A Lightweight Representation of News Events on Social Media

Mauricio Quezada
mquezada@dcc.uchile.cl
Department of Computer Science,
Universidad de Chile & IMFD
Chile

Barbara Poblete
bpoblete@dcc.uchile.cl
Department of Computer Science,
Universidad de Chile & IMFD
Chile

ABSTRACT

The sheer amount of newsworthy information published by users in social media platforms makes it necessary to have efficient and effective methods to filter and organize content. In this scenario, off-the-shelf methods fail to process large amounts of data, which is usually approached by adding more computational resources. Simple data aggregations can help to cope with space and time constraints, while at the same time improve the effectiveness of certain applications, such as topic detection or summarization. We propose a lightweight representation of newsworthy social media data. The proposed representation leverages microblog features, such as redundancy and re-sharing capabilities, by using surrogate texts from shared URLs and word embeddings. Our representation allows us to achieve comparable clustering results to those obtained by using the complete data, while reducing running time and required memory. This is useful when dealing with noisy and raw user-generated social media data.

CCS CONCEPTS

• **Information systems → Document collection models**.

KEYWORDS

social media, news events, graphs, modeling, topic detection, summarization, word embeddings, clustering, anchor text

ACM Reference Format:
Mauricio Quezada and Barbara Poblete. 2019. A Lightweight Representation of News Events on Social Media. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331300

1 INTRODUCTION

The overload of information in social media streams makes it difficult for users to obtain a comprehensive account of different emerging events reported by users. The ability to identify and summarize relevant information about newsworthy topics that are accounted for by users in social media can be of great use to society. This is particularly true in the case of high impact and breaking news, like natural disasters and crisis situations [1]. In general, social media platforms present their content to the consumer in one of two ways: (1) by displaying user posts in reverse chronological order, or (2) by displaying posts in a personalized fashion, showing first what they deem more interesting to the user. However, if the user is seeking information about a specific event, the aforementioned approaches can result also in unnecessary exposure to redundant, irrelevant, or even misleading information. In this context, delivery of key pieces of information to the user is important, and even critical in some cases; for instance, social media is often used as a complementary source of information for disaster management and response [1, 12] and also to obtain information about newsworthy events not yet reported by formal news outlets.

Despite the usefulness of social media as a worldwide news information source, its consumption is not without challenges. These challenges include, among many others, correctly assessing information veracity and relevance to a specific topic, as well as dealing with a huge volume of data with variable quality. In particular, we study Twitter[1], a popular microblog platform in which users can post text-based messages. These messages, called *tweets* can include links to images, videos, to external webpages or to other tweets, making this platform's content multimodal. The URLs shared in Twitter often link to varied types of media (articles, images, or videos), which have been found useful for identifying conversation topics [9]. In the context of news events, users tend to quickly reshare information if the event sparks high interest [7], which results in many (near-)duplicate messages. Prior work has acknowledged the need to develop techniques for dealing exclusively with this type of social media data.

In this work we focus on the problem of producing summarized representations of social media data related to news events without significant loss of information. We present a compact representation for social media content related to news events. This representation allows us to preserve information about the topics involved in an event, and to identify relevant content, while reducing the volume of data. This can be especially useful for online data processing about developing news and news information seeking tasks. To achieve this, our model annotates shared URLs with the text of the messages in which they appear. Here we use messages as *anchor text* or *surrogate text*. We first leverage this idea by identifying relevant URLs that are shared in social media during a news event. We discard URLs that are too general with respect to an event (e.g. a general report), or generic (e.g. linking to the homepage of a news outlet) as we deemed them as non-relevant. Then we aggregate all the anchor texts associated with relevant URLs into *documents*, and all the conversations around those documents as well.

As a preliminary way to study the usefulness of our event representation, we applied it to three news events as a case study. We

[1] https://twitter.com (Accessed: Feb 19, 2019)

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331300

observed that our representation reduces the amount of records needed to describe an event by one order of magnitude, compared to using the raw tweets. We were also able to identify sub-topics by clustering the documents using dense word representations, such as neural network–based word embeddings. We show that the clustering based sub-topic detection task using our proposed representation yields similar external quality metrics to those obtained using the entire data. Hence, indicating that we are preserving relevant aspects of the information while considerably reducing running time. In summary, our contributions are the following:

(1) We present a novel compact representation for news event information in microblogs, based on shared URLs. This representation reduces data volume.
(2) We study the usefulness of our representation through three case studies. We do so by introducing a methodology to represent news events in Twitter and find sub-topics. We show that our approach displays comparable performance to baseline methods at a fraction of the computational resources.

2 RELATED WORK

Two lines of research are relevant to our work: the utility of anchor texts in microblogs and topic detection methods using different aggregation strategies.

About the usefulness of anchor texts on Twitter, Raux et al. [11] used anchor texts from tweets pointing to a predefined set of URLs to characterize general topics, by clustering a bipartite graph of words and URLs. We instead focus on news events, and tweets related to these news, which have their own particularities, in contrast to long lasting general topics. Mishne and Lin [9] studied the contribution of anchor texts compared to the text of the websites behind the URLs, concluding that anchor text add new terms not seen in the website content, by looking also at the conversations around the sharing tweets. Alonso et al. [2] had similar findings examining Facebook posts. In another work by Alonso et al. [3], the authors designed a social search engine using the propagation of shared URLs as cues to measure virality, and anchor texts to augment metadata of the search results with social content. In a previous work [10], we showed a proof of concept of automatic summarization of documents using anchor texts.

Topic modeling of tweets is an active line of research, and there are some studies which investigate the effectiveness of aggregation of tweets to improve topic detection [4, 6, 8]. Hong and Davison [6] analyzed the effects of different aggregation strategies of tweets when finding topics with Latent Dirichlet Allocation (LDA). They found that some schemes yield better results at some tasks, such as classification problems related to tweets. In a similar fashion, Mehrotra et al. [8] observed that aggregating hashtags (also described as *pooling* of tweets by hashtags) is a more effective strategy to identify topics from tweets, but at the cost of longer running times due to the duplication of tweets. Finally, Alvarez-Melis and Saveski [4] found that adding the threads of conversation into the pooling is more effective than pooling by hashtag in their observations. We instead aggregate by shared URLs and conversations, with the goal of generating a compact representation without major loss of information.

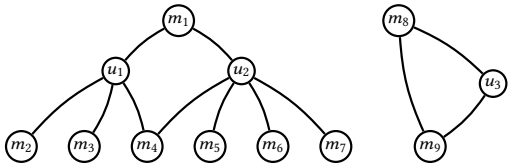


Figure 1: Example representation of messages. Messages m_1 and m_4 shares URLs u_1 and u_2, while m_2 only shares u_1; m_8 and m_9 share or reply one to another. Each connected component is a *document*.

3 EVENT REPRESENTATION

We introduce our model and the methodology for applying it to messages that describe events in social media.

Our model is based on the assumption that *most of the social media posts that discuss a specific event and contain a URL, are messages that cover a particular portion or sub-topic of the event.* For example, in the case of an earthquake, tweets that refer to damages to buildings may share a URL to an external news report; similarly, a tweet that discusses the magnitude of the event may link to the official seismological report. We define an event as a tuple $\mathcal{E} = (M, U)$, where M is the set of messages that discuss a real-world occurrence, and U is the set of *topic-specific* URLs that are mentioned or shared by messages in M. We denote the URLs U as *topic-specific*, assuming that each URL in U identifies only one of the different topics within the real-world event. Our representation is a graph $\mathcal{G} = (V, E)$, where $V \subseteq M \cup U$ is the set of nodes, comprised of *messages* and *URLs*, and there is an edge $(i, j) \in E$ if at least one of the following conditions hold: (1) $m = i$ is a message and $u = j$ is a URL and m shares u, (2) $m_1 = i$ and $m_2 = j$ are messages and m_1 re-shares m_2, or (3) $m_1 = i$ and $m_2 = j$ are messages and m_1 replies to m_2 (see an example in Figure 1). Finally, a *document* is a connected component of $\mathcal{G}$. In this case, a document is defined as a collection of messages that discuss only one aspect of the event.

The assumption of U to be a set of topic-specific URLs may not hold in all cases, for example, for URLs that address general aspects of an event (e.g., a summary report of an entire event), or that are generic (e.g., refer to a online news website's root URL), or irrelevant to the event (e.g., spam, or unrelated information). We deal with these cases in our methodology for generating event models, presented below. In that sense, our model and methodology focus on identifying URLs that are specific to one aspect of an event, and use those URLs to aggregate similar messages. Our goal is to represent underlying sub-topics of an event by just using shared URLs. Please refer to the poster for more examples[2].

Given a set of event-related social media messages, we propose the following methodology for representation generation:

Filter out generic URLs: We identify generic URLs (too general with respect to the event) as those that co-occur with several different other URLs in the same messages. Highly connected URLs do not contribute to specific topics: links that are very generic (e.g., `cnn.com`) or very general to the event (e.g., general reports). All of the messages mentioning these URLs would fall into the same component, regardless of their differences in content. We removed

[2] https://users.dcc.uchile.cl/~mquezada/papers/sigir2019_poster.pdf

URLs that co-occurred with three or more different URLs across messages. This threshold yielded the best results in our case studies.

Representation generation: The generation step consists of grouping messages that end up in the same component, as described in the previous section. For that, a straightforward method is to compute all pairs of messages and pairs of URLs and messages that fulfill the conditions stated in the representation definition and then find connected components using a union-find algorithm. The URLs are the non-generic ones identified in the previous step, as an approximation to the topic-specific URLs.

Vector representation of documents: Finally, we aggregate messages into documents in order to produce a vector representation, using neural network-based word embeddings. This procedure generates dense document representations.

4 CASE STUDIES

We describe the case studies, the data, and the experiment we performed to validate our representation.

4.1 Datasets

For the case studies, we selected three events from the data made available by Kalyanam et al. [7]. Namely, a terrorist attack (2015 Libya Hotel Attack), a long-lasting event (2014 Oscar Pistorius Trial), and a natural disaster (2015 Nepal Earthquake). We report duration in days (corresponding to the amount of days encompassing at least 95% of the tweets), total tweets, retweets, and unique resolved URLs (Table 2). We also show example tweets (Table 1).

4.2 Experimental setting

To generate the representation for each event, we discarded all tweets that have more than two URLs or more than three hashtags, as we consider them as potential spam tweets. We resolved every shortened URL mentioned in the tweets by following redirects. Even though the tweet meta-data may include the URLs as before they were shortened by the Twitter platform, they are often shortened by additional external services (e.g., `bit.ly`) even more than once. Also, we removed all query strings from the URLs, with some exceptions, which for some sites they are relevant for identifying the resource (e.g., `?id=`, `?fbid=`, `?v=`, etc.).

We discarded all URLs that co-occurred with 3 or more other URLs in the same tweets. Then, we computed the representation for each event by identifying connected components in the graph of tweets and URLs, considering only components with URLs. This resulted in 2 957 documents for the Libya event, 20 984 for the Nepal event, and 9 092 for the Pistorius event.

We used fastText [5] to produce dense vectors from the documents. We choose fastText due to its capability to encode subword information into the embeddings and to encode some out-of-vocabulary words, resulting in better quality embeddings for rare or uncommon words. For the generation of document vectors, we took all the tweets in a document, and obtained the vector of each word in each tweet. Then, we took the sum of the vectors of the words in the document. We trained 300-dimension word embeddings using the dataset of Kalyanam et al. [7] consisting of of 193 million event-related tweets (3 billion words). Note that the training of vectors can be done off-line and is done only once.

4.3 Validation of sub-topic detection task

To validate our representation in the case studies, we identified topics in our events using our representation and using raw tweets.

Sixteen people –mainly Computer Science undergrad and grad students– labeled tweets independently using a custom Web interface to produce a ground-truth. The interface displayed a tweet and a list of labels, and each user could assign one or more labels to a tweet, mark the tweet as non relevant, or skip it. Some tweets may refer to more than one sub-topic, and we preferred that users felt free to assign as many labels as they prefer. We imposed a limit of three evaluations per tweet. We manually generated the list of labels by looking into news reports in the Web. The tweets were chosen in such a way that there were roughly no underrepresented sub-topics. For this, we manually produced a list of keywords for each label, and for each label ranked the tweets using Okapi BM25; in the interface, a random label was chosen and the corresponding ranked tweet was presented. Finally, to assign a ground-truth label to a tweet, we selected the label users chose the most. This resulted in 401 labeled tweets (1339 labels in total) for the Libya event, 368 (531 labels) for Nepal, and 85 (362 labels) for Pistorius.

As our goal is to compare event representations, we chose k-means as a simple baseline to validate the effectiveness of our model. To find sub-topics, we ran k-means with different numbers of clusters, using our representation and raw tweets. For the baseline, we considered each tweet as a document, that is, we computed the sum of word vectors for each tweet individually. We report normalized mutual information, purity, and entropy for each event, using the available labels in both settings (Figure 2). Note that the measures were done only on the labeled tweets, that is, as if the unlabeled tweets did not exist in the clustering solution.

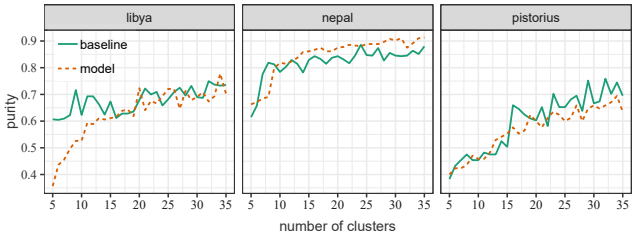
We observe that with our representation, the clustering outperforms the baseline in most cases. In the case of the Nepal event, our representation had better purity, NMI and entropy. In the case of Pistorius, the measures are very similar. However, in the case of Libya, our representation matches the baseline after a certain number of clusters, except in the entropy measure. We believe this is because the Libya event has more unrelated (spam and irrelevant) tweets, and the model captures more information about unrelated topics, as they use more URLs. And in terms of running time, we observed that k-means under the representation runs one order of magnitude faster than the baseline (Figure 3). These results suggest that our representation is capable of preserving topical information about the target event, with reduced time required to identify this kind of information.

5 DISCUSSION AND FUTURE WORK

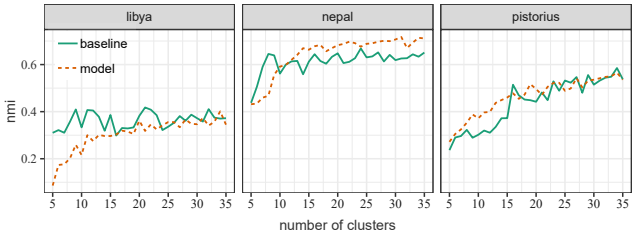
Our proposed representation allows us to identify sub-topics in an event much faster than traditional or off-the-self methods. However, there is room for improvement in terms of the results. For this, we need to perform a large scale evaluation, although it is hard to find ground truth data for large, raw, uncurated social media messages [2]. For this case scenario, our methodology aims to help processing large quantities of noisy, un-curated data around news events more effectively. In future work we are interested in studying how to formally define and identify topic-specific URLs in the events, which we believe are key to creating a useful representation.

Table 1: Sample tweets for each event.

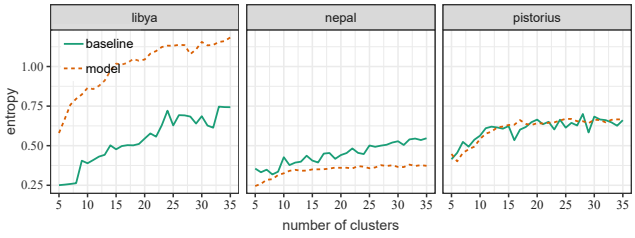
Event	Sample tweets
Libya Hotel Attack	*Gunmen possibly linked to Islamic State attack hotel popular with foreigners in Libyan capital Tripoli -officials* <URL> *#Libya forces surround luxury hotel in #Tripoli where gunmen have taken hostages after car bomb attack - Reuters/AP*
Oscar Pistorius Trial	*Oscar Pistorius pleads not guilty Monday to all 4 charges against him, marking start of Olympian's murder trial.* <URL> *RT* <mention>*: Oscar Pistorius murder trial set to begin in South Africa:* <URL>
Nepal Earthquake	*Buildings down & roads out after major 7.5 magnitude earthquake hits #Nepal. Quake could be felt as far as Delhi* <URL> *BREAKING: 15-year-old girl dead near Nepal border after quake brings house wall down, reports Reuters. Read more:* <URL>



(a) Purity.



(b) Normalized Mutual Information.



(c) Entropy.

Figure 2: External clustering measures for target events.

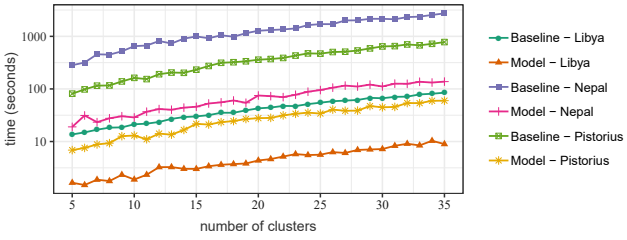


Figure 3: Running times for clustering.

Table 2: Datasets for case studies.

Name	Duration	Tweets	Retweets	URLs
Libya Hotel Attack	8 days	28 616	12 280 (43%)	3 385
Nepal Earthquake	1 day	522 434	363 102 (70%)	22 661
Oscar Pistorius Trial	70 days	113 189	26 307 (23%)	9 335

ACKNOWLEDGMENTS

This work was supported by the Millennium Institute for Foundational Research on Data (IMFD). M. Quezada was also supported by CONICYT PCHA/Doctorado Nacional 2015/21151445.

REFERENCES

[1] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. *International AAAI Conference on Web and Social Media* (2018).

[2] Omar Alonso, Sushma Bannur, Kartikay Khandelwal, and Shankar Kalyanaraman. 2015. The World Conversation: Web Page Metadata Generation From Social Sources. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, 385–395.

[3] Omar Alonso, Vasileios Kandylas, Serge-Eric Tremblay, Jake M. Hofman, and Siddhartha Sen. 2017. What's Happening and What Happened: Searching the Social Web. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*. ACM, New York, NY, USA, 191–200.

[4] David Alvarez-Melis and Martin Saveski. 2016. Topic Modeling in Twitter: Aggregating Tweets by Conversations. *ICWSM* 2016 (2016), 519–522.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *CoRR* abs/1607.04606 (2016). arXiv:1607.04606

[6] Liangjie Hong and Brian D. Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)*. ACM, New York, NY, USA, 80–88.

[7] Janani Kalyanam, Mauricio Quezada, Barbara Poblete, and Gert Lanckriet. 2016. Prediction and characterization of high-activity events in social media triggered by real-world news. *PloS one* 11, 12 (2016), e0166694.

[8] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, New York, NY, USA, 889–892.

[9] Gilad Mishne and Jimmy Lin. 2012. Twanchor Text: A Preliminary Study of the Value of Tweets As Anchor Text. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 1159–1160.

[10] Mauricio Quezada and Barbara Poblete. 2013. Understanding Real-World Events via Multimedia Summaries Based on Social Indicators. In *Collaboration and Technology*. Springer Berlin Heidelberg, Berlin, Heidelberg, 18–25.

[11] Stéphane Raux, Nils Grünwald, and Christophe Prieur. 2011. Describing the Web in less than 140 Characters. *International AAAI Conference on Web and Social Media* (2011).

[12] Hernan Sarmiento, Barbara Poblete, and Jaime Campos. 2018. Domain-Independent Detection of Emergency Situations Based on Social Activity Related to Geolocations. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '18)*. ACM, New York, NY, USA, 245–254.