attacks).

Typically, assessing news articles for polarity involves classifying text for three-valued sentiment: positive, neutral, and negative (e.g., [12]). However, recently researchers have taken more fine grained approaches towards modeling complex emotions (e.g., Lin and Margolin's [21] work on quantifying the diffusion of fear, sympathy, and solidarity during the Boston bombings and Schulz et al.'s [34] work on finer grained multi-valued sentiment classification). Moreover, recently Vargas et al. [40] showed that there are marked differences between the overall tweet sentiment and the sentiment expressed towards the subjects mentioned in the tweets. Mejova et al. [24] took a data-driven approach to understand how controversy interplays with emotional expression and biased language in the news using crowdsourcing, and found that for controversial issues, negative affect and biased language is prevalent, while the use of strong emotion is tempered.

Given the foregoing, we found these factors to be important for assessing Overall Sympathy of a crisis news tweet: sympathy (is the tweet sympathetic or not?), sentiment (is the tweet negative, neutral, or positive?), religious reference (does the tweet make an explicit reference to a religious entity(s)?), and political reference (does the tweet make an explicit reference to a political entity(s)?). Together, these components allow us to understand in higher resolution whether Arab and Western media covered the two attacks differently.

3.3 News Media Bias and Communication

Trumper et al. [33] examined biases in online news sources and social media communities around them, and by analyzing 80 international news sources during a two-week period, they showed that biases are subtle but observable, and follow geographical boundaries more closely than political ones. Sert et al. [44] proposed to leverage user comments along with the content of the online news articles to automatically identify the latent aspects of a given news topic, to be used as a first step in detecting the news resources that are biased towards certain subsets of these latent aspects. Park et al. [29] took a different approach towards media bias with NewsCube, where they automatically provide readers with multiple classified viewpoints on a news event of interest.

Dallmann et al. [9] investigated a dataset covering all political and economical news from four leading online German newspapers over four years, and showed that statistically significant differences in the reporting about specific parties can be detected between the analyzed online newspapers. Looking at how news sources tackle controversial issues, Mejova et al. [24] took a data-driven approach to understand how controversy interplays with emotional expression and biased language in the news using crowdsourcing as a data collection method. Interestingly, they found that when it comes to controversial issues, the use of negative affect and biased language is prevalent, while the use of strong emotion is tempered.

3.4 Twitter for Crisis and Controversy Understanding

Twitter has shown to be a rich resource to study media bias and controversy, especially in the aftermath of major events, whether political, religious, or natural (e.g., [36, 1, 4]). Morgan et al. [26] found that Twitter users share news in similar ways regardless of outlet or perceived ideology of outlet, and that as a user shares more news content, they tend to quickly include outlets with opposing viewpoints. Younes et al. [45] looked at how traditional media outlets and social media differ in the coverage of an event, and focused on coverage patterns of two sources (NYTimes articles and tweets) during the Egyptian uprising in January, 2011. To discover such patterns, they proposed a simple media bias measurement model for day-to-day news items built on top of topic models. They found that traditional news sources have a wider disparity in the ranks and hence a strong presence of media bias.

Wei et al. [42] proposed an empirical measure to quantify mainstream media bias based on sentiment analysis and showed that it correlates better with the actual political bias in the UK media than pure quantitative measures based on media coverage of various political parties. They then studied media behavior on Twitter during the 2010 UK General Election, and showed that while most information flow originated from the media, they seem to lose their dominant position in shaping public opinion during this general election. Olteanu et al. [27] investigated several crises using public Twitter data – including natural hazards and human-induced disasters – in a systematic manner and found that tweets expressing sympathy and emotional support constituted on average 20% of the crisis-related datasets. The four crises in which the messages in this category were most prevalent (>40%) all pertained to instantaneous disasters (which included terror attacks).

Given the foregoing, we also used Twitter to study crisis news media bias, and we approach this with attempts at capturing the multidimensional character of sympathy in unexpected, human-induced crises by crowdsourcing tweet annotations.

4. METHODOLOGY

To perform this study, we employ the following methodology (pipeline shown in Fig. 1): (1) Twitter Data Collection & Preprocessing (2) Data Processing (3) Crowdsourcing Annotation. Each step is described in detail below.

[8] http://www.merriam-webster.com/dictionary/sympathy; last retrieved: 25.04.2017

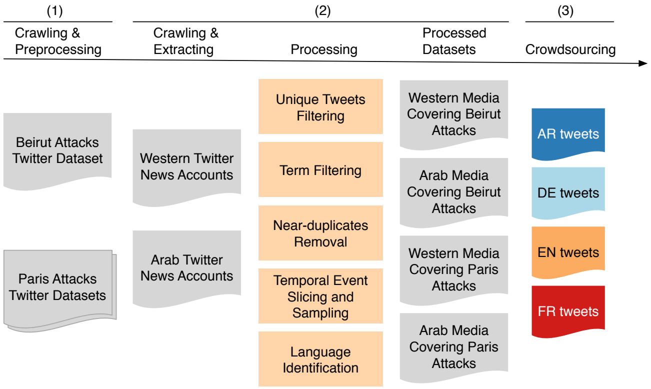

(1)
Crawling & Preprocessing
(2)
Crawling & Extracting
Processing
Processed Datasets
(3)
Crowdsourcing
Beirut Attacks Twitter Dataset
Paris Attacks Twitter Datasets
Western Twitter News Accounts
Arab Twitter News Accounts
Unique Tweets Filtering
Term Filtering
Near-duplicates Removal
Temporal Event Slicing and Sampling
Language Identification
Western Media Covering Beirut Attacks
Arab Media Covering Beirut Attacks
Western Media Covering Paris Attacks
Arab Media Covering Paris Attacks
AR tweets
DE tweets
EN tweets
FR tweets

Figure 1: Overview of methodological pipeline.

4.1 Twitter Data Collection & Preprocessing

4.1.1 Beirut dataset

We collected 906,538 tweets pertaining to the Beirut bombings on November 12, 2015 shortly after news of the attacks (approx. 21:00 CET), using the Twitter Streaming API and with these hashtags: #lebanon, #beirut, #beirut2paris, #beirutattacks, #beirutbombing. A limitation here is that our data was drawn from hashtags written in Latin alphabet. Despite that we were able to collect a sufficient number of tweets to run our analyses, where 62.1% of Arab media tweets in our final Beirut dataset consisted of Arabic tweets. The dataset was pruned for duplicates. Collection spanned 3.31 days, from 2015-11-12 18:51:07 UTC till 2015-11-16 02:17:04 UTC. Our dataset had a total of 667,073 (73.58%) retweets and 610,879 unique users. After removing retweets, we ended up with a dataset of 239,093 unique tweets. The top five hashtags for this dataset are shown in Table 1. What is immediately striking here are the high occurrences of the hashtags #paris and #parisattacks, where we attribute this due to the overlap between people's attention to the Paris attacks after having heard about the Beirut attacks.

Beirut		Paris	
Count	**Hashtag**	**Count**	**Hashtag**
95,280	#beirut	5,930,054	#parisattacks
66,898	#paris	5,359,273	#paris
56,021	#lebanon	1,597,903	#prayforparis
41,245	#parisattacks	844,384	#bataclan
21,546	#isis	760,320	#porteouverte

Table 1: Top most frequent hashtags for each dataset.

4.1.2 Paris datasets

Own dataset: We collected a total of 5,339,452 tweets during the two days (13th and 14th) after the November 2015 Paris attacks using the Twitter Streaming API and with the following hashtags: #paris, #france, #parisattacks, #prayforparis, and #porteouverte. Some of these JSON records (0.03%) were poorly structured and others were duplicates, and thus removed. This resulted in a total of 5,337,840 tweets. Collection spanned 1.17 days, from 2015-11-14 13:30:49 UTC till 2015-11-15 18:06:54 UTC. Our Paris dataset had a total of 4,045,046 (75.78%) retweets and 2,538,348 unique users.

Nick Ruest (NH) collection: In addition to our collection, we use a larger dataset created by Nick Ruest[9], who collected tweets shortly after the attacks occurred (approx. 23:00 UTC on November 13) with the following hashtags: #paris, #parisattacks, #prayforparis, and #porteouverte. We hydrated (collected metadata for tweets based on tweet IDs) the dataset on 2015-11-19 using the Twitter public API, and collected a total of 12,788,201 tweets. Tweet volume was less (attrition rate: 14.4%) than the original collection (N=14,939,154), which is common given that some tweets are removed (either by Twitter or by users). This collection spanned 33.96 days, from 2015-11-04 21:14:39 UTC till 2015-12-08 20:54:03 UTC. NH's Paris dataset had a total of 9,742,241 (76.18%) retweets and 4,127,762 unique users.

Merged Paris dataset: For later analysis, we merged both our own Paris dataset together with NH's. We expanded the Paris dataset (by merging with the NH dataset) because it spanned only 1.17 days after the Paris attacks, while the Beirut dataset spanned 3.31 days. This temporally aligns the two datasets. Furthermore, the additional dataset allowed us to run experiments with a larger sample size. Merging datasets resulted in a total of 16,868,318 tweets. Our dataset consisted of 75.6% retweets, resulting in a total of 4,110,291 unique tweets with 5,772,262 unique users. This merged Paris dataset (with retweets and duplicates removed) is used for all subsequent analyses, and will be referred to simply as the Paris dataset.

4.2 Data Processing

To answer our questions about media coverage bias (specifically differences in overall news sympathy), we apply multiple processing steps. We first crawl Twitter news (and blog) accounts, extract the news tweets from our datasets, identify the language of the tweets, slice our data by time to ensure that the two event timelines match in duration, and finally draw samples to ensure our data is sizeable and ready for human computation. Details of each step is described below.

Region	Country (population per million)
Middle East (6)	Egypt (89.6), Iraq (34.8), Saudi Arabia (30.9), United Arab Emirates (9.1), Jordan (6.6), Lebanon (4.5)
Western (5)	USA (318.9), Germany (80.9), France (66.2), United Kingdom (64.5), Spain (46.4)

Table 2: Regions and countries of interest for our analysis. Population estimates drawn from the World Bank (http://www.worldbank.org/; last retrieved: 25.04.2017)

4.2.1 Crawling and Verifying Twitter news accounts

Our first step was to identify and collect influential Western and Arab news accounts on Twitter. The list of countries chosen across the Middle East and the Western world are shown in Table 2. With respect to the Middle East, we chose countries that were geographically near Lebanon, and that did not have explicit and/or visible political nor religious conflicts with

[9] http://ruebot.net/post/look-14939154-paris-bataclan-parisattacks-porteouverte-tweets; last retrieved: 25.04.2017

Lebanon at the time of collection (e.g., Syria was excluded due to the ongoing conflict at the time of data collection). For Western countries, we based our decision on population size, language (English being most prominent), and proximity to France. Furthermore, based on the language distribution of our data (shown later in Table 4), we decided to include Spain, and not Italy, despite Italy's larger population. Despite that Twitter is dominated by English language users[10], we wanted to ensure that we were collecting news media tweets from both English as well as the native language of the countries of interest. To find news media[11] accounts on Twitter from these countries, we followed a two-step approach:

Crawling: We found a seed set of accounts automatically (using Twitter's relevance-based Search API) by crawling user accounts (Table 3) with news related queries (e.g., 'France news' for English queries; 'Nouvelles France' for native language queries). This first step deliberately takes a crude computational approach as curating news organizations by experts may be subject to bias, and could exclude unfamiliar news accounts that possibly became highly active during the time of crisis (e.g., bloggers new to the scene). To ensure some measure of influence, we chose to only retrieve user accounts that matched our query with at least 5,000 followers. Despite earlier research that showed that a high number of followers does not always mean an influential user [6], we used follower count as a simple heuristic to gather prominent news accounts. Our query returned six results (which were kept) with a follower count less than 5,000 (Max=4,935, Min=4,473), with the rest above 5,000 (Max=31.4M, CNN Breaking News). We did not set a limit on account creation date nor on Twitter verification, as our experiments showed that: (a) some new bloggers and news agencies with accounts created only a year earlier (2014) appeared to be quite active in reporting events (b) even major news accounts were sometimes not Twitter verified, wherein we could potentially miss important news accounts if we enabled this filter.

Verification: Returned accounts were manually inspected to ensure they comprise news media outlets and blogger accounts. This was done by cross-checking whether names occur in public lists (e.g., Wikipedia pages 'News media in {Country}'[12]) and if blogger accounts, whether they cross-link to a webpage. We had a total of unique 208 news media accounts, where 93.3% (194/208) of our dataset consists of news outlet accounts, and the remainder 8 journalist and 6 blog accounts. Furthermore, there was some overlap in accounts for Western media (38/117) coverage and for Arab media (38/91) coverage of Paris and Beirut. The final list of unique crawled news organizations (N=208) is provided as a supplementary dataset to this paper (see Supplementary Material B)[13].

4.2.2 Extracting and Cleaning News Tweets

After gathering a set of news accounts across countries using both English and native language queries, we then matched these user IDs with all IDs in our Beirut and Paris datasets. The full set of queries used, the total number of Twitter news accounts found, and the amount and percentage extracted from both datasets are shown in Table 3. This process resulted in four datasets: (1) Arab media covering the Beirut attacks (N=2,766) (2) Arab media covering the Paris attacks (N=2,728) (3) Western media covering the Beirut attacks (N=245) (4) Western media covering the Paris attacks (N=9,245). The datasets combined resulted in 14,984 tweets.

As an additional step, we made sure that within each dataset, there was no mention of the other set of attacks (e.g., we removed all mentions of Paris from the Beirut dataset), and that all tweets pertained to the events in question. Even though the two attacks happened a day apart, where we would expect cross-pollination across messages, we deliberately chose not to include tweets that reference both the Paris and Beirut attacks, as this may influence our attempts at investigating Twitter media bias within each dataset separately.

For the Beirut dataset, we filtered out tweets that included these terms: paris, parís france, parisattacks, bataclan, parisattacks, porteouverte. For the Paris dataset, we filtered out the following terms: beirut, lebanon, beirutattacks, لبنان [Lebanon], بيروت [Beirut]. Finally, to ensure that our dataset contains only unique tweets without any near duplicates (as this would cause redundancy later in the annotation task), we removed all partial duplicates from our resulting datasets. This was done by applying the Levenshtein distance [20] string similarity metric on the tweet texts of each dataset, with a threshold set to 0.1. This reduced the size of our dataset to 12,814 tweets.

4.2.3 Language Identification

To prepare our Beirut and Paris datasets for analysis of sympathy, we need to be able to identify the language of the tweets so crowdworkers can annotate them. To do so, we used the langid.py [22] language identification Python package, and computed the (percentage) distribution of languages. We used langid.py as it provided us with classification probabilities while Twitter's 'lang' value does not provide such a metric, so we could manually adjust the 'lang' of low confidence tweets. To deal with any misclassifications from langid.py, we disqualified any tweets with a normalized classification probability of < 0.95, reducing our dataset to 10,460 unique tweets. For the remainder of the tweets, we manually inspected and reclassified all tweets with a normalized probability of > 0.95.

We found that across all datasets combined, the languages of tweets were either English (67%), German (13.9%), Arabic (10.1%), French (8.3%), or Spanish (0.7%). Given the low percentage of Spanish in our dataset, we decided to exclude any Spanish tweets from all subsequent analyses. The language distribution per dataset is shown in Table 4.

[10] http://www.beevolve.com/twitter-statistics/#a3; last retrieved: 25.04.2017

[11] We use Wikipedia's definition of news media, which includes blog accounts: https://en.wikipedia.org/wiki/News_media ; last retrieved: 25.04.2017

[12] E.g., https://en.wikipedia.org/wiki/News_media_in_the_United_States ; last retrieved: 25.04.2017

[13] We show the name, user name, user description, the country / query used to retrieve the account, and follower count at the time of crawling – all of which are publicly available data. Additionally, we include tweet count and mean ONS score.

			Beirut	Paris
Country	**Query**	**Total**	**Found (%)**	**Found (%)**
France	‘France news’	51	4 (7.8%)	26 (51.0%)
	‘Nouvelles France’	8	0 (0%)	6 (75%)
Germany	‘Germany news’	24	4 (16.7%)	10 (41.7%)
	‘Deutsche Nachrichten’	18	6 (33.3%)	17 (94.4%)
Spain	‘Spain news’	35	1 (2.9%)	12 (34.3%)
	‘Noticias de España”	4	2 (50.0%)	4 (100.0%)
USA	‘USA news’	150	21 (14.0%)	50 (33.3%)
UK	‘UK news’	207	22 (10.6%)	65 (31.4%)
Lebanon	‘Lebanon news’	54	25 (46.3%)	25 (46.3%)
	‘اخبار لبنان’	6	1 (16.7%)	1 (16.7%)
Jordan	‘Jordan news‘	25	2 (8.0%)	2 (8.0%)
	‘اخبار الاردن’	11	1 (9.1%)	2 (18.2%)
UAE	‘UAE news’	42	5 (11.9%)	14 (33.3%)
	‘اخبار الامارات’	13	1 (7.7%)	1 (7.7%)
Saudi Arabia	‘Saudi Arabia news’	8	1 (12.5%)	2 (25.0%)
	‘اخبار السعوديه’	20	1 (5.0%)	2 (10%)
Egypt	‘Egypt news’	95	25 (26.3%)	44 (46.3%)
	‘اخبار مصر’	30	1 (3.3%)	7 (23.3%)
Iraq	‘Iraq news‘	41	14 (34.1%)	15 (36.6%)
	‘اخبار العراق’	10	1 (10.0%)	1 (10.0%)

Table 3: Twitter news account queries and number found in each dataset.

Lang.	**Western media (Beirut attacks)**	**Western media (Paris attacks)**	**Arab media (Beirut attacks)**	**Arab media (Paris attacks)**
EN (%)	80.2	68.4	37.9	71.7
AR (%)	0	0	62.1	28
DE (%)	6.1	19.4	0	0
FR (%)	0	11.5	0	0.33
ES (%)	13.7	0.7	0	0

Table 4: Language distribution for each dataset.

4.2.4 *Temporal event slicing and sampling*

Given that our Beirut and Paris datasets differed temporally in coverage of the attacks, it would be unfair to compare sympathy as tweets posted 5 days after the attacks may differ for example than tweets posted two weeks after the attacks. Since we were constrained by the size and coverage duration of our Beirut dataset, we used the coverage length of that dataset as a seed to slice the Paris dataset. For our processed Beirut dataset, our earliest coverage started on 2015-11-12 18:52:30 UTC (approx. 4 hours after the Beirut attacks that took place around 18:00 EET) and went until 2015-11-16 02:00:10 UTC. This amounts to exactly 3.3 days. Thereafter, we applied the same time slice for the processed Paris dataset, where earliest coverage started from 2015-11-13 21:15:20 UTC (approximately one hour after the Paris attacks, which started at 20:20 UTC) until 2015-11-17 02:30:23 UTC, giving exactly 3.22 days. This time slicing further reduced the total size of our combined datasets to 7,768 unique tweets: Western media coverage of Beirut (N=131), Western media coverage of Paris (N=5,298), Arab media coverage of Beirut (N=287), and Arab media coverage of Paris (N=1,566).

Finally, in order to send our tweets for annotation by crowdworkers, we only needed a sufficient sample from each of our four resulting datasets to avoid lengthy crowdwork time and costs, and to later test classifier performance on unlabeled data (Section 5.2). Therefore, we drew a random sample of 1,000 tweets from the Paris datasets. However, random sampling may miss important tweets that occurred on specific days within our 3.3 days. Therefore, we split each dataset into separate buckets of approximately 24 hours, and drew normalized random samples from each bucket to eliminate bias in drawing more samples from a day that happens to have more records. The normalization constant was calculated by dividing the size of the desired sample draw (1,000) by the total number of rows in each dataset. For each bucket, the sample drawn was the number of records in that bucket multiplied by the normalization constant, and rounded to ensure all day buckets cap at 1,000 records. This process reduced the size of the Paris attacks datasets (Western and Arab media coverage) each to N=1,000. The final language distribution of our language-specific datasets ready for annotation is shown in Fig. 2.

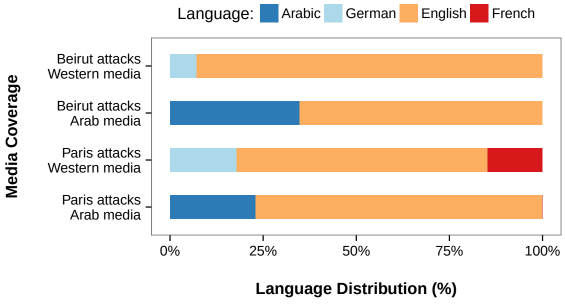


Figure 2: Overall language distribution of our datasets to be annotated.

4.3 Crowdsourced Overall Sympathy Annotation

To annotate our datasets, we employed crowdsource workers through the CrowdFlower[14] platform. Characteristics of the annotation task are detailed below.

4.3.1 *Annotation Task Characteristics*

We had a total of 4 language-specific datasets (all translated from English by native speakers). A language-specific dataset here covered both the Beirut and Paris attacks, which meant there was an assumption that a worker had to be familiar with both events to accurately annotate the tweets. For each language-specific annotation task, we provided instructions and examples in that target language. For each tweet, a worker had to label it for sympathy (sympathetic, unsympathetic), sentiment (positive, neutral, negative), religious reference (yes, no), and political reference (yes, no).

For worker selection, we ensured workers spoke the target language. Also, we did not set a restriction on worker location, as a worker could speak a target language (e.g., Arabic) yet reside in a non-Arabic speaking country. In our case, there could be a risk that workers who are not local to Lebanon

[14] http://crowdflower.com/; last retrieved: 25.04.2017

Lang.	# Tweets	Label	*a*=3 (%)	Fleiss' Kappa
EN	1732	Sentiment	69.3	0.27
		Sympathy	83.4	0.42
		Religious	85.4	0.46
		Political	85.2	0.47
AR	354	Sentiment	69.1	0.32
		Sympathy	82.7	0.29
		Religious	88.2	0.42
		Political	77.9	0.43
FR	147	Sentiment	73.1	0.22
		Sympathy	82.4	0.32
		Religious	92.7	0.36
		Political	90.7	0.58
DE	185	Sentiment	73.1	0.28
		Sympathy	85.5	0.38
		Religious	91.3	0.39
		Political	92.0	0.45

Table 5: Trusted worker agreement (*a*=3) scores across languages (from CrowdFlower) and our own computed Fleiss' Kappa scores.

perhaps do not know when a political reference is being made – however, given that we require at least 3 judgments per tweet, this risk is mitigated. While we set target language requirements for workers, we did not make our tweets into images (cf., [3]), which can be a limitation. Additionally, following standard guidelines from CrowdFlower, 10-15 tweets per language-specific task were classified by the authors of this paper. We did not trust the assessment of any worker who differed significantly from our own (cut-off point of less than 70% agreement).

Workers were presented with the original tweet, and included media items (image or video), and were asked to label that tweet. While we are aware of the potential ethical concerns on behalf of Twitter users in displaying their name (cf., [27]), in our case we were only displaying tweets from news organizations, who are presumably aware and even encourage publicizing their content. Importantly, omitting the username of our tweets would risk misrepresenting the original tweet and its overall sympathy.

Trusted workers took on average (across all languages) 57 seconds (interquartile mean) to label each tweet. We collected labels from at least 3 different trusted workers per tweet and task[15], where the final label of the tweet was determined by majority vote. We followed the guidelines of CrowdFlower, and set a limit of no worker labeling more than 300 items in our rating task. Workers were paid 10 cents per page, where each page contained 5 tweets. This amounted to approximately $10 per 100 tweets. In the end, we had a total of 2,390 x 4 x 3 = 28,680 labels (excluding the 'not applicable' checkbox).

4.3.2 Task Description

Below are the instructions given during the annotation phase to crowdsource workers, that directly precede a fully worked out example (not shown here). These same instructions and worked out example were additionally translated to Arabic, French, and German, by native speakers.

Overview In this task we want your help rating tweets, that were posted during November, 2015. The events in question are the November 13, 2015 Paris attacks and the November 12, 2015 Beirut attacks, and anything related to them. Please read the tweets carefully. If the tweet contains Media (photo or video), please take the media into account when forming your judgment. It is important that you form an impression based on the text of the tweet and the accompanying media if available.If the tweet onlyor additionally contains a URL link (e.g., http://...), then you may click on that and use it to help you form a judgmentonly if you are completely unable to rate the tweet based on the text and media alone. In other words, URL links should serve only as backup measures. Please keep in mind that all the questions asked are about the tweet text and accompanying media, not the contents of the URL the tweet provides. Also remember, some tweets do not have a Media link nor a URL link, but only text.

We ask you to rate each tweet along each of the aspects shown below:

(A) **Not applicable [checkbox]:** In case the tweet is not applicable or relevant at all to the November 2015 events related to France or Lebanon, please check the box "Not applicable; not readable...". This will hide the other questions, which means you skip rating that specific tweet.

(1) **Overall sentiment rating [positive, neutral, negative]**

- Positive sentiment means some aspects of the overall tweet and accompanying media uncovers a positive mood or sentiment, such as happiness, support, hope, enthusiasm, kindness, praise, recommendations or a favorable comparison. Example: "NewYork in solidarity with France #ParisAttacks"
- Neutral sentiment means that the overall tweet and accompanying media is only informative in nature and provides no hint as to the mood of the text or media.
- Negative sentiment means that overall some aspects of the tweet and accompanying media uncover a negative mood or sentiment such as sadness, hate, violence, discrimination, criticism, insults or a negative comparison. Example: "Relatives search for missing, grief pours out on social media after Paris attacks "

(2) **Sympathy rating [sympathetic, unsympathetic]**

- Sympathetic if the tweet text and accompanying media highlights or shows sympathy to the affected individual(s) or subject(s) of the tweet,then it would be rated as positive sympathy. This includes thoughts, prayers, gratitude, sadness, solidarity, and so on concerning affected individuals. Example: "Watch the world stand in solidarity with France and sing La Marseillaise following the #ParisAttacks".

  Note: Sometimes the tweet might appear neutral, but the tweet media shows an image that evokes sympathy or lack of sympathy. In such a case, the tweet is sympathetic.
- Unsympathetic if the tweet is factual and shows no sympathy with the affected individual(s) or subject(s) of the tweet, then it would be rated as unsympathetic. This includes lack of sympathy, neutral, insensitive, uncaring, indifference, coldness, lack of solidarity, etc. Example: "Follow FRANCE 24's live blog for all the latest on the #ParisAttacks"; "SYRIAN PASSPORT FOUND NEAR BODY OF ONE OF PARIS SUICIDE BOMBERS."

Note: Tweets that are purely factual (links to news articles without comment) are not necessarily unsympathetic – consider whether the fact/news/image itself is sympathetic towards the topic.

Note2: While Sympathy and Sentiment might be the same sometimes, this is not always the case. E.g., a tweet that is sympathetic (labeled as such because it shows an image that makes one sympathize with the victims affected) could be rated as having a negative sentiment as it evokes sadness.

(3) **Religious reference to individuals / groups / organizations [yes, no]**

- Yes means some aspects of the overall tweet and accompanying media make explicit reference to named or unnamed religious individuals, groups, or organizations. Example: "France blames ISIS for 'act of war,' vows 'merciless' response to #ParisAttacks"
- No means no aspects of the overall tweet and accompanying media make explicit reference to named or unnamed religious individuals, groups, or

[15] Note that if a tweet has different labels from all 3 workers, the CrowdFlower platform brings in additional workers.

organizations. Example: “Mourners are gathering in Paris to remember loved ones killed in terror attacks”

(4) **Political reference to individuals / groups / organizations [yes, no]**

- Yes means some aspects of the overall tweet and accompanying media make explicit reference to named or unnamed political individuals, groups, or organizations. Example: “Iranian President cancels trip to France after terror attacks ”
- No means no aspects of the overall tweet and accompanying media make explicit reference to named or unnamed political individuals, groups, or organizations. Example: “Stay safe: Americans in France urged to be vigilant in wake of ParisAttacks”

Note: It is debatable whether one should treat religious terror organizations as political ones. For this task, tweets that refer to terrorist organizations should not be labeled as having political reference.

4.3.3 Annotation Quality

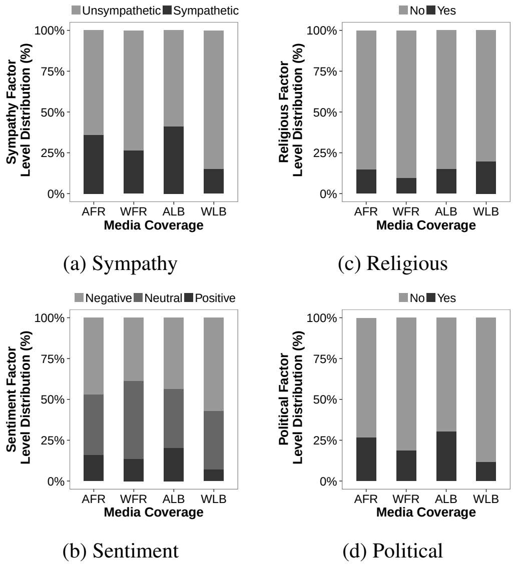
As we know from sentiment analysis research (cf., Pang and Lee, [28]), sentiment classification and additionally sympathy classification can be subjective. This is exacerbated by the fact that a tweet can have either positive or negative sentiment, yet still be labeled as sympathetic. To this end, we provided detailed instructions and examples of Positive, Negative, and Neutral tweets (see Section 4.3.2) to ensure that workers correctly label the data. The foregoing notwithstanding, our inter-rater agreement scores drawn from CrowdFlower (shown in Table 5) are promising, with the lowest being 69.1% for sentiment classification of Arabic tweets, which is in line with previous work, and highlights the reliability of large-scale crowdsourced social media annotation [27, 10]. As a further measure, we also computed Fleiss’ Kappa for the labeled tweets, and found reasonable agreement scores (Table 5), with sentiment expectedly exhibiting lowest agreement across languages.

As a further test, we followed the approach by Olteanu et al. [27] and independently (N=2) rated a random sample of 15 tweets from each language dataset[16] (total N=60) and computed unweighted Cohen’s Kappa for each factor except sentiment (which was weighted). Our ratings reached substantial agreement on religious (κ=0.76, CI: [0.53,0.98]), political reference (κ=0.71, CI: [0.48,0.93]), sentiment (κ=0.70, CI: [0.52,0.88]) and sympathy (κ=0.71, CI: [0.50,0.91]) labels. Thereafter, we took our agreement ratings and compared their joint label with those provided by workers, and we reached reasonable agreement for all factors: political reference (κ=0.63, CI: [0.35,0.91]), religious reference (κ=0.65, CI: [0.36,0.94]), sentiment (κ=0.64, CI: [0.42,0.87]), and sympathy (κ=0.59, CI: [0.37,0.82]).

5. ANALYSIS & CLASSIFICATION SETUP

5.1 Measuring Overall News Sympathy

The percentage level distributions across each factor investigated are shown in Fig. 3. To measure overall sympathy of the tweets, that factored in sentiment, sympathy, political and religious reference, we needed a single composite score to describe the data. Such a score helps us capture the multidimensional nature of sympathy, as well as provide us with a method to turn our nominal factors into a single continuous variable, which simplifies subsequent analysis. Our modeling makes the assumption that religious references in a tweet can negatively affect a tweet’s overall sympathy, given that this was the topic of controversy in published news articles, and also in line with Preece & Ghozati’s [31] findings that religious online communities exhibit a lack of empathy (or hostility). To reduce subjectivity in modeling overall crisis news sympathy, we began by exploring the measured factors.



(a) Sympathy (c) Religious

(b) Sentiment (d) Political

Figure 3: Level distributions for all factors tested on each dataset: Arab media coverage of Paris attacks (AFR), Western media coverage of Paris attacks (WFR), Arab media coverage of Beirut attacks (ALB), Western media coverage of Beirut attacks (WLB).

5.1.1 Exploratory Factor Analysis

We ran exploratory factor analysis (EFA) following best practices [7] to better understand our labeled data. Since Bartlett’s Sphericity Test was significant (χ^2(2,390) = 22535, p<0.001) and Kaiser-Meyer-Olkin was greater than 0.5 (KMO=0.56), our data allowed for EFA. Given our assumptions that sentiment and sympathy pertain to a tweet’s sympathy, and that religious and political reference can negatively slant a tweet’s sympathy, we tested two factors corresponding to each set. Furthermore, since we assumed that factors would be related, we used oblique rotation (‘oblimin’) along with standard principal axes factoring. The standardized loadings are shown in Table 6, which show that the first factor ties sympathy and sentiment together, while the second factor (no significant loadings) still shows a clustering of politics and religion.

	Factor 1	Factor 2
Sentiment	**0.49**	-0.01
Sympathy	**0.39**	0.02
Religious	-0.25	0.24
Political	0.18	0.26

Factor loadings of 0.3 and above are marked in bold.

Table 6: Exploratory factor analysis (EFA) applied to our labeled data.

[16] We ensured that the tweets were translated by native speakers.

5.1.2 Modeling Overall News Sympathy

To empirically assign weights to our scoring function, we further tested correlations[17] between each of our variables (shown in Table 7). There was a weak yet significant negative correlation between religious reference and sentiment, and between religious reference and sympathy.

	Sentiment	Sympathy	Religious
Sentiment			
Sympathy	0.19*** (0.49)		
Religious	-0.10*** (0.11)	-0.12*** (0.12)	
Political	0.06** (0.09)	0.09*** (0.09)	0.02 (0.02)

*p<.05, **p<.01, ***p<.001

Table 7: Spearman correlation matrix with Cramer's V (ϕ_c) correlations shown in parentheses.

With respect to politics and sentiment, there was a very weak yet significant correlation, and also a weak yet significant correlation for political reference and sympathy. Lastly, there was a strong positive correlation between sentiment and sympathy (shown by Cramer's V), indicating we are likely measuring the same concept. Given these correlations and in line with news media statements, we decided to give an extra penalty on a tweet's overall sympathy if it contained reference to a religious entity. It is important to mention here that which religion under which specific context (in our case a terror attack) could affect the modeling outcome. Furthermore, since we found very weak correlations for political reference and since it was not explicitly mentioned as a factor in news articles, we chose not to include it into our final model. Moreover, political reference did not have any grounding in prior literature to warrant inclusion. However, it was interesting to observe that Arab media contained more tweets with political reference than Western media (Fig. 3). Finally, since sentiment and sympathy positively correlate, both of these factors are not penalized, and under few cases may cancel each other out. Therefore, we modeled Overall News Sympathy (ONS) as:

$$ONS = \frac{\sum_{i=1}^{n}(s_i + y_i) - \sum_{i=1}^{n}(\alpha \cdot r_i)}{\psi} \tag{1}$$

where: $s \in \{-1, 0, 1\}$, $y \in \{-1, 1\}$, and $r \in \{-1, 1\}$. Sympathy s and sentiment y account directly into the ONS score, whereas religious r reference accounts negatively. The values for sympathy and sentiment come directly from the crowdsourcing results, where -1 indicates negative sympathy or sentiment, and 1 indicates positive sympathy or sentiment. Neutral sentiment is represented by the value 0. For religious reference, we translated from true or false to 1 or -1, respectively. In our model, we have the built in assumption that a negative sentiment tweet which exhibits sympathy would be treated as neutral; this treatment is a result of our observed data[18].

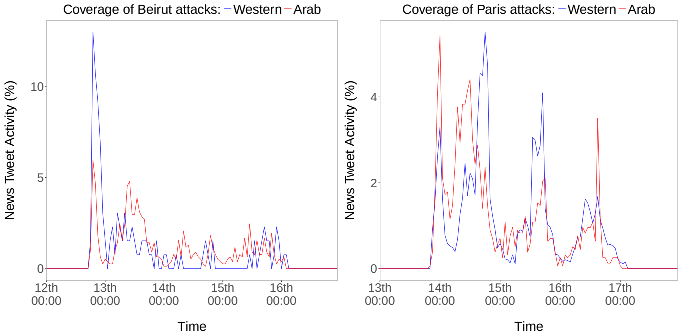


Figure 4: Hour by hour normalized tweet activity volume across the days after the Beirut attacks (left) and Paris attacks (right). *Best seen in color.*

We defined a weight of α=0.12 for religious reference to balance its impact on the score. This is based on our preceding assumption and correlation analyses that religious reference negatively affects the tweet sentiment and sympathy. This should have some effect due to observed negative correlations, without extremely mitigating the influence of actual sympathy and sentiment ratings. We normalize the ONS score to range from -1 to 1 using constant ψ:

$$\psi = (s_{MAX} + y_{MAX}) - (\alpha \cdot r_{MIN}) \tag{2}$$

This gives us: $ONS \in \mathbb{R}, -1 \leq ONS \leq 1$

5.2 Classifying Overall News Sympathy

The crowdsourced annotation task was not able to cover the entire corpus due to a number of practical constraints (e.g., the cost and availability of workers). In order to generalize the analysis to unlabeled data, we also developed machine learning models that recognize the overall crisis news sympathy level of tweets. The learning task was designed as a binary classification problem where a model aims to classify if a given tweet is sympathetic (positive ONS score (> 0)) or unsympathetic (negative ONS score (< 0)). A classification model was built for each language, and each model was trained using the data annotated through crowdsourcing and applied to the rest.

For the classification model, we adopted a convolutional neural network (CNN) combined with word2vec embeddings (cf., [18]). This work extended the recent successful applications of deep learning models in NLP tasks to sentence classification and reported that the model achieves high performance even with a very simple architecture (i.e., one convolutional layer) and little tuning. This CNN architecture has a single channel (word2vec embeddings) and three layers: the embedding layer which translates the words of a sentence to corresponding word2vec embeddings; the convolutional layer that applies filters over sliding windows of words and extracts the feature through max-pooling; the final layer performs dropout regularization [38] and classifies the result using softmax (refer to [18] for more details of the model).

[17] We use Spearman's rho in addition to Cramer's V (since our data is not normally distributed) by making an assumption that our dichotomous variables exhibit a monotonic relationship, and therefore can be treated as ordinal variables. Moreover, Cramer's V is symmetric, and does not show negative relationships.

[18] Tweets labeled as unsympathetic made up 96.2% of negative and neutral sentiment labels.

The word2vec embeddings were pre-trained for each language [25]. Given a text corpus, word2vec learns a lower-dimensional vector (100 dimensions in our experiment) representation of words that preserves the semantic distance between them. As we deal with tweets for a particular topic, we used the entire set of collected tweets for training the embeddings instead of using an available general corpus. We briefly describe the configuration of the model (the details can be found in the TensorFlow implementation we used [5]). The model is trained through stochastic gradient descent with the Adadelta update rule [46]. It trains for 100 epochs using shuffled mini-batches of 64 instances. Three different filter sizes (3, 4, and 5 words) were used for convolution, and 128 filters were made for each size. Dropout rate was set to .5.

For evaluation of the approach, we ran 10 fold cross validation (Table 8). We chose balanced accuracy[19] as the metric of choice to account for class imbalances in our data, and report the precision and recall for each class. Overall, the weighted average of the balanced accuracies across languages was 70.4% (shown in Table 8).

Overall balanced accuracy: 0.704		
EN balanced accuracy: 0.744		
	Positive	Negative
Precision	0.617	0.870
Recall	0.585	0.882
AR balanced accuracy: 0.504		
	Positive	Negative
Precision	0.813	0.195
Recall	0.698	0.293
FR balanced accuracy: 0.661		
	Positive	Negative
Precision	0.561	0.761
Recall	0.406	0.831
DE balanced accuracy: 0.747		
	Positive	Negative
Precision	0.813	0.195
Recall	0.698	0.293

Table 8: Balanced accuracies (Positive ONS/Sympahetic, Negative ONS/Unsympathetic) for unlabeled news media tweets (N=5,378) across each language.

6. RESULTS

6.1 Coverage Bias

We firstly looked at the normalized tweet volume from Western and Arab media across both attacks (N=7,768), where we visualize this daily and hourly in Fig. 4. From the graphs, we can see that for the Beirut attacks, there was more coverage from Arab media, while the inverse for the Paris attacks, which shows more Western media coverage. To test this, we ran a Chi-square test with Yates' continuity correction across all days to compare the difference between Arab and Western media coverage. In line with our hypothesis, our test revealed a statistically significant difference in tweet activity volume between how much Western (M=0.909) and Arab media (M=5.37) covered the Beirut attacks and by how much Western (M=36.79) and Arab media (M=10.87) covered the Paris attacks (χ^2(1, N=7,768) = 1489, p<0.001, ϕ=0.44, odds ratio=0.05). From this, we accept the alternative hypothesis that there was indeed coverage bias across both attacks.

Furthermore, did Western and Arab media follow a similar pattern of posting tweets? To answer this, we ran correlation analyses between hourly tweet volume for both attacks across media coverage, and found that for the Beirut attacks, Western and Arab media exhibited a medium-sized correlation (Spearman (ρ) = 0.608, p<0.001; ϕ_c=0.698) and for the Paris attacks exhibited a significant and strong correlation (Spearman (ρ) = 0.943, p<0.001; ϕ_c=0.918). Here we see that with respect to tweet activity volume, Western and Arab media are engaged at approximately similar time points, even though this effect was stronger for the events in Paris.

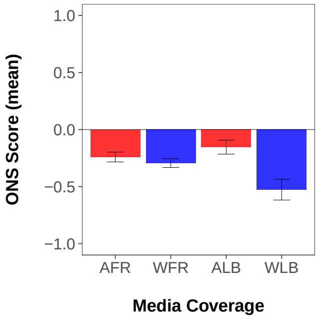


Figure 5: Overall News Sympathy (ONS) scores for Arab media coverage of Paris attacks (AFR), Western media coverage of Paris attacks (WFR), Arab media coverage of Beirut attacks (ALB), Western media coverage of Beirut attacks (WLB).

6.2 Overall News Sympathy

Examples of the highest and lowest ranked tweets with respect to our ONS score are shown in Table 10 and Table 11, respectively. The means and confidence intervals for ONS scores of Arab and Western media for coverage of each attack are shown in Fig. 5. It is worth noting that our ONS score strongly correlates with the Sympathy labels provided by workers (Spearman (ρ) = 0.815, p<0.001), which shows that this score does correspond to sympathy treated as a single variable. Since our data is not normally distributed, we ran Mann-Whitney U tests to compare the difference between overall crisis news coverage sympathy between Arab and Western media for each of the datasets on the Paris and Beirut attacks. We found a statistically significant difference in ONS scores between Western (*Md*= -0.7) and Arab (*Md*= -0.4) media in coverage of the Beirut attacks (Z=5.79, p<0.001, r=0.25), however not between Western (*Md*= -0.4) and Arab (*Md*= -0.4) media in coverage of the Paris attacks (Z=1.49 p=0.134, r=0.03).

	Paris	Beirut
Arab	0.617 (N=708)	0.976 (N=338)
Western	0.264 (N=4270)	0.186† (N=112)

†Value drawn from labeled (ground truth) data.

Table 9: Ratio of tweets with a positive ONS (i.e., sympathetic) score.

With respect to our classification results covering our unlabeled data (which excludes Spanish (N=61)), we found (Table 9) that each media coverage (Arab, Western) is overall more

[19] Arithmetic mean of class-specific accuracies.

sympathetic towards the country affected in their respective region. It is important to mention here that given the low quantity of unlabeled tweets from Western media coverage of Beirut attacks (N=1), we instead report the ratio of positive ONS from the crowd labeled (ground truth) data. Based on these results, it appears that Western media was more sympathetic towards Paris, while Arab media was more sympathetic towards Beirut.

Dataset	Screen Name	Text	Media	URL
Western media (Beirut attacks)	@France24_en	“IN THE PAPERS - ‘United by tragedy’ ”	yes	yes
Arab media (Paris attacks)	@youm7	[‘The high activity of the hashtag #paris shows the value of Twitter #French_president’] هاشتاج #باريس يتصدر قائمة الأكثر تداولا بـ تويتر ، #الرييس_الفرنسي	yes	no
Western media (Paris attacks)	@YahooNewsUK	“For those concerned about loved ones after #ParisAttacks”	yes	no
Arab media (Beirut attacks)	@Live961	[‘In pictures... the leader stands in solidarity with Lebanon.’] بالصور.. ، الزعيم ، يتضامن مع لبنان #lebanon	no	yes

Table 10: Examples of top ranked tweets in terms of ONS across Western and Arab media.

Dataset	Screen Name	Text	Media	URL
Western media (Beirut attacks)	@NBCNews	“ISIS claims responsibility for deadly Beirut explosions that killed at least 37”	yes	yes
Arab media (Paris attacks)	@Arab_News	“#AFP: #Russia FM says Paris attacks “justify” need to combat #Daesh, Al-Nusra #terror groups. #IS #ISIS #ParisAttacks”	no	no
Western media (Paris attacks)	@tagesspiegel	“#Erdogan kämpft gegen #IS: Karikatur von Klaus Stuttmann. #Türkei #Syrien #G20 #G20Turkey #ParisAttacks #Paris” [‘Erdogan battling against IS: Cartoon by Klaus Stuttmann’]	yes	no
Arab media (Beirut attacks)	@AJENews	“Hezbollah chief vows to continue fight against ISIL after deadly Beirut bombings”	yes	yes

Table 11: Examples of lowest ranked tweets in terms of ONS across Western and Arab media.

6.3 Sympathetic Tweet Propagation

As an additional analysis, we wanted to find out whether overall sympathy of news tweets has any network effects, specifically whether higher ONS scores resulted in greater information diffusion on Twitter. Combining all data together, we tested whether our score was correlated with the number of retweets. We found a very weak yet significant negative correlation (Spearman (ρ) = -0.04, p<0.05). Given the low correlation, we had to reject our hypothesis that such sympathetic tweets would result in higher information propagation.

To understand further what kind of content was most retweeted, we retrieved the top five tweets with the highest retweets in our combined data (shown in Table 12). What is interesting to observe here is that while retweeting behavior appears to be impartial as to whether a tweet has a high ONS score or not, it does appear (within our dataset) to depend on how polarized the score is. This is likely due to the changing user information needs in the days following an attack.

# RTs	Screen Name	Text	Media	URL	ONS
9,530	@BBCBreaking	“Sydney Opera House lit up by French tricolour amid #ParisAttacks global tributes”	yes	yes	0.528
5,787	@CNN	“At least 149 people were killed in #Paris shootings and bombings, French officials said”	yes	yes	-0.886
5,570	@nytimes	“When a double suicide bombing rocked Beirut, there was no global outpouring of sympathy”	yes	yes	-0.886
5,498	@CNN	“Ten ”horrific” minutes of shooting according to this witness inside the #Paris theater: ”	no	yes	-0.886
4,850	@FoxNews	“WATCH: The Eiffel Tower went dark tonight in memory of the victims of the terrorist attacks in #Paris.”	no	yes	1.000

Table 12: Examples of the top 5 most retweeted tweets across both coverage of the Paris and Beirut attacks.

7. DISCUSSION

7.1 News Sympathy Under Crises

The selection of foreign news by domestic news media has the power media to shape the public perception about those countries [41]. Not just the selection, but also how a piece of news is reported has the capacity to evoke compassion, which could potentially lead to various charitable acts, such as fund-raising to provide monetary support [19]. When an unexpected crisis such as a terror attack on Beirut or Paris strikes, with sufficient quality coverage, it has the power to instigate collective worldwide public action.

In this work, we examined how Western and Arab media covered the Beirut and Paris attacks on Twitter. One cannot deny that the attacks in Paris were more newsworthy than the attacks in Beirut. This is what news flow theory would predict [35, 43], since Paris fulfills the criteria of a familiar, powerful foreign country, close geographically to other European states, and had an unexpected, human-induced crisis occur. This is also what the newsworthiness theory by Galtung and Holmboe [13] would predict, as newsworthiness depends on frequency, intensity, unambiguity, meaningfulness, consonance, unexpectedness, continuity of an event, and / or some unique characteristics of an actor involved. Despite the predictive power of such theories, with the growth of online communities and interactions, traditional means of news reporting can now be influenced – as exemplified by media bias outcries on Twitter and across prominent news articles.

We showed that with respect to coverage, the volume of tweet activity between Western and Arab media differed, with Western media overall reporting the attacks in Beirut less. However, if we inspect only the first day of the Beirut attacks (Fig. 4), we see that in fact there was more coverage (in normalized volume percentage) of the Beirut attacks from Western media. This becomes an interesting fact, considering that Western media on Twitter appeared to exhibit a sympathy bias.

To revisit the question that drove this research: were the Beirut attacks ignored? For coverage volume, we found that there was indeed less coverage (in terms of normalized coverage volume), possibly due to temporal proximity with the Paris attacks. However, Beirut received more coverage from Arab media. Concerning how the attacks were covered, we can state that the Beirut attacks were covered differently, based on the observed lower overall sympathy towards the Beirut attacks when it came to Western media coverage on Twitter, but not Arab media. However, we also find that while Western media was more sympathetic towards Paris, Arab media was more sympathetic towards Beirut.

Regardless of whether this could have been predicted by news flow or newsworthiness theory, our results suggest a call to action in reevaluating how overall sympathy bias may manifest in journalistic foreign news coverage. It is however important to mention here that these findings should be interpreted with respect to our modeling decisions (importantly that we factor in religious reference) and overall methodology. Finally, we were not able to observe differences in how news tweets are spread by Twitter users across the network based on overall sympathy. We find that instead, the best predictor of retweeting activity is the number of followers (Spearman (ρ) = 0.76, $p<0.001$), which fits our intuitions and understanding of how Twitter works [6].

7.2 Approach Considerations

The results of our study must be considered in light of a number of considerations and limitations to our approach. First, we have analyzed a snapshot of Twitter data in the aftermath of the Beirut and Paris attacks, where we made several decisions in how we treated the data, and later assumptions about what we took to be an adequate measurement of sympathy. However, sympathy is subjective, and a composite score that draws on other subjective elements (such as sentiment) means we are limited by these assumptions. Nevertheless, we adopted a data-driven approach to help ensure our modeling decisions were based on properties of our data, wherein the insights gained fit our intuitions and helped us interpret the data. Furthermore, it was not viable to test whether overall news sympathy differed across every country in the Western world, and compare with news coverage of every Arab country. Instead, we focused on geographic regions that made sense to test in the context of these attacks, and in this regard, our work provides an approximation of the differences in news coverage across Western and Arab media.

Another potential limiting factor is what is dubbed as the 'hostile media effect' [30], which is a perceptual theory of mass communication that refers to the tendency for individuals with a strong preexisting attitude on an issue to perceive neutral media coverage of a topic is biased against their perspective, and instead adopts the antagonists' point of view. In the context of our work, this phenomenon could have surfaced in crowdworkers, where workers from one country could have held biases when rating tweets. This is amplified by the fact that we did not conceal the screen names of news accounts (which we did to ensure the tweets were as close to real-world conditions as possible). However, given that each tweet received at least three trusted judgments, we have confidence that our annotations are likely to be untainted by such a bias. Finally, our approach is limited to Twitter data, which is shown to contain inherent biases [16]. Here, we did not test other mediums. However, the large quantity of tweets and prevalence of reactions on this divided issue of Arab versus Western news coverage provided an initial step to test our approach.

8. CONCLUSION AND FUTURE WORK

We presented a data-driven approach to tease out differences between Western and Arab Twitter media reporting of the November 2015 Paris and Beirut attacks. By combining text-mining techniques and crowdsourcing, we tested whether Western and Arab media differed with respect to coverage bias and overall news sympathy. Based on properties of our data and media statements (namely, that religious reference plays a role in sympathy), we derived the Overall News Sympathy (ONS) score, and based on this score we found that Western media tweets were less sympathetic when covering the Beirut attacks. Further, based on our labeled data, we trained a deep neural network to predict ONS scores from unlabeled data, and results further supported our ground truth analysis that each Twitter media (Western, Arab) were more sympathetic to attacks in their respective regions (Paris and Beirut, respectively).

As a more general framework, our work contributes to an understanding of Twitter media bias, and factors that may influence it, which are not necessarily limited to the studied attacks. We believe the methods we adopted are more widely applicable to other areas of computational journalism, and can serve as useful tools to better understand, expose, and design around Twitter media bias. For future work, we intend on exploring the dynamics of opinion formation and interaction when both the media and polarized individual viewpoints interact, within and across Arab cultures. We believe there is much to gain in adopting a cross-cultural lens on opinion formation, as this has far-reaching impact in shaping national and foreign public discourse on the cultural differences between Arabs and non-Arabs.

9. REFERENCES

1. Jisun An, Haewoon Kwak, Yelena Mejova, Sonia Alonso Saenz De Oger, and Braulio Gomez Fortes. 2016. Are You Charlie or Ahmed? Cultural Pluralism in Charlie Hebdo Response on Twitter. In *Proc. ICWSM '16*. 2–11. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/12997

2. David P. Baron. 2006. Persistent media bias. *Journal of Public Economics* 90, 1-2 (January 2006), 1–36. https://ideas.repec.org/a/eee/pubeco/v90y2006i1-2p1-36.html

3. Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proc. LSM '12*. ACL, 65–74.

4. Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and Network Dynamics Behind Egyptian Political Polarization on Twitter. In *Proc. CSCW '15*. ACM Press, 700–711. DOI: http://dx.doi.org/10.1145/2675133.2675163

5. D. Britz. Implementing a CNN for Text Classification in Tensorflow. (????). https://github.com/dennybritz/cnn-text-classification-tf

6. Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P. Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. In *Proc. ICWSM '10*.

7. Anna B Costello and Jason W Osborne. 2005. Best practices in exploratory factor analysis. *Practical Assess., Research & Eval.* 10, 7 (2005), 1–9.

8. D D'Alessio and M Allen. 2000. Media bias in presidential elections: a meta-analysis. *Journal of Comm.* 50, 4 (2000), 133–156. http://dx.doi.org/10.1111/j.1460-2466.2000.tb02866.x

9. Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. 2015. Media Bias in German Online Newspapers. In *Proc. HT '15*. ACM, New York, NY, USA, 133–137. DOI: http://dx.doi.org/10.1145/2700171.2791057

10. Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and Assessing Social Media Information Sources in the Context of Journalism. In *Proc. CHI '12*. ACM, 2451–2460. DOI: http://dx.doi.org/10.1145/2207676.2208409

11. Nicholas Diakopoulos and Arkaitz Zubiaga. 2014. Newsworthiness and Network Gatekeeping on Twitter: The Role of Social Deviance.. In *ICWSM '14*. AAAI.

12. Júlio Cesar dos Reis, Fabrício Benevenuto, Pedro Olmo S. Vaz de Melo, Raquel O. Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the News: First Impressions Matter on Online News., In Proc. ICWSM '15. *CoRR* (2015).

13. Johan Galtung and Mari Holmboe Ruge. 1965. The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers. *Journal of Peace Research* 2, 1 (1965), 64–90. DOI: http://dx.doi.org/10.1177/002234336500200104

14. C.J. Glynn, S. Herbst, G.J. O'Keefe, and R.Y. Shapiro. 2004. *Public Opinion*. Westview Press. https://books.google.de/books?id=b3OqBAAAQBAJ

15. Jennifer Golbeck and Derek Hansen. 2011. Computing Political Preference Among Twitter Followers. In *Proc. CHI '11*. ACM, 1105–1108. DOI: http://dx.doi.org/10.1145/1978942.1979106

16. Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Soc. Net.* 38 (2014), 16–27.

17. E.S. Herman and N. Chomsky. 1988. *Manufacturing consent: the political economy of the mass media*. Pantheon Books. https://books.google.de/books?id=Up5sAAAAIAAJ

18. Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1746–1751. http://aclweb.org/anthology/D/D14/D14-1181.pdf

19. Haewoon Kwak and Jisun An. 2014. *SocInfo*. Springer, Chapter A First Look at Global News Coverage of Disasters by Using the GDELT Dataset, 300–308. DOI: http://dx.doi.org/10.1007/978-3-319-13734-6_22

20. V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Phys. Doklady.* 10, 8 (Feb. 1966), 707–710.

21. Lin, Yu-Ru and Margolin, Drew. 2014. The ripple of fear, sympathy and solidarity during the Boston bombings. *EPJ Data Sci.* 3, 1 (2014), 31. DOI: http://dx.doi.org/10.1140/epjds/s13688-014-0031-z

22. Marco Lui and Timothy Baldwin. 2012. Langid.Py: An Off-the-shelf Language Identification Tool. In *Proc. ACL '12 System Demonstrations*. ACL, Stroudsburg, PA, USA, 25–30. http://dl.acm.org/citation.cfm?id=2390470.2390475

23. Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016. #ISISisNotIslam or #DeportAllMuslims?: Predicting Unspoken Views. In *Proc. WebSci '16*. ACM, New York, NY, USA, 95–106. DOI: http://dx.doi.org/10.1145/2908131.2908150

24. Yelena Mejova, Amy X. Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and sentiment in online news. *Symposium on Computation and Journalism* (2014).

25. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ICLR* (2013).

26. Jonathan Scott Morgan, Cliff Lampe, and Muhammad Zubair Shafiq. 2013. Is News Sharing on Twitter Ideologically Biased?. In *Proc. CSCW '13*. ACM, New York, NY, USA, 887–896. DOI: http://dx.doi.org/10.1145/2441776.2441877

27. Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In *Proc. CSCW '15*. ACM, New York, NY, USA, 994–1009. DOI: http://dx.doi.org/10.1145/2675133.2675242

28. Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135. DOI: http://dx.doi.org/10.1561/1500000011

29. Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: Delivering Multiple Aspects of News to Mitigate Media Bias. In *Proc. CHI '09*. ACM, New York, NY, USA, 443–452. DOI: http://dx.doi.org/10.1145/1518701.1518772

30. Richard M. Perloff. 2015. A Three-Decade Retrospective on the Hostile Media Effect. *Mass Communication and Society* 18, 6 (2015), 701–729. http://dx.doi.org/10.1080/15205436.2015.1051234

31. Jennifer J. Preece and Kambiz Ghozati. 2001. Experiencing Empathy Online, In The Internet and Health Communication: Experience and Expectations. *The Internet and Health Communication: Experience and Expectations.* (2001), 237–260. http://www.ifsm.umbc.edu/~

32. Vincent Price and David Tewksbury. 1997. News values and public opinion: A theoretical account of media priming and framing. *Progress in communication sciences* (1997), 173–212.

33. Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social media news communities: gatekeeping, coverage, and statement bias. In *Proc. CIKM '13*. ACM, 1679–1684. DOI: http://dx.doi.org/10.1145/2505515.2505623

34. Axel Schulz, T Thanh, H Paulheim, and Immanuel Schweizer. 2013. A fine-grained sentiment analysis approach for detecting crisis related microposts. *ISCRAM 2013* (2013).

35. Elad Segev. 2015. Visible and invisible countries: News flow theory revised. *Journalism* 16, 3 (2015), 412–428. DOI: http://dx.doi.org/10.1177/1464884914521579

36. David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. 2009. Tweet the Debates: Understanding Community Annotation of Uncollected Sources. In *Proc. WSM '09*. ACM, New York, NY, USA, 3–10. DOI: http://dx.doi.org/10.1145/1631144.1631148

37. P.J. Shoemaker and S.D. Reese. 1996. *Mediating the Message: Theories of Influences on Mass Media Content*. Longman. https://books.google.de/books?id=E_HtAAAAMAAJ

38. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

39. Kate Starbird and Leysia Palen. 2012. (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In *Proc. CSCW '12*. ACM, 7–16. http://dl.acm.org/citation.cfm?id=2145212

40. Saúl Vargas, Richard Mccreadie, Craig Macdonald, and Iadh Ounis. 2016. Comparing Overall and Targeted Sentiments in Social Media during Crises. In *Proc. ICWSM '16*.

41. Wayne Wanta, Guy Golan, and Cheolhan Lee. 2004. Agenda setting and international news: Media influence on public perceptions of foreign nations. *Journalism & Mass Comm. Quarterly* 81, 2 (2004), 364–377.

42. Zhongyu Wei, Yulan He, Wei Gao, Binyang Li, Lanjun Zhou, and Kam-fai Wong. 2013. Mainstream Media Behavior Analysis on Twitter: A Case Study on UK General Election. In *Proc. HT '13*. ACM, 174–178. DOI: http://dx.doi.org/10.1145/2481492.2481512

43. HD Wu. 2000. Systemic determinants of international news coverage: a comparison of 38 countries. *Journal of Comm.* 50, 2 (2000), 110–130. http://dx.doi.org/10.1111/j.1460-2466.2000.tb02844.x

44. Sevgi Yigit-Sert, Ismail Sengor Altingovde, and Özgür Ulusoy. 2016. Towards Detecting Media Bias by Utilizing User Comments. In *Proc. WebSci '16*. ACM, New York, NY, USA, 374–375. DOI: http://dx.doi.org/10.1145/2908131.2908186

45. Arjumand Younus, Muhammad Atif Qureshi, Suneel Kumar Kingrani, Muhammad Saeed, Nasir Touheed, Colm O'Riordan, and Pasi Gabriella. 2012. Investigating Bias in Traditional Media Through Social Media. In *Proc. WWW '12 Companion*. ACM, New York, NY, USA, 643–644. DOI: http://dx.doi.org/10.1145/2187980.2188168

46. Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR* abs/1212.5701 (2012). http://arxiv.org/abs/1212.5701