A Multi-Source Collection of Event-Labeled News Documents

Ida Mele
ISTI-CNR, Pisa, Italy
ida.mele@isti.cnr.it

Fabio Crestani
USI, Lugano, Switzerland
fabio.crestani@usi.ch

ABSTRACT

In this paper, we present a collection of news documents labeled at the level of crisp events. Compared to other publicly-available collections, our dataset is made of heterogeneous documents published by popular news channels on different platforms in the same temporal window and, therefore, dealing with roughly the same events and topics.

The collection spans 4 months and comprises 147K news documents from 27 news streams, i.e., 9 different channels and 3 platforms: Twitter, RSS portals, and news websites. We also provide relevance labels of news documents for some selected events. These relevance judgments were collected using crowdsourcing. The collection can be useful to researchers investigating challenging news-mining tasks, such as event detection and tracking, multi-stream analysis, and temporal analysis of news publishing patterns.

CCS CONCEPTS

• **Information systems → Test collections**.

KEYWORDS

test collections; news streams; event detection and analysis

ACM Reference Format:
Ida Mele and Fabio Crestani. 2019. A Multi-Source Collection of Event-Labeled News Documents. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3341981.3344253

1 INTRODUCTION

Although news collections are crucial for research on news mining, there is a lack of datasets gathering news documents from different sources, with fine-granularity labels representing the relevance of a news document to a particular event, and spanning more than one month. Multi-source datasets ensure heterogeneity of information, which is important for different challenging tasks ranging from detection of untrusted sources of information (e.g., news reported only by one channel could be fake news) to the analysis of the news channels' publishing patterns (e.g., which channel reported a news before the others and on which platform).

News collections should provide labels representing crisp events reported in the news documents (e.g., *Attack at the Brussels Airport* or *Earthquake in Ecuador*). Such labels can be used for evaluating the effectiveness of event-detection approaches. Another important feature of news datasets is their time span. In particular, collections that cover more than a month allow not only to assess the effectiveness of event detection techniques but also to track how the events evolve over time (e.g., bomb attack followed by the arrests of the terrorists) that can be useful for news recommendation and summarization [7].

In this paper, we present our effort in building a multi-source collection of news documents with fine-granularity event labels. We used this dataset for our research on event detection and tracking as well as for cross-linking news streams [9, 10]. The collection consists of news published by 9 popular channels (e.g., BBC, CNN, NBC) using 3 different platforms (i.e., Twitter, RSS portals, and news websites) over a period of 4 months. The news documents are represented by tweets, RSS feeds, and news articles, hence they are heterogeneous in term of writing styles and length.

Other publicly-available collections are mostly made of homogeneous documents (e.g., only tweets or only news articles). Such documents are unlabeled or the labels represent coarse topics (e.g., *sport* or *politics*); however, document-event relevance labels are needed to assess the performance of event-detection tools. To overcome this limitation, we randomly selected 57 events which are described by a set of keywords. We retrieved news documents potentially relevant to these events and used crowdsourcing for collecting judgments on the relevance of a news document to an event. Since crowdsourcing evaluations are costly in term of time and money, we believe that our collection, providing not only the content of news documents but also the relevance labels, can be very useful for researchers who may want to assess their event detection and tracking tools but they lack collections for performing the assessment. Furthermore, having one collection used by different research teams for quantifying the performance of their approaches makes easier and more fair the comparison among different techniques.

The contribution of this paper is two-fold: (1) We present the first multi-source collection of news documents (i.e., tweets, RSS feeds, and news articles) downloaded from 27 different news streams (i.e., 9 channels and 3 platforms) for 4 months; (2) We provide labels for the news documents based on the events they describe. The events were discovered automatically and, for a subset of 57 events, a crowdsourcing evaluation was conducted to collect labels on the relevance of the news documents to the events.

2 RELATED COLLECTIONS

Several news and tweet collections have been published for helping IR and NLP research, in particular, for text classification, news clustering, event detection and tracking. These collections present some limitations when dealing with cross-linking news streams, e.g., for analyzing publishing patterns, plagiarism detection, and detecting

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344253

trustworthy sources. Indeed, most of them do not gather news from multiple sources but rather have homogeneous documents from a limited number of platforms. Moreover, the labeling is at the level of general topics (e.g., *politics, sport,* and *finance*) which can be used for text categorization but not for event detection and tracking. In this section, we discuss the novelty of our collection compared to some other publicly-available datasets.

Homogeneous-Document Collections. Many collections comprises one type of document, e.g., news articles. An early collection gathered by the Carnegie Group and Reuters (*Reuters-21578 Text Categorization Collection*[1]) was made of news published by Reuters in 1987. Its labels are general topics such as *Money/Foreign Exchange*, *Shipping*, used for evaluating a text-categorization system. Similarly, the *20 Newsgroups* dataset[2] consists of 20K documents, published in 1997, and partitioned across 20 different newsgroups, each one representing a different topic (e.g., *sci.electronics* and *sci.med*). These collections are homogenous and their labels are too coarse to be used for event mining.

Other datasets were provided for different tasks. For example, Nova Search[3] released news collections for assessing the quality of online news articles in term of fluency, richness, novelty, etc. Some TREC collections, consisting of blog and microblog data, were provided for blog distillation and opinion mining[4], or for real-time filtering tasks[5].

Another example of homogenous dataset is represented by tweet collections. Petrovíc et al. [12] tested their technique for first-story detection on a collection of 50M English tweets. They asked some experts to select tweets relevant to 27 manually picked events and created a subset of 3K relevant tweets. Also, McMinn et al. [8] built a test collection for evaluating event-detection techniques. The authors crawled around 120M English tweets and collected labels for over 150K tweets covering more than 500 events which were identified using automatic and manual ways. Compared to these tweet datasets, our collection gathers different types of documents (tweets, RSS feeds, and news articles), allowing to work with documents that are heterogeneous in length and writing styles.

Multiple-Source Document Collections. Datasets from different sources mainly consist of news and blog articles. For example, the *LivingKnowledge news and blogs annotated subcollection*[6], used for Temporalia Task and the *TREC Knowledge Based Acceleration 2014 Stream Corpus*[7]. These collections are annotated at the level of sentence splitting, named entities, and time, while our collection provides labels on the event relevance of news documents.

Another more recent dataset was released by *Signal Media* [8] for the NewsIR'16 workshop. It contains over 1M news and blog articles published in September 2015. At a later time, the collection was enriched with related tweets [9]. Although the dataset is made of documents from different sources including major newswires as well as local blogs, it has no labels on the topic/event relevance of documents. Moreover, it covers only 1 month that makes difficult the event tracking.

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[2] http://qwone.com/~jason/20Newsgroups/
[3] http://novasearch.org/datasets/
[4] http://trec.nist.gov/data/blog.html
[5] http://trec.nist.gov/data/microblog.html
[6] http://ntcir.nii.ac.jp/Temporalia/NTCIR-11-Temporalia/Document-Collection/
[7] http://trec-kba.org/data/fakba1/index.shtml
[8] http://research.signalmedia.co/newsir16/signal-dataset.html
[9] https://research.signal-ai.com/datasets/signal1m-tweetir.html

3 COLLECTION POTENTIAL APPLICATIONS

In the following, we discuss the features of our collection and outline its possible uses for supporting different areas of research. Our collection is made of documents labeled at the level of fine-granularity events rather than general topics. This can foster research on event detection and tracking [4, 8, 11] as well as event segmentation [6] and novel-event detection [12].

Moreover, the dataset covers several months allowing to track the evolution of events (e.g., Muhammed Ali's death followed by his funeral). This is beneficial for recommending news to users, story reconstruction, and news summarization [7]. Also, it can foster research on dynamic topic modeling which tries to model the topic/event evolution over time [1, 2].

Another feature of our collection is that it is made of heterogeneous documents (i.e., tweets, RSS feeds, and news articles) that differ in term of length and writing styles. It can be used for assessing the performance of techniques on heterogeneous text. As an example, information extraction is more challenging on tweets since they are short and provide less textual context.

Lastly, our data consists of news published by different sources, allowing to cross-link the streams that have reported news about the same event. This has several potential applications, such as using links as an endorsement to discover untrusted sources of information and fake news, or analyzing the temporal publishing patterns of newswires [9, 10].

4 BUILDING THE COLLECTION

In this section, we describe how we collected the data, extracted the events, and conducted the crowdsourcing evaluation.

4.1 Data Gathering

We built the dataset of heterogeneous news documents by monitoring 9 different news channels: American Broadcasting Company (`abc`), Al Jazeera (`alj`), British Broadcast (`bbc`), Canadian Broadcast (`cbc`), Cable News Network (`cnn`), NBC News (`nbc`), Reuters (`reu`), United Press International (`upi`), Xinhua China Agency (`xin`). For each channel we downloaded data from 3 different platforms: Twitter, RSS portals, news websites. The gathering was conducted for a period of 4 months (from March 1 to June 30, 2016).

For collecting tweets, we monitored the Twitter accounts of some selected newswires using the `getUserTimeline` method provided by the Twitter REST APIs[10]. It allows to download the latest tweets posted or retweeted by a Twitter user. For each tweet, it provides all the information about it (author, text, timestamp, etc.) in JSON format. Data was downloaded every 24 hours and repeated tweets were removed, if needed. For the RSS feeds and news articles, we collected the up-to-date news every 15 minutes and removed duplicated feeds/news links when needed. Each document consists of a title, content, link, timestamp, channel, and platform. Note that the time when a news document has been published can reflect a

[10] https://dev.twitter.com/rest/public

different time zone, so we converted all the publishing timestamps to UTC, ensuring the alignment of data from different sources.

It is worth to notice that we could not enrich existing news collections with tweets and RSS feeds since these cannot be downloaded anymore. In particular, the RSS platforms are updated at a high speed, so feeds about novel events quickly replace the old ones. Regarding Twitter, its REST APIs provide *search* and *timeline* functions. The *search* is used to retrieve tweets with a given set of terms back up to 7 days. The *timeline* collects up to 3,200 most recent tweets published by a given account. Due to this limitation, one can get outdated tweets only from less active accounts, while the news channels are very active, so their latest tweets are about more recent events.

4.2 Event Detection and News Retrieval

To discover popular events happened in the 4 months of our data, we first split the time-sorted dataset into time buckets of fixed size (i.e., 24 hours). Then, we run an NLP tool [13] for extracting named entities and event-descriptive keywords. The most frequent co-occurrences of keywords K_e are used to represent the event e. For example, *earthquake, ecuador, strikes* represent the earthquake happened in Ecuador in April 2016.

To determine if an event is new or can be merged with another one, we computed the overlapping of keywords in consecutive time buckets. This overlapping depends on the length of the list of keywords, if the percentage of common keywords is high (i.e., 70%), then the two events are merged. Iterating this operation results in events with flexible temporal windows. For example, attention towards Brussels terror attacks lasted more compared to mudslides in Nairobi. Our approach only merge events in consecutive time buckets ($\delta = 24$ hours) since when two temporal buckets are distant, very likely the events are distinct even if they may have high-overlapping keywords. As an example, the Japan earthquake and the 5th anniversary from Japan Tsunami have high keywords overlapping but the former refers to the earthquake that hit the South of Japan in mid April and the latter to the anniversary, marked in March 2016, of the tsunami which devastated Japan in 2011.

After having detected the events, we checked on Wikipedia their veracity, namely, if they really occurred and the temporal window is estimated correctly. Given the large number of news documents, we could not collect relevance labels for all the discovered events. Hence, we selected a subset of 57 events and retrieved the potentially relevant news documents. In particular, we represent news documents and even keywords as tf-idf vectors and compute their cosine similarity. The news were also filtered by time to make sure to have only the ones published in the temporal window of the event. Then, we conducted a crowdsourcing evaluation to determine the relevance of the news documents to the events.

4.3 CrowdFlower Evaluation

The evaluation was conducted using the CrowdFlower[11] crowdsourcing platform. We collected human judgements for the pairs {event, news} where the former is represented by the event keywords, while the latter is a potentially relevant news document (i.e., a tweet, an RSS feed, or a news article) and its timestamp (i.e., when it was published). The evaluation task consisted in showing the pair {event, news} to the evaluators and asking them to determine whether the news document was about the event or not. They could also select "I can't decide" and move to the next question if unsure about the answer.

Before starting the task, the evaluators read the instructions describing the scope of the evaluation and how to perform it. Also, some examples were shown to make clear that the news document should be relevant to a specific event. For example, the news "A 6.4 earthquake hits southern Japan with reports of collapsed buildings and people injured" could be relevant to the event *Japan earthquake* but not to *5th anniversary from Japan Tsunami*. The evaluators were asked to pay attention to all the event keywords and to check whether the news document matches the event or not.

To guarantee high quality results we created 200 *gold questions* (i.e., questions with known answers). These questions were used as a quiz during the training phase. Although the task is quite easy and does not require a particular training of the evaluators, we decided to have a short training phase in order to avoid spammers, so the evaluators were asked to successfully complete at least 3 out of 5 gold questions. Moreover, the gold questions were randomly shown during the evaluation to detect low-quality answers, sloppy workers, and malicious activities (e.g., spam). Once the evaluators started the evaluation, they had to maintain a minimum accuracy of 70% to be considered "trusted evaluators" and allowed to continue the task. Only labels from trusted evaluators were included in our collection. We collected relevance judgements for 4.3K pairs {event, news}. For each of them we collected judgements from at least 3 different trusted evaluators. The news document was considered truly relevant to the event if at least 2 out of 3 evaluators agreed.

To measure the inter-annotator agreement we computed the Fleiss' Kappa, κ, which measures how consistent are the assessors' ratings [3]. We got $\kappa = 0.45$ corresponding to a *moderate agreement* (according to the table for interpreting κ values provided in [5]). Note that Fleiss' Kappa does not take into account the trustworthiness of the assessors, while the CrowdFlower's confidence gives the level of agreement between multiple contributors weighted by the contributors' trust scores. For our task, we got a high average confidence of 0.91, this is expected due to the high quality of our labels as well as the easiness and objectiveness of the task.

5 COLLECTION

Our collection[12] consists of two sources:

1. News Streams. We monitored 27 (i.e., 9 channels and 3 platforms) streams of news for 4 months: from March 1 to June 30, 2016. For each news stream (identified by a channel and a platform) we provide a timestamp-sorted list of news documents published by that channel on that platform.

Overall, we collected ~147K news documents, out of which ~24K are news articles, ~43K are RSS feeds, and ~80K are tweets. Notice that the news articles are less than the RSS feeds as some of the feeds did not have the link or it was impossible to download the page content of the news article.

2. Events and Relevance Judgements. We identified events with the approach explained in Section 4.2, then we randomly selected

[11] http://www.crowdflower.com/

[12] Dataset is available here: http://hpc.isti.cnr.it/~idamele/NewsEventData.html

Table 1: Example of events and news documents with relevance

Event	Source	Time	News document	Relevant	Confidence
attack, brussels-airport, isis, maelbeek-metro	abc NEWS	2016-03-22 07:31	Brussels attacks airport metro rocked explosions killing least 34 ...	Yes	1.0
	cbc NEWS	2016-03-22 08:32	Brussels attacks: Explosive device, chemicals found as police carry out raids...	Yes	0.6
	upi RSS	2016-03-22 09:04	Blasts kill at least 13 at Brussels airport, subway station	Yes	1.0
	reu RSS	2016-03-20 11:40	Paris suspect's lawyer to sue French prosecutor: media BRUSSELS...	No	0.6
japan, kumamoto, earthquake, damage, victims	xin TWT	2016-04-15 20:31	Japan's quake upgraded to magnitude-7.3, 1 death confirmed	Yes	1.0
	cnn RSS	2016-04-14 15:57	Japan quake destroys 19 houses; people still trapped	Yes	1.0
	cnn NEWS	2016-04-16 07:05	Japan earthquakes dozens killed race against clock find survivors...	Yes	0.7
	cnn NEWS	2016-04-18 12:46	Ecuador earthquake: Rescuers race to find survivors as death toll climbs...	No	0.7
oklahoma, tornado, storms, destruction, rescue	cnn TWT	2016-05-09 21:44	Large and dangerous tornado spotted on the ground in central Oklahoma	Yes	1.0
	bbc RSS	2016-05-10 03:56	Tornado hits Oklahoma. At least one person is killed after a large tornado...	Yes	1.0
	bbc NEWS	2016-05-10 03:56	One dead after Tornado hits Oklahoma. At least one person was killed...	Yes	0.7
	nbc NEWS	2016-05-10 23:51	10 Injured as Large Tornadoes Range Across Western Kentucky...	Yes	0.6
bonnie, carolina, tropical-depression, storm, flooding	cnn TWT	2016-05-27 21:27	Forecasters issue a tropical storm warning for the coast of South Carolina	Yes	1.0
	cbc RSS	2016-05-28 09:58	U.S. tropical storm warning issued for South Carolina	Yes	1.0
	cbc NEWS	2016-05-29 15:05	Weakened but rainy tropical depression Bonnie hits South Carolina...	Yes	1.0
	nbc NEWS	2016-05-29 18:35	Five Dead, at Least Three Missing in Texas and Kansas Flooding...	No	0.6

57 events for the evaluation. An event is identified by a set of event-descriptive keywords. We used these keywords to get the potentially relevant news documents. Each pair {event, news} was shown to at least 3 different evaluators to determine whether the news is relevant or not to the event. The data consists of 4.3K labeled pairs {event, news}, for each of them we provide the news id, the event keywords, the relevance judgement with the confidence score.

The distribution of news documents per event is shown in Figure 1. As expected, popular events, such as *Brexit*, *Zika virus diffusion*, and *Brussels attacks*, have more relevant news compared to other minor events, such as the *volcano eruption in Indonesia*. The average number of relevant news per event is ~75 (see red line).

Table 1 shows some examples of events and news documents with their publishing sources and timestamps. We also report the relevance of the news to the event and the confidence score. Due to the space limit, we can only show a few examples of events with a subset of their news documents (i.e., tweets, RSS feeds, and excerpts of the original news articles).

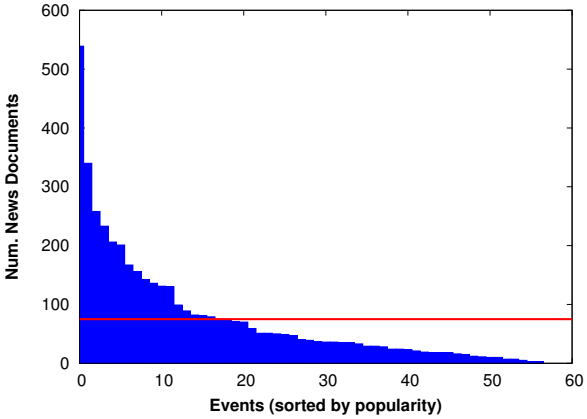

Figure 1: Distribution of news documents per event

6 CONCLUSIONS

This paper described our effort in building a multi-source collection of news documents and evaluating them on the relevance to real-world events. The collection consists of news streams, event keywords, and event-labeled news documents. It spans 4 months and gathers around 147K news documents from 27 different news streams. We automatically detected the events discussed in the news documents and the event-relevance labels are provided for a subset of 57 events. To determine the relevance of a document to an event we performed a crowdsourcing evaluation on CrowdFlower. This collection will foster research on event detection, multi-stream mining, trustworthy analysis, and dynamic topic modeling.

REFERENCES

[1] S. A. Bahrainian, I. Mele, and F. Crestani. 2017. Modeling Discrete Dynamic Topics. In *SAC '17*. ACM, New York, NY, USA, 858–865.
[2] D. M. Blei and J. D. Lafferty. 2006. Dynamic Topic Models. In *ICML '06*. ACM, New York, NY, USA, 113–120.
[3] J. L. Fleiss. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin* 76, 5 (1971), 378–382.
[4] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. 2005. Parameter Free Bursty Events Detection in Text Streams. In *VLDB '05*. VLDB Endowment, 181–192.
[5] J. R. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.
[6] C. Li, A. Sun, and A. Datta. 2012. Twevent: Segment-based Event Detection from Tweets. In *CIKM '12*. ACM, New York, NY, USA, 155–164.
[7] Z. Li, J. Tang, X. Wang, J. Liu, and H. Lu. 2016. Multimedia News Summarization in Search. *ACM Trans. Intell. Syst. Technol.* 7, 3, Article 33 (2016), 33:1–33:20 pages.
[8] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. 2013. Building a Large-scale Corpus for Evaluating Event Detection on Twitter. In *CIKM '13*. ACM, New York, NY, USA, 409–418.
[9] I. Mele, S. A. Bahrainian, and F. Crestani. 2017. Linking News Across Multiple Streams for Timeliness Analysis. In *CIKM '17*. ACM, New York, NY, USA, 767–776.
[10] I. Mele, S. A. Bahrainian, and F. Crestani. 2019. Event mining and timeliness analysis from heterogeneous news streams. *Information Processing & Management* 56, 3 (2019), 969–993.
[11] I. Mele and F. Crestani. 2017. Event Detection for Heterogeneous News Streams. In *NLDB '17*. Springer International Publishing, Cham, 110–123.
[12] S. Petrović, M. Osborne, and V. Lavrenko. 2010. Streaming First Story Detection with Application to Twitter. In *HLT '10*. Association for Computational Linguistics, Stroudsburg, PA, USA, 181–189.
[13] A. Ritter, S. Clark, Mausam, and O. Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *EMNLP '11*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1524–1534.