frontiers in Earth Science

ORIGINAL RESEARCH
published: 15 May 2020
doi: 10.3389/feart.2020.00141

Analysis of Online News Coverage on Earthquakes Through Text Mining

Stephen Camilleri[1]*, *Matthew R. Agius*[2,3]* *and Joel Azzopardi*[1]*

[1] *Dipartiment tal-Intelliġenza Artifiċjali, Fakultà tat-Teknoloġija tal-Informatika u l-Komunikazzjoni, L-Università ta' Malta, Msida, Malta,* [2] *Dipartiment tal-Ġeoxjenza, Fakultà tax-Xjenza, L-Università ta' Malta, Msida, Malta,* [3] *Dipartimento di Scienze, Università Roma Tre, Rome, Italy*

News agencies work around the clock to report critical news such as earthquakes. We investigate the relationship between online news articles and seismic events that happen around the world in real time. We utilize computer text mining tools to automatically harvest, identify, cluster and extract information from earthquake-related reports, and carry out cross-validation on the mined information. Earthquake parameters retrieved from the United States Geological Survey (USGS) Application Programming Interface (API) are organized into earthquake events, with each event consisting of daily earthquake readings taking place in a particular geographical location. The results are then visualized on a user-friendly dashboard. 268,182 news reports published by 23 news agencies from different parts of the world and 14,717 earthquakes of magnitude ranging from 4 to 8.2 listed in the bulletin were processed during a 1-year study between 2018 and 2019. 1.25% of the analyzed articles had the word "quake" and 0.4% were clustered and then mapped to an earthquake event. The use of multilingual news sources from 16 countries (6 languages) gives the advantage of reducing potential news bias originating from English-written reports only. The mapping of articles with an earthquake catalog helps verify earthquake reports and determine relationships. We find that the distribution of the reported seismicity is from earthquakes that occur on or very close to land. We propose a general relationship between the number of news agencies, the earthquake magnitude and the anticipated number of published articles. News reports tend to mention higher earthquake magnitudes than those in the USGS earthquake catalog, and the reports on earthquakes can last from a few days to a couple of weeks following the earthquake.

Keywords: big data and analytics, information extraction, earthquakes, news agencies, online news analysis

OPEN ACCESS

Edited by:
Remy Bossu,
European-Mediterranean Seismological Centre, France

Reviewed by:
Mark Quigley,
The University of Melbourne, Australia
Maud H. Devès,
UMR 7154 Institut de Physique du Globe de Paris (IPGP), France

***Correspondence:**
Stephen Camilleri
stephen.camilleri.03@um.edu.mt
Matthew R. Agius
matthew.agius@um.edu.mt
Joel Azzopardi
joel.azzopardi@um.edu.mt

Specialty section:
This article was submitted to Geohazards and Georisks, a section of the journal Frontiers in Earth Science

Received: *28 December 2019*
Accepted: *16 April 2020*
Published: *15 May 2020*

Citation:
Camilleri S, Agius MR and Azzopardi J (2020) Analysis of Online News Coverage on Earthquakes Through Text Mining. Front. Earth Sci. 8:141. doi: 10.3389/feart.2020.00141

INTRODUCTION

Many researchers have tried to identify the factors that determine the level of coverage news agencies give following major earthquake events (Suzanne, 2006; Eisensee and Strömberg, 2007; Stomberg, 2012; Le Texier et al., 2016). These studies do not specify that data was automatically gathered, clustered and processed for information extraction in real time and, moreover, these studies have been limited to specific earthquakes or focused on a particular geographical region. This makes it difficult to quantify how quickly news agencies react to such earthquake events, how accurate the news agencies are when reporting such events, what earthquake features are mostly mentioned in the news as well as other parameters such as how long an earthquake event remains mentioned in the news and the extent of global coverage given to the earthquake.

Early studies investigated the correlation between earthquake events and television news coverage (Adams, 1986; Simon, 1997; Van Belle, 2000). They carried out a regression analysis to determine the most important earthquake-related features that are taken into consideration when allocating TV time slots for different seismic events. The three studies identified a correlation between TV news coverage reporting on earthquake events versus the logarithm of initial number of people estimated to be killed/affected (Adams, 1986; Simon, 1997; Van Belle, 2000) and how far the catastrophic event took place from a specific city such as New York (Adams, 1986; Simon, 1997). Interestingly, some of the identified relationships seem peculiar, such as the one between the amount of news coverage given in relation to the ties the United States has with the stricken country (Simon, 1997), or the popularity of the earthquake-hit country with American tourists (Adams, 1986; Van Belle, 2000). Similarly, Van Belle (2000) identified an increase in coverage when the impacted country had better social and cultural ties with the United States, which tend to lead to increased assistance to the earthquake-affected country following the aftermath of the earthquake (Heeger, 2007). Heeger (2007) went farther, stating that while a strong correlation was found between students who followed earthquake news on TV and the financial assistance Americans provided, the relationship between the amount of time watching the event unfold and the financial contributions given, was weak.

Other studies focused on the correlation between earthquake events and newspaper coverage. For example, Suzanne (2006) analyzed 64 daily and weekly publications in 9 countries, to determine the basis on which western media opt to cover disaster-related events, as well as the difference in coverage between Europe and the United States. Suzanne (2006) sustains that there is no relationship between the severity of the natural event taking place and the media's attention. On the other hand, Adams (1986) and Eisensee and Strömberg (2007) claimed that the amount of news coverage allocated on other local current affairs that are being reported affects the news coverage given on catastrophic events.

Others state that the reporting of earthquake news is based on the devastation left by the seismic event on the affected country. Devastation is expressed in terms of damage to the infrastructure, culture, economy, labor, and environmental consequences that follow, as well as political strength that could help minimize the impact, and social adaptedness (as analyzed by Brown (2012) and Dhakal (2018) in multiple sources). Suzanne (2006) mentions that the motivation of the news coverage in the West is politically derived, i.e., to help the victims of the disaster in return for political votes to gain favorable publicity to hold key worldwide events (e.g., the Olympics). As a result, Suzanne (2006) identified a correlation between news coverage and its effect on the Western market. Culture too is considered as being a factor for American and Japanese influence in determining whether an article is written about an earthquake event according to Stomberg (2012), who analyzed American and Japanese newspapers over a period of 36 days.

More recently, the use of "tweets" (short messages on online social media platform Twitter[1]) assisted in providing more real-time information on the geographical region of the "news" where the earthquake took place and the amount of structural damage caused (Earle et al., 2010, 2012; Sakaki et al., 2010; Liang et al., 2013; Avvenuti et al., 2015; Bossu et al., 2015; Hicks, 2019). Panagiotou et al. (2016) noticed a relationship between the time Twitter users tweeted and the time the event took place, highlighting the importance of social media users acting as news collaborators. Liang et al. (2013) analyzed how fast the news spread over a period of 90 minutes and managed to identify a correlation between retweet densities and tweeting count per user versus the distance from the earthquake's epicenter. Similarly, Avvenuti et al. (2015) studied the relationship between the earthquake's magnitude and how tweets are spread around the geographical area hit by the earthquake. They took into consideration unique Twitter users and the mean value of tweets submitted following the earthquake event. However, this highly depends on the population of the area (Earle et al., 2010). Other studies have investigated user traffic on dedicated earthquake websites to gauge the interest of the general public (Bossu et al., 2008, 2014) and also to understand how long the general public stay interested (Quigley and Forte, 2017).

Recently, Devès et al. (2019) analyzed the articles published by worldwide newspapers in 2015 in English, Spanish and French. They found that the press covered a very small number of earthquake events. Coverage was mostly dedicated to 3 major earthquakes that happened in that year (e.g., Nepal). They found that the duration of the coverage was very short, with news focus on short-term issues: the event magnitude, tsunami alerts, human losses, material damage and rescue operations. Longer-term issues linked to the recovery, restoration, reconstruction, mitigation and prevention were barely addressed.

Our study aims to automate and run in real time the entire process to investigate relationships between earthquakes and online news coverage. It also aims to decrease the potential bias from online news reports arising from a limited number of sources (typically those in the English language and associated with the western world) and from a limited focus on specific earthquake events or geographical region. This is done by (i) harvesting data from as many online news sources as possible by downloading directly from their respective websites or through Application Programming Interface (API) or Really Simple Syndication (RSS) endpoints; (ii) use text mining tools to identify, cluster and extract information from earthquake-related news; (iii) cluster daily earthquake parameters retrieved from the United States Geological Survey (USGS) international seismological bulletin API endpoint into earthquake events based on the geographical location; (iv) map the news clusters and earthquake locations in near real time, and (v) provide analysis on the results and possible relationships between news coverage and earthquake events. Novel approaches adopted here are that the software filters, clusters, and mines information from news

[1] http://www.twitter.com

sources automatically and in real time; and that the sources are from several, multilingual websites, which will help reduce potential bias originating from English written newspapers only.

DATA AND METHOD

The main objective in the algorithm is to automatically create a dataset of harvested news articles mapped to actual earthquakes that have occurred. This is done by (i) cleaning and translating news articles to English; (ii) identifying news articles referring to earthquakes; (iii) grouping news articles referring to earthquakes into clusters; (iv) extracting earthquake parameters such as the date, location, magnitude and other parameters such as number of casualties and quantifiable structural damage; and (v) cross-validating the information extracted from multiple articles across each cluster.

The prototype, named QuakeNews Analyser, was programmed to download news published on the websites of news agency and traditional newspapers through the API/RSS endpoints and from hyperlinks within the endpoints to scrape the content from news agencies websites. All the harvested news content is initially pre-processed by applying text cleaning techniques (Guy et al., 2010; Oh et al., 2010; Piskorski and Atkinson, 2011; Azzopardi and Staff, 2012; He et al., 2013; Asghar et al., 2014; Khumoyun et al., 2016). This typically entails discarding headers, footers, embedded images and JavaScript, as well as removing special characters and HTML tags. The language of the news content is also detected and translated to English if the language is otherwise. Translated news articles containing the word "quake" were then identified for clustering on the basis of a keyword-based search, while the rest of the articles were stored for statistical purposes. From manual evaluation, when alternate words such as "seismic," "shaking" etc. were used, the word "quake" or "earthquake" was also found in the text (as it commonly referred to by the general public). News articles were filtered using the bag-of-words model (frequency of the words in a news article) and words that were weighted. In this way, words which appear frequently in many articles (words such as "the," "so," and "there") were given less importance than those which were explicitly found in certain articles. Earthquake attributes from the text written in each article were then automatically extracted and articles were grouped into clusters using the Term Frequency-Inverse Document Frequency (TF-IDF) and No-*K*-Means approaches. TD-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents, and No-*K*-Means is used in the analysis of data mining by partitioning observations into clusters based on a similarity threshold value. These rigorous approaches were aimed to minimize the possibility that a news cluster contains articles reporting different earthquakes.

Information extraction is then carried out on each of the translated clustered articles, where six common earthquake-related features were extracted: the event magnitude, the date and geographic location of the event, the number of casualties, people injured, and structural quantifiable damage caused by the seismic event. The values of the features extracted from all the articles within each cluster were then cross validated. In order to validate the news articles, a list of earthquakes is compiled from those issued by USGS and aggregated into a list of events in such a way that an event can represent multiple earthquake readings taking place on the same day within the same country. A minimum earthquake magnitude threshold of 4 was chosen because this is the typical lower-bound magnitude of felt earthquakes (e.g., Coburn and Spence, 2002), and thus with a higher likelihood that moderately sized earthquakes were reported in newspapers. Each earthquake event from USGS is then mapped against an extracted magnitude, range of dates and list of locations retrieved from each news cluster (containing one or many news articles mentioning the same earthquake event).

There are computational challenges and limitations when exploiting the text mining tools (e.g., Radinsky et al., 2012; Asghar et al., 2014), particularly problems when extracting information from unstructured sources such as websites (Vannella et al., 2014). One challenging problem arises from the use of part of speech taggers, including incorrect tagging of words, vague classification of entities (Asghar et al., 2014; Jurafsky and Martin, 2014), language-related issues (Pinto et al., 2016), extraction of temporal expressions (Mani and Wilson, 2000; Kisilevich et al., 2010; Bögel et al., 2014; Derczynski and Gaizauskas, 2015) and geographical locations (Kisilevich et al., 2010; Piskorski and Atkinson, 2011) as discussed by Gupta (2016). Another case is with interpreting the content written by different journalists especially when using technical terms (Liu, 2010), the translation of words denoting numerical values into numbers (Miner et al., 2012), and the mapping of articles referring to earthquakes with the characteristics provided by USGS (Le Texier et al., 2016). Another limitation the current prototype has is the 5,000 character limit imposed by Google translator library (GoogleTrans) on identifying the language of the news article and translating it. Generally, the number of words for a text made up of 5,000 characters is approximately 500 to 1,000 words. Studies have shown that key facts are usually placed in the beginning of the content (e.g., Bell and Garrett, 1998; Tanev et al., 2008; Piskorski and Atkinson, 2011). It is anticipated that at least the word "quake" and other relevant information (date, location, magnitude, damage, etc.) are mentioned in the first 5,000 characters of the translated articles.

Depending on the complexity of the texts, articles may result being mapped to the wrong earthquake event. Rigorous text analysis are in place to cross validate such cases; however, this resulted in a reduced amount of mapped articles with listed events. A detailed technical description of the algorithm and tests performed on the datasets are in Camilleri et al., 2019 and Camilleri, 2019.

In summary, we extracted news reports from 23 international news agencies in 6 different languages (**Table 1**). The news sources were chosen based on whether the data could be retrieved from RSS/API for free, on the popularity of the news agency, and on which country the news agency is focused so as to widen the coverage of news articles published across six continents. We ran the prototype in real time for a period of 12 months from the 10th of January 2018 to the 10th of January 2019. A total of 14,717 earthquakes listed in the USGS bulletin with

TABLE 1 | List of news agency internet sites mined for data.

Country	Source name	Data provider	RSS/API
Argentina	La Nacion (local)	La Nacion	RSS
Argentina	La Nacion (world)	La Nacion	RSS
Australia	ABC (local)	ABC	RSS
Australia	ABC (world)	ABC	RSS
Brazil	Grupo Globo (local)	Grupo Globo	RSS
Brazil	Grupo Globo (world)	Grupo Globo	RSS
Chile	Soy Chile (local)	Soy Chile	RSS
Chile	Soy Chile (world)	Soy Chile	RSS
China	China Daily (local)	China Daily	RSS
China	China Daily (world)	China Daily	RSS
China	Shanghai Daily (local)	Shanghai Daily	RSS
China	Shanghai Daily (world)	Shanghai Daily	RSS
China	Xinhua Net	Google News	API
Germany	Die Zeit	Die Zeit	API
India	India Today (local)	India Today	RSS
India	India Today (world)	India Today	RSS
Italy	TGCOM24 (local)	TGCOM24	RSS
Italy	TGCOM24 (world)	TGCOM24	RSS
Japan	Japan Times	Japan Times	RSS
Qatar	Al Jazeera	Google News	API
Russia	TASS	TASS	RSS
South Africa	News24 (local)	News 24	RSS
South Africa	News24 (world)	News 24	RSS
Spain	El Mundo	Google News	API
Spain	El Pais (local)	El Pais	RSS
Spain	El Pais (world)	El Pais	RSS
Sudan	Sudan Tribune	Sudan Tribune	RSS
United Kingdom	BBC	Google News	API
United Kingdom	Metro End	Google News	API
United Kingdom	Reuters	Google News	API
United Kingdom	The Guardian	The Guardian	API
United States	Associated Press	Google News	API
United States	USA Today	Google News	API

a magnitude between 4 and 8.2 were retrieved and aggregated into a list of 7,359 earthquake events. Many of these earthquake events include aftershocks (or sequence of earthquakes) that happen on the same day in the same country. At the same time, a total of 268,182 articles were analyzed; 3,355 articles (1.25%) had the word "quake" of which 1,042 articles (0.4%, or 31% of 3,355) were grouped into clusters and mapped to the earthquake events. These resulted in successfully mapping 698 earthquake events with articles. Here we discuss the outcome of the results in the context of earthquakes and the relationship with the online news media.

RESULTS

This study has widened considerably the source of news coverage for investigation when compared to previous studies reaching out to news agency websites spread across 16 different countries from six continents (North America, South America, Europe, Africa, Asia, and Australia - refer to **Table 1**) and translating 6 languages to English in the process. The study produced some interesting observations that shed more light on how news agencies reacted with recent earthquakes.

We compared the amount of earthquake reporting between news agencies. Out of the 3,355 articles containing the word "quake," Japan Times, TGCOM24, ABC, and El Pais wrote the highest number of articles: 16.4, 11, 10.5, and 10.3% (**Figure 1**). The most reported earthquake-feature in the extracted news articles was the magnitude parameter (22%), followed by the number of casualties (14.9%), quantifiable structural damage caused (5.8%), and the number of injuries (5.7%).

One has to be careful not to over interpret these numbers. Some agency portals provided users with two different API/RSS endpoints – one disseminating local news and the other disseminating international news (e.g., La Nacion, ABC, and Grupo Globo), while others publish the local and international news through one API/RSS endpoint (e.g., Die Zeit, Japan Times, Al Jazeera, and The Guardian). QuakeNews Analyser is programmed to extract data from both sources when available, which may result in some agencies publishing more reports that

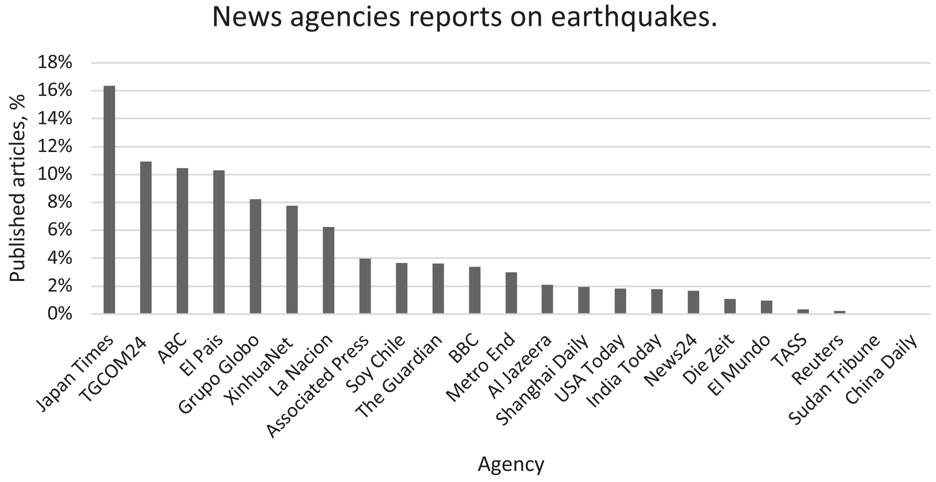


FIGURE 1 | Comparison of published earthquake-related articles per news agency.

include many local earthquakes than other agencies who report only international earthquakes.

Figure 2 shows the distribution of earthquake epicenters from the analyzed reports, and the location of the reporting online news agencies. The distribution of the reported seismicity is from earthquakes that occur on or very close to land with most of the earthquakes that happen along mid-ocean ridges not reported. This follows the trend observed in Le Texier et al. (2016).

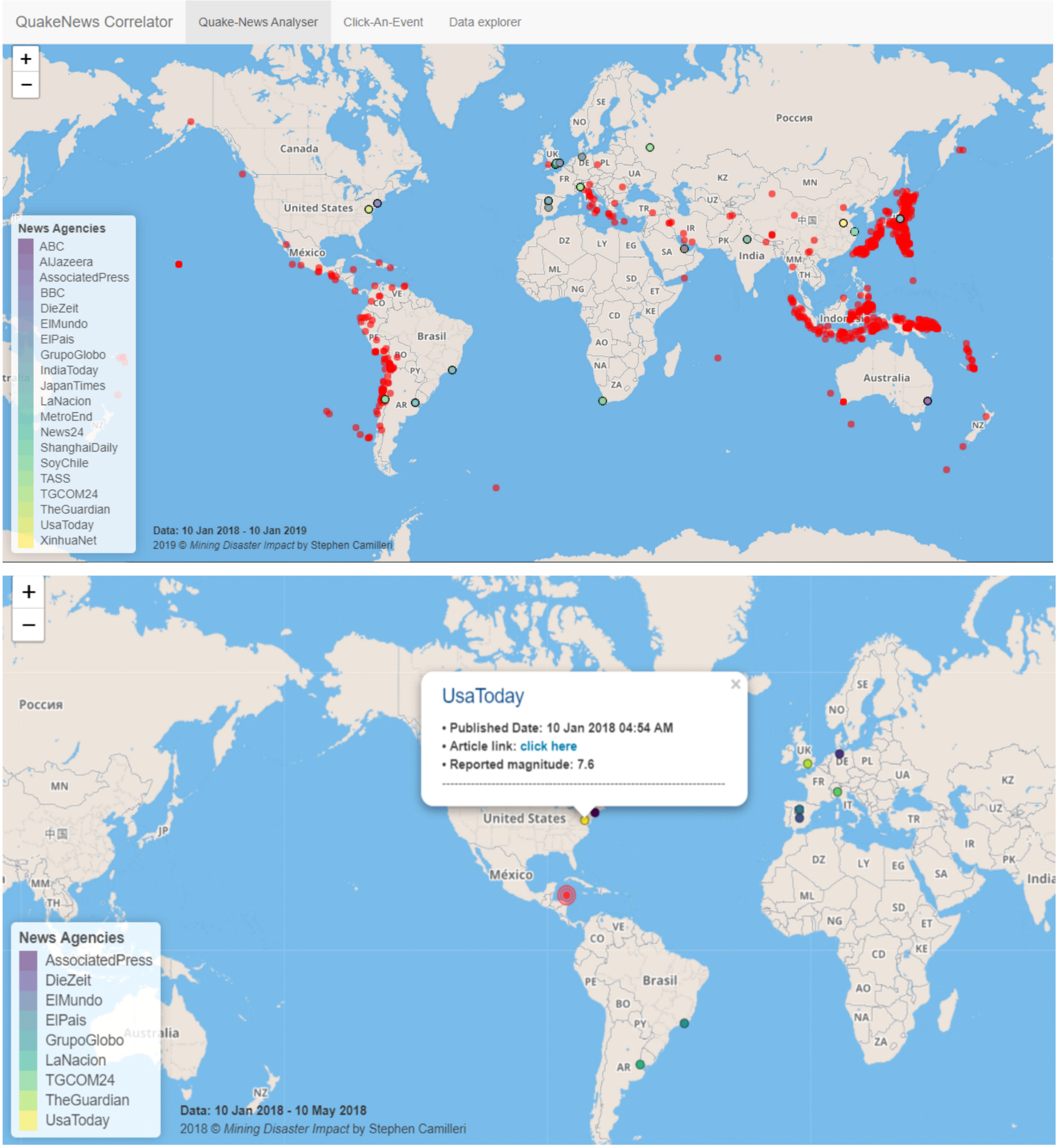
Figure 3 shows the timeline of detected reports binned in weeks. The number of published online articles per week are on the same order of magnitude (tens to hundreds) of those reported by Devès et al. (2019) for daily media coverage (with the exception of the Nepal earthquake in reference). Spikes in the number of news articles coincide with notable earthquakes such as the week following the 7.5 magnitude earthquake on September 28, 2018 near Palu, Indonesia, which also triggered a devastating tsunami (weeks 39–41). Similarly, but for a smaller peak was the Hawaiian earthquake on May 4, 2018, with a magnitude of 6.9, which also involved a series of volcanic eruptions over a number of weeks (weeks 18–24). Another peak in the number of published articles at the end of the year is also related to a volcanic eruption, this time from Etna in Italy, following a series of earthquakes. Interestingly, the greatest earthquake for 2018 had a magnitude of 8.2 on August 19 (week 34) located beneath Fiji, however this was under reported probably because of its very deep hypocenter (>500 km) which led to a few numbers of local felt reports and minimal risk to generate a tsunami. Other deep earthquakes (>100 km depth) are mostly not reported (e.g., during the Hawaii volcanic eruption, **Figure 3**).

The discrepancy between the number of articles which had the word "quake" and the number of articles mapped to an earthquake event from USGS (**Figure 3A**) is mainly due to reports mentioning earthquakes in a vague or general context making it difficult for the article to be mapped. For example, the Hawaiian earthquake was followed by articles reporting the volcanic activity and briefly mentioning the seismic activity leaving out crucial details that enable mapping that article with an earthquake event. On the other hand, thanks to the rigorous parsing of articles that extracts key earthquake parameters (e.g., location, time, and magnitude) we focus our results on the mapped articles, making it possible to investigate relationships of published articles with magnitude.

In general, the expected trend of an increased number of published articles for the larger magnitude earthquakes holds (e.g., Le Texier et al., 2016) keeping in mind that there are far less earthquakes of higher magnitude. In **Figure 4A** we divide the number of articles with the number of global earthquakes of 2018 in the respective magnitude category, clearly showing a higher ratio of published articles for the larger magnitude earthquakes. We propose a power-law relationship between the number of news agencies (A), the earthquake magnitude (M), and the anticipated number of published articles during the 1-year study:

$$Predicted\ articles\ =\ A^{(M/M_t)}$$

where M_t is the minimum earthquake magnitude threshold reported (**Figure 4B**). Unlike Le Texier et al. (2016), who developed a model to explain the number of mentions each earthquake of magnitude greater than 5 gets by its geophysical characteristics (earthquake magnitude, depth, localization and concentration), our model takes into consideration the earthquake magnitude of any size as well as the number of news agencies available. Thus, for typical minimum reported earthquake magnitude M_t of 4 and 23 news agencies one would expect about 23 articles for earthquakes of this magnitude. Similarly, for earthquakes of magnitude 7 or more one would expect about 240 news reports annually. These accumulative reports could either be from unique news agencies or perhaps multiple reports from a fewer number of agencies. This simple, direct relationship assumes that the global distributed seismicity for the various magnitude ranges remains the same. The



Data: 10 Jan 2018 - 10 Jan 2019
2019 © *Mining Disaster Impact* by Stephen Camilleri

UsaToday

- Published Date: 10 Jan 2018 04:54 AM
- Article link: click here
- Reported magnitude: 7.6

Data: 10 Jan 2018 - 10 May 2018
2018 © *Mining Disaster Impact* by Stephen Camilleri

FIGURE 2 | Graphic user interface of QuakeNews Analyser. Top map shows the location of 698 earthquake epicenters (red dots), and the location of news agencies' headquarters who reported about these events (color shaded dots). Bottom map shows an example of the news coverage for a magnitude 7.6 earthquake that occurred on Great Swan Island off the coast of Honduras on the 9th of January 2018 (red marker) (see also **Figure 3**). Colored dots show the agencies which reported the event. Tab shows details of a published article by USA Today.

relationship is likely to be a lower bound because of our limited dataset, strict parsing of articles and careful mapping of articles with earthquake events. Furthermore, the application does not take into consideration the republishing of articles via the diverse media streams and use of online social media.

DISCUSSION

Earthquakes and Their Coverage

The majority of events during the study period took place along the seismically active Pacific Rim, in the regions of

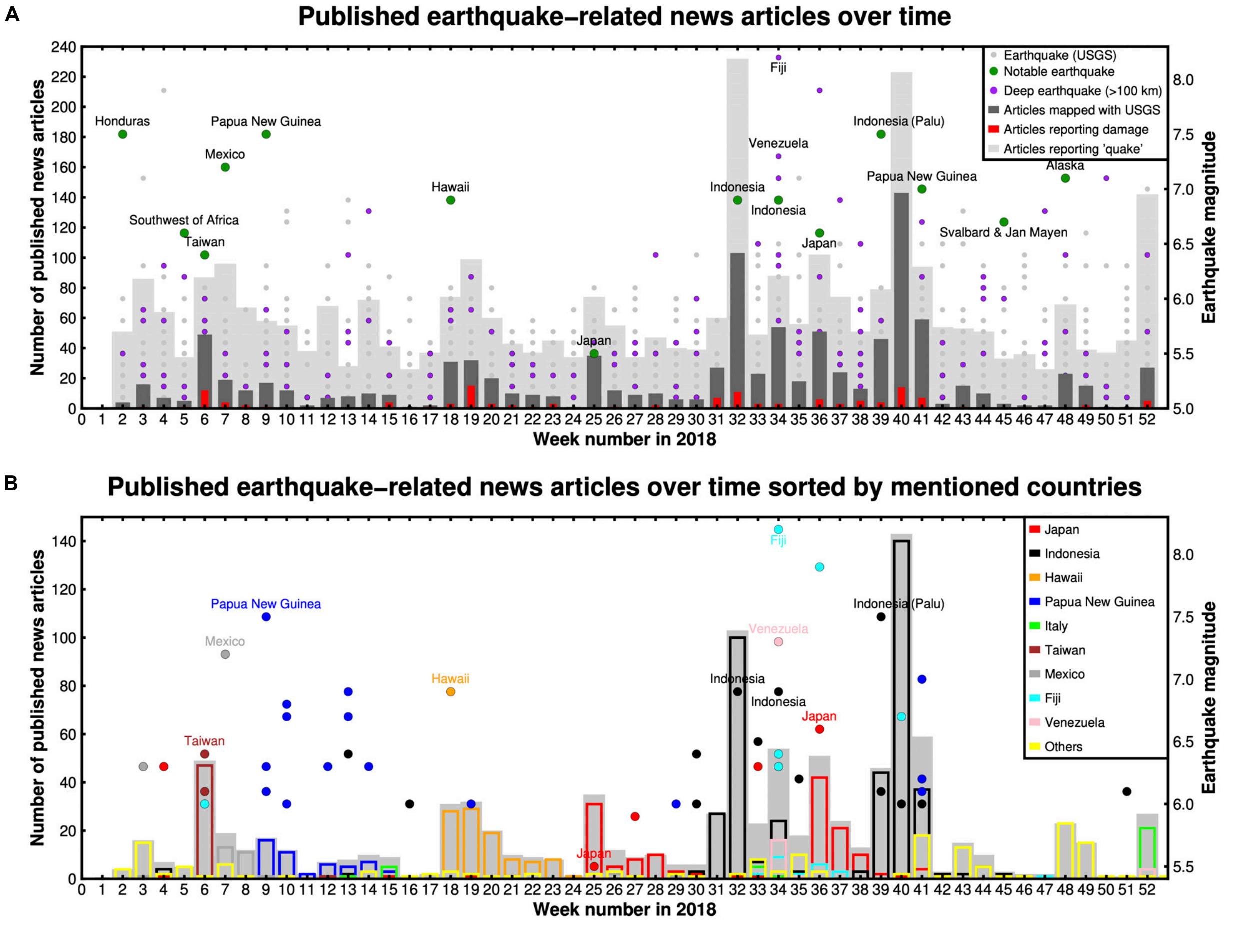


FIGURE 3 | Weekly bins of detected earthquake-related news articles during 2018. **(A)** Light gray bars represent the number of articles which had the word "quake" in the text (3,355). Dark gray bars represent the number of articles which were mapped with an earthquake event (1,042). Red bars represent the number of articles which reported earthquake damage. Green dots are notable earthquakes corresponding to peaks in the bins or discussed in the text. Purple dots are deep earthquakes (> 100 km) with strong magnitudes. Gray dots are other earthquakes listed in the USGS catalog that occurred during the year. **(B)** Same as top graph but sorted by color-coded countries referred to in the articles. Dots are earthquakes color-coded by country and have a magnitude ≥ 6 (except for Japan weeks 25–30 with lower magnitude M ≥ 5.5). Note that the *y*-axes have been expanded for better viewing, and that some events may happen at the end of a week resulting in articles flowing onto the next week bin.

Japan (282 earthquake events), Indonesia (155 seismic events), Papua New Guinea (85 earthquake events), and Chile (57 seismic events), making up 83% of the seismic events (579 out of 698 earthquake events) which were mapped to news clusters (**Figure 2**). One of the most mentioned single event was the magnitude 7.5 earthquake in Papua New Guinea that occurred on the February 25, 2018 (week 9), with 50 articles over a span of 48 days (7 weeks, **Figure 3B**). Similarly, Taiwan's magnitude 6.4 earthquake on the 6th of February 2018 was mentioned in 47 articles, however, over a span of 5 days only.

Not all major earthquakes got the same news coverage. For example, the magnitude 7.1 earthquake that took place in Anchorage, Alaska on the 30th of November 2018 (week 48, **Figure 3**), did not receive as much coverage despite the many aftershocks over the following days. In all, the event was mentioned in 15 articles, with news agencies reporting a magnitude ranging between 5.7 and 7.0. One possible reason for such lack of coverage is that Alaska is sparsely populated and that there were no casualties even though a few buildings suffered structural damage. In other cases, strong earthquakes did not get any news coverage at all. For example, the magnitude 6.6 earthquake in Southwest of Africa on the 28th of January 2018 (week 5, **Figure 3A**) and the magnitude 6.7 in Jan Mayen on 9 November 2018 (week 45) are two such cases. The most likely reason for the lack of coverage is that the events took place in uninhabited, remote areas.

While many news agencies are quick to report an earthquake and update readers with the latest news within a few hours of the earlier reports (e.g., **Figure 5**), levels of interest from news agencies on the aftermath of the earthquake varies. The trend on the duration an earthquake event is reported correlates with the damage caused rather with the magnitude of the earthquake. Very large magnitude earthquakes are at times either reported

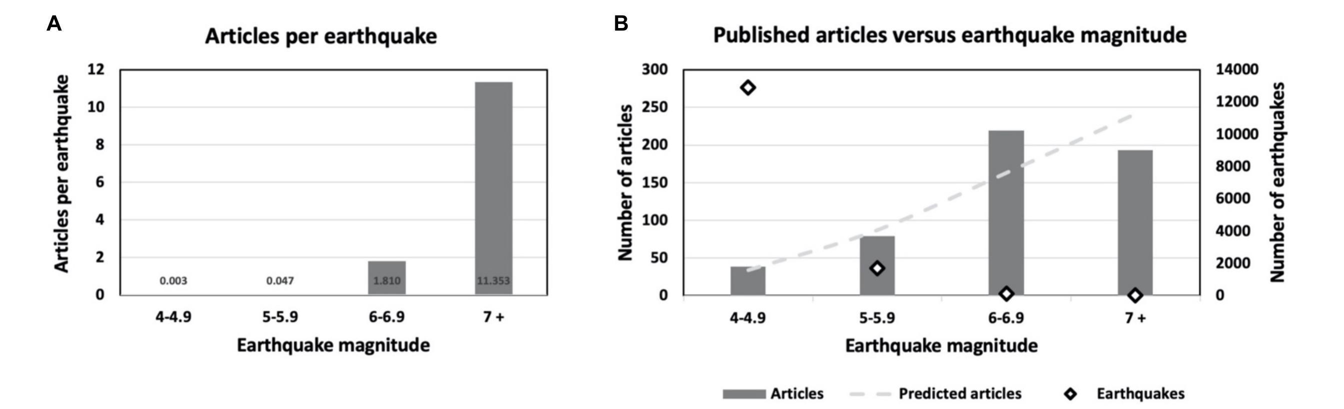


FIGURE 4 | Graphs showing the relationship of published articles with the number of global earthquakes and earthquake magnitude. **(A)** Ratio of number of articles per earthquake sorted by magnitude. **(B)** Number of articles with the corresponding magnitude ranges. Dashed line is the modeled number of articles (Correlation coefficient of 0.88, and R-squared value of 0.78). Diamond symbols show the number of earthquakes listed in the 2018 USGS catalog.

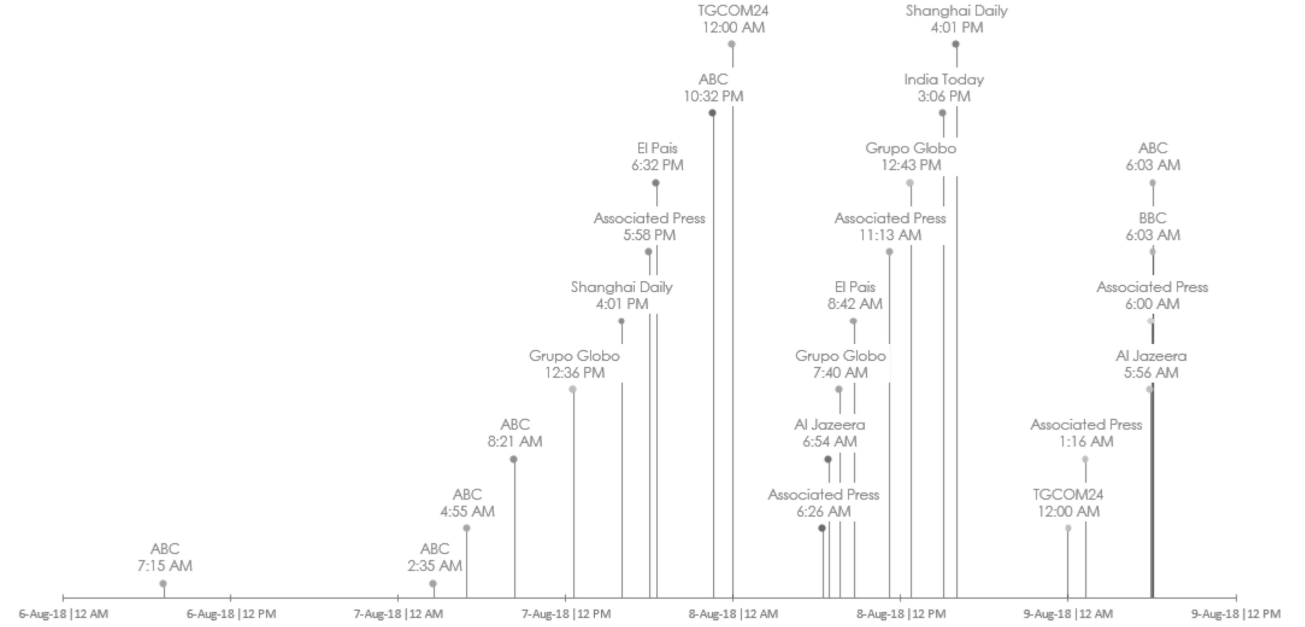


FIGURE 5 | Timeline showing the reports from different news agencies as Indonesia was experiencing earthquakes frequently. The date represents the timestamp when the article was published by the news agency.

briefly or none at all (**Figure 3**), whereas earthquakes with a lower magnitude but cause some damage tend to have a temporal decay that lasts from a few days to a few weeks, example: Taiwan 1 week, Mexico 2 weeks (7–8), Hawaii 7 weeks (18–24). During 2018 there were two cases of long-enduring earthquake sequences which kept the attention of the media, one in Papua New Guinea and one in Indonesia. In the case of the latter, the seismic activity took place between the 30th of June and the 12th of October 2018, with reports continuing over the following weeks (week numbers 26–45, **Figure 3B**). Throughout this period, 98 earthquake events out of 104 earthquakes listed by USGS were mapped against at least one news cluster, with the number of news articles published by news agencies totaling 400. During the three and a half months there was a significant earthquake recorded almost every day somewhere in the country, of which 42 earthquakes were equal or above magnitude 5, and 6 were above magnitude 6. Probably, the extensive coverage was due to the seismic activity claiming thousands of lives, injuring thousands of people and causing a large amount of structural damage (**Table 2**). **Figure 5** shows an example of a timeline of some extracted reports between the 6th to the 8th of August 2018 during the earthquake sequence.

Earthquake Magnitude

In the case of reported earthquake magnitudes, in general, they either match those listed by USGS or are higher (**Figure 6**).

TABLE 2 | Example of news reports when Indonesia was frequently experiencing earthquakes.

Source	Title	Date	Magnitude	Damage	Link[1]
ABC	"People were screaming": Witness describes chaos when quake hit Gili Islands	06-08-2018 07:15:42	6.9		http://www.abc.net.au/news/2018-08-06/witness-describes-chaos-when-earthquake-hit-the-gili-islands/10078808
ABC	How the Lombok earthquake happened	07-08-2018 04:55:39	9.1		http://www.abc.net.au/news/2018-08-07/what-creates-quake-risk-on-lombok/10082912
ABC	Survivors pulled from rubble in Indonesia's quake-hit Lombok	07-08-2018 22:32:58	7	230	http://www.abc.net.au/news/2018-08-08/rescuers-pull-people-out-alive-from-rubble-in-indonesias-lombok/10087980
TGCOM24	Terremoto in Indonesia, si aggrava il bilancio: morti salgono a 347	08-08-2018 00:00:00	6.9		http://www.tgcom24.mediaset.it/mondo/terremoto-in-indonesia-si-aggrava-il-bilancio-morti-salgono-a-347_3156866-201802a.shtml
Associated press	Food, aid reaching quake-stricken parts of Indonesian island	08-08-2018 06:26:58	7		https://apnews.com/e481d46a399c4d5b83ae96cff5df7193
Associated press	The Latest: Death toll rises to 131 in Indonesian quake	08-08-2018 11:13:30	7	156000	https://apnews.com/6abd9ea6f3f04bad8ed83590adebcc1f
Shanghai Daily	Quake leaves 156,000 homeless	08-08-2018 16:01:00	6.9	131	http://www.shanghaidaily.com/world/Quake-leaves-156000-homeless/shdaily.shtml
TGCOM24	Indonesia, nuovo forte terremoto sull'isola di Lombok: magnitudo 5.9	09-08-2018 00:00:00	5.9		http://www.tgcom24.mediaset.it/mondo/indonesia-nuovo-forte-terremoto-sull-isola-di-lombok-magnitudo-5-9_3156949-201802a.shtml
Associated press	Quake put life on hold in damaged, hungry Indonesian village	09-08-2018 01:16:06	7		https://apnews.com/924cdf5ef27a47cdbd21138d7aecddbe
ABC	Another strong quake hits Indonesia's Lombok, witnesses say buildings have collapsed	09-08-2018 06:03:35	6.2		http://www.abc.net.au/news/2018-08-09/another-strong-quake-hits-indonesia-lombok-buildings-col/10102422

[1] *Internet links last accessed on December 16, 2019. Entries correspond to the timeline in* ***Figure 5****. The date represents the timestamp when the article was published by the news agency.*

For example, the magnitude of the earthquake that struck the Great Swan Island off the coast of Honduras on the 9th of January (**Figure 2**) was reported as 7.6 by United States Today, whereas the USGS magnitude is 7.5 (**Figure 3**). There are various factors that may influence this outcome. For instance, many news agencies quote earthquake magnitudes published by USGS, while others report the earthquake magnitude from other seismic monitoring agencies. Different earthquake agencies are likely to report slightly different magnitude values either because of different type of estimate (e.g., local, body wave, surface wave, and moment magnitude) or because using different parameters such as the number of seismic stations (e.g., Chung and Bernreuter, 1981; Kanamori, 1983). In other instances, some reporters tend to give rounded "ceiling" values for the earthquake magnitude reporting a higher than the actual value. Another reason could be the news agencies tendency to exaggerate the news of small-magnitude earthquakes to attract the public's interest (e.g., Dhakal, 2018). Also, the earthquake magnitudes are sometimes revised by seismologists resulting in a different value other than what was initially reported by USGS. In the case of aftershocks, reporters sometimes remind the readers of the larger magnitude of the earthquake sequence rather than the magnitude of the recent earthquake. Additionally, QuakeNews Analyser might extract the wrong magnitude when a past event is mentioned in the article as is the case for the Lombok earthquake reported by ABC, where the software picked a magnitude of 9.1, which was referring to the 2004 Indian Ocean tsunami (**Table 2**).

It has also been noted that the magnitude is not always reported in news articles because news editors uses arbitrary words such as "strong" to describe the energy of the earthquake particularly for reporting on past events. Also, the magnitude parameter may have been missed because of the limited number of words allowed for translation by the Google API, adopted by xthis prototype.

Earthquake Damage

With regards to quantifying damage caused by an earthquake solely based on reports, it is a challenging task. One has to first define what is "damage." It can take different forms such as deaths, displaced people, collapsed buildings, etc. Thus, unlike the magnitude, which is usually described with a number gauged by a seismograph and is one of the most reported parameters to describe an earthquake, one has to look for specific words to capture the context of the report and determine the level of damage. Secondly, reports on damage may change from day to day following an earthquake as initial reports tend to report "minimal" amount of damage. Then, as the story unfolds, the damage is better assessed however

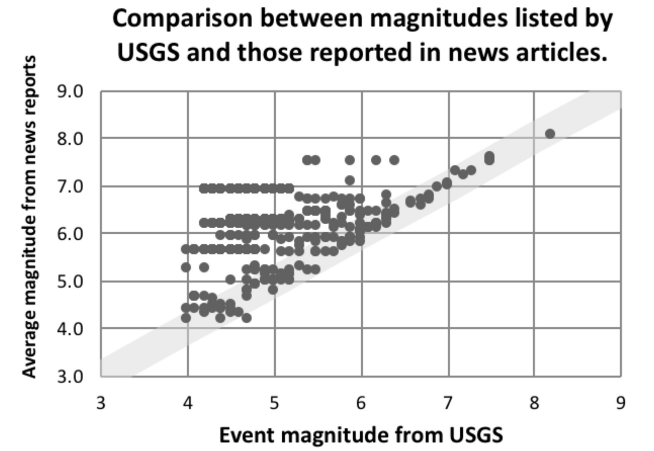


FIGURE 6 | Comparison between magnitudes published by USGS and the average of magnitudes reported by different agencies. The thick gray line represents the expected 1-to-1 relationship.

fewer agencies continue reporting details. Furthermore, these latter reports may lack earthquake details that make mapping the article to the original earthquake event difficult. A more comprehensive algorithm than that presented here, which takes into account a wider range of grammar and words related to damage, is necessary to extract complete information on damage. Nonetheless, in **Figure 3**, we show the trends of articles which report damage. The number of detected reports mentioning damage are much less than those reporting on the earthquakes however they follow similar trends in the peaks and coincide with earthquakes. The number of articles mentioning damage related to an earthquake event can last from 1 week to couple of weeks following a significant earthquake (e.g., Taiwan, Mexico, and Hawaii, **Figure 3**), and contribute to the duration on how long the event is mentioned in the press (see section "Earthquakes and Their Coverage"). **Table 3** summarizes the number of reported casualties and the number of reporting articles also sorted by magnitude. The devastating earthquakes in Indonesia during 2018 dominated the list of casualties.

Alternative Uses

Enhanced applications of tools such as this can have alternative uses of far more important implications than just statistical analysis aimed at finding which earthquake is being reported. For example, it can be used to automatically map the felt intensity of an earthquake-struck region based on the macroseismic scale, traditionally limited to people filling in local felt report forms (e.g., Wald et al., 2012; Bossu et al., 2016; Van Noten et al., 2017). Another use could be to detect mistaken news reports about wrongly reported seismic activity such as the case with Kenya's "crack" in 2018[2,3,4], whereby the reports can be validated automatically with earthquake bulletins. Such applications can also be used for other natural disasters such as tsunamis, hurricanes, forest fires, etc., simply by changing the searched keywords.

Another alternative use could be for global campaigns aimed at relief efforts as might be necessary in the case of large-scale disasters (earthquake or otherwise). The spread of news reports across the world is indicative of the attention the disaster brought on to the world. In general, the international community responds positively to calls for international aid and provide support to the affected community, however, the focus on the devastation fades out as new stories catch the media's attention (e.g., Devès et al., 2019). Thus, the gathering of information on news reporting in real time can help plan campaigns for relief efforts or fundraising particularly when the news reports are declining but the attention is still necessary. Similarly, such a tool can provide useful information to keep on-going educational campaigns for rare, large-scale disasters like tsunamis, which may need the occasional

[2] https://face2faceafrica.com/article/africa-splitting-two-tear-kenyas-rift-valley-video (last accessed March 12, 2020)

[3] https://www.theguardian.com/science/blog/2018/apr/06/africa-is-slowly-splitting-in-two-but-this-crack-in-kenya-rift-valley-has-little-to-do-with-it (last accessed March 12, 2020)

[4] https://www.forbes.com/sites/davidbressan/2018/04/05/seismologists-are-not-happy-how-media-reported-the-kenya-crack/ (last accessed March 12, 2020)

TABLE 3 | Table shows the number of casualties (when extracted) and the related number of articles published, and also sorted by reported magnitude.

Number of casualties	Number of articles, which extracted casualties value	Number of articles (magnitude not mentioned)	Number of articles (M 4–5)	Number of articles (M 5–6)	Number of articles (M 6–7)	Number of articles (M 7–8)
1	10	1	0	0	2	7
1–10	66	13	1	1	42	9
10–100	105	28	0	4	40	33
100–1000	80	31	0	4	21	24
1000–10000	49	31	0	2	2	14
10000–100000	2	1	0	0	1	0
100000+	3	0	0	0	1	2

Initial reports tend to report minimal number of damage.

article as part of maintaining a good level of preparedness within the community in general.

CONCLUSION

We investigate the relationship between earthquakes and online news portals. We use an inhouse developed software that automatically extracts earthquake reports in real time from 23 news agencies and authored in 6 different languages available at the time in order to reduce bias from reports. Out of 268,182 articles collected during a 1-year time period, 1.25% had the word "quake" and 0.4%, were mapped to the earthquake events listed in the USGS earthquake bulletin to validate its authenticity and establish relationships.

We find that the distribution of the reported seismicity is from earthquakes that occur on or very close to land with most of the earthquakes that happen along mid-ocean ridges not reported, as has been noted in previous studies. Our results also confirm that the number of published articles online about an earthquake depends on its magnitude, on the duration of the seismicity which can be linked to the same main shock and, of course, on the number of news agencies considered. Based on the news agencies and reports analyzed here, we propose a lower bound relationship between the number of news agencies, the earthquake magnitude and the anticipated number of published articles online in a year. We find that, in general, reports mention higher earthquake magnitudes than those in the USGS earthquake catalog, and the reports on earthquakes can last for a few days to a couple of weeks following the earthquake.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

SC has worked on the design of the software, acquired the data, made the analysis, worked on the interpretation of the data, and wrote the manuscript. MA has conceived the project, contributed to the analysis and interpretation of the data, and revised the manuscript. JA contributed to the design of the work, data acquisition, analysis and interpretation of the data, and revised the manuscript.

FUNDING

MA has received funding from the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie Grant Agreement No. 843696.

ACKNOWLEDGMENTS

The authors would like to thank two reviewers and Pauline Galea for helping improve the manuscript. Some figures were made using Generic Mapping Tools software (Wessel et al., 2013; Agius, 2018).

REFERENCES

Adams, W. C. (1986). Whose lives count? TV coverage of natural disasters. *J. Commun.* 36, 113–122.

Agius, M. R. (2018). "Getting started with GMT: an introduction for seismologists," in *Moment Tensor Solutions*, ed. S. D'Amico (Cham: Springer Natural Hazards, Springer). Available at: https://link.springer.com/chapter/10.1007/978-3-319-77359-9_31

Asghar, M. Z., Khan, A., Ahmad, S., and Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. *J. Basic Appl. Sci. Res.* 4, 181–186.

Avvenuti, M., Del Vigna, F., Cresci, S., Marchetti, A., and Tesconi, M. (2015). "Pulling information from social media in the aftermath of unpredictable disasters," in *2015 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, Piscataway, NJ: IEEE, 258–264.

Azzopardi, J., and Staff, C. (2012). Incremental clustering of news reports. *Algorithms* 5, 364–378.

Bell, A., and Garrett, P. D. (1998). *Approaches to Media Discourse*. Hoboken, NJ: Wiley-Blackwell.

Bögel, T., Strötgen, J., and Gertz, M. (2014). Computational narratology: extracting tense clusters from narrative texts. *Proc. Ed. Lang. Resour. Eval. Conf.* 14, 950–955.

Bossu, R., Landès, M., Roussel, F., Steed, R., Mazet-Roux, G., Martin, S. S., et al. (2016). Thumbnail-based questionnaires for the rapid and efficient collection of macroseismic data from global earthquakes. *Seismol. Res. Lett.* 88, 72–81.

Bossu, R., Laurin, M., Mazet-Roux, G., Roussel, F., and Steed, R. (2015). The importance of smartphones as public earthquake-information tools and tools for the rapid engagement with eyewitnesses: a case study of the 2015 Nepal earthquake sequence. *Seismol. Res. Lett.* 86, 1587–1592.

Bossu, R., Lefebvre, S., Cansi, Y., and Mazet-Roux, G. (2014). Characterization of the 2011 Mineral, Virginia, earthquake effects and epicenter from website traffic analysis. *Seismol. Res. Lett.* 85, 91–97.

Bossu, R., Mazet-Roux, G., Douet, V., Rives, S., Marin, S., and Aupetit, M. (2008). Internet users as seismic sensors for improved earthquake response. *Trans. Am. Geophys. Union* 89, 225–226.

Brown, H. L. (2012). *Representations of Haiti in Western news media: Coverage of the January 2010 earthquake in Haiti*. PhD thesis, Georgia State University, Atlanta.

Camilleri, S. (2019). *Mining Disaster Impact*. Master's thesis, University of Malta, Malta.

Camilleri, S., Azzopardi, J., and Agius, M. R. (2019). "Investigating the relationship between earthquakes and online news," in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Sardinia, 1–8.

Chung, D. H., and Bernreuter, D. L. (1981). Regional relationships among earthquake magnitude scales. *Rev. Geophys.* 19, 649–663.

Coburn, A., and Spence, R. (2002). *Earthquake Protection. 2nd End*. Hoboken, NJ: Wiley.

Derczynski, L., and Gaizauskas, R. (2015). "Temporal relation classification using a model of tense and aspect," in *Proceedings of Recent Advances in Natural Language Processing*, Madrid: Incoma Ltd, 118–122.

Devès, M. H., Texier, M. L., Pecout, H., and Grasland, C. (2019). Seismic risk: the biases of earthquake media coverage. *Geosci. Commun.* 2, 125–141.

Dhakal, S. P. (2018). Analysing news media coverage of the 2015 Nepal earthquake using a community capitals lens: implications for disaster resilience. *Disasters* 42, 294–313. doi: 10.1111/disa.12244

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., and Vaughan, A. (2010). OMG earthquake! can twitter improve earthquake response? *Seismol. Res. Lett.* 81, 246–251.

Earle, P. S., Bowden, D. C., and Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Ann. Geophys.* 54, 708–715.

Eisensee, T., and Strömberg, D. (2007). News droughts, news floods, and US disaster relief. *Q. J. Econ.* 122, 693–728.

Gupta, D. (2016). "Event search and analytics: detecting events in semantically annotated corpora for search and analytics," in *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, New York, NY: ACM, 705–705.

Guy, M., Earle, P., Ostrum, C., Gruchalla, K., and Horvath, S. (2010). Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. *Adv. Intell. Data Anal.* 9, 42–53.

He, W., Zha, S., and Li, L. (2013). Social media competitive analysis and text mining: a case study in the pizza industry. *Int. J. Inform. Manag.* 33, 464–472.

Heeger, B. (2007). *Natural Disasters and CNN: The Importance of TV News Coverage for Provoking Private Donations for Disaster Relief*. Rhode: Brown University, Providence.

Hicks, S. P. (2019). Geoscience analysis on Twitter. *Nat. Geosci.* 12, 585–586. doi: 10.1038/s41561-019-0425-4

Jurafsky, D., and Martin, J. H. (2014). *Speech and Language Processing*, 2nd Edn. London: Pearson.

Kanamori, H. (1983). Magnitude scale and quantification of earthquakes. *Tectonophysics* 93, 185–199.

Khumoyun, A., Cui, Y., and Lee, H. (2016). Real-time information classification in twitter using storm. *Bangk. Int. Conf.* 49, 1–4.

Kisilevich, S., Mansmann, F., and Keim, D. (2010). "P-DBSCAN: A density-based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos," in *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research and Application* , Vol. 38, New York, NY: ACM, 1–4.

Le Texier, M., Devès, M. H., Grasland, C., and De Chabalier, J. B. (2016). Earthquakes media coverage in the digital age. *Espace Géogr.* 45, 5–24. doi: 10.3390/ijerph16183239

Liang, Y., Caverlee, J., and Mander, J. (2013). "Text vs. images: on the viability of social media to assess earthquake damage," in *Proceedings of the 22nd International Conference on World Wide Web*, New York, NY: ACM, 1003–1006.

Liu, D. (2010). *A Comparative look at the Coverage of the Sichuan Earthquake in Chinese and American Newspapers*. Ames, LA: Iowa State University. Master's thesis.

Mani, I., and Wilson, G. (2000). "Robust temporal processing of news," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA: Association for Computational Linguistics, 69–76.

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., and Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Cambridge, MA: Academic Press.

Oh, O., Kwon, K. H., and Rao, H. R. (2010). "An exploration of social media in extreme events: rumor theory and Twitter during the Haiti earthquake 2010," in *Proceedings of the International Conference on Information Systems*, Saint Louis, MO, 231.

Panagiotou, N., Katakis, I., and Gunopulos, D. (2016). "Detecting events in online social networks: definitions, trends and challenges," in *Solving Large Scale Learning Tasks. Challenges and Algorithms*, eds S. Michaelis, N. Piatkowski, and M. Stolpe (Berlin: Springer), 42–84.

Pinto, A., Oliveira, H. G., and Alves, A. O. (2016). "Comparing the performance of different NLP toolkits in formal and social media text," in *OpenAccess Series in Informatics*, Vol. 51, (Wadern: Dagstuhl Publishing), 1–16.

Piskorski, J., and Atkinson, M. (2011). "Frontex real-time news event extraction framework," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM, 749–752.

Quigley, M. C., and Forte, A. M. (2017). Science website traffic in earthquakes. *Seismol. Res. Lett.* 88, 867–874.

Radinsky, K., Davidovich, S., and Markovitch, S. (2012). Learning to predict from textual data. *J. Artif. Intell. Res.* 45, 641–684.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, New York, NY: ACM, 851–860.

Simon, A. F. (1997). Television news and international earthquake relief. *J. Commun.* 47, 82–93.

Stomberg, D. R. (2012). *Disasters in the Media: A Content Analysis of the March 2011 Japan Earthquake/Tsunami and Nuclear Disasters*. Fort Collins, CO: Colorado State University. PhD thesis.

Suzanne, F. (2006). The CARMA report: western media coverage of humanitarian disasters. *Polit. Q.* 77, 1–17.

Tanev, H., Piskorski, J., and Atkinson, M. (2008). "Real-time news event extraction for global crisis monitoring," in *International Conference on Application of Natural Language to Information Systems*, Berlin: Springer, 207–218.

Van Belle, D. A. (2000). New York times and network TV news coverage of foreign disasters: the significance of the insignificant variables. *Journal. Mass Commun. Q.* 77, 50–70.

Van Noten, K., Lecocq, T., Sira, C., Hinzen, K. G., and Camelbeeck, T. (2017). Path and site effects deduced from merged transfrontier internet macroseismic data of two recent M4 earthquakes in northwest Europe using a grid cell approach. *Solid Earth* 8, 453–477.

Vannella, D., Flati, T., and Navigli, R. (2014). "WoSIT: A word sense induction toolkit for search result clustering and diversification," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Stroudsburg, PA: Association for Computational Linguistics, 67–72.

Wald, D. J., Quitoriano, V., Worden, C. B., Hopper, M., and Dewey, J. W. (2012). USGS "Did You Feel It?" internet-based macroseismic intensity maps. *Ann. Geophys.* 54, 688–707.

Wessel, P., Smith, W. H. F., Scharroo, R., Luis, J., and Wobbe, F. (2013). Generic mapping tools: improved version released. *Trans. Am. Geophys. Union* 94, 409–410. doi: 10.1002/2013EO450001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.