



TweetStream2Story: Narrative Extraction from Tweets in Real Time

Mafalda Castro^{1,2(✉)}, Alípio Jorge^{1,2}, and Ricardo Campos^{1,3}

¹ INESC TEC, Porto, Portugal

`mafalda.r.castro@inesctec.pt`

² FCUP, University of Porto, Porto, Portugal

`amjorge@fc.up.pt`

³ Polytechnic Institute of Tomar, Ci2 - Smart Cities Research Center,
Tomar, Portugal

`ricardo.campos@ipt.pt`

Abstract. The rise of social media has brought a great transformation to the way news are discovered and shared. Unlike traditional news sources, social media allows anyone to cover a story. Therefore, sometimes an event is already discussed by people before a journalist turns it into a news article. Twitter is a particularly appealing social network for discussing events, since its posts are very compact and, therefore, contain colloquial language and abbreviations. However, its large volume of tweets also makes it impossible for a user to keep up with an event. In this work, we present TweetStream2Story, a web app for extracting narratives from tweets posted in real time, about a topic of choice. This framework can be used to provide new information to journalists or be of interest to any user who wishes to stay up-to-date on a certain topic or ongoing event. As a contribution to the research community, we provide a live version of the demo, as well as its source code.

Keywords: Narrative extraction · Natural language processing · Twitter

1 Introduction

Social media is a powerful tool that can provide great insights into a variety of topics. Using Twitter posts as a source for extracting narratives may bring us different information than a news article does, from the people who experience an event first-hand. The Twitter platform is a very helpful tool for journalists [4, 7], however, its colloquial language and the large volume of *tweets* (6000 *tweets* are posted every second, on average [12]) makes it impractical to keep up with an event. For this reason, obtaining the most relevant tweet posts turns out to be of the utmost importance. To achieve this, researchers have presented a variety of methods regarding the automatic summarization of *tweet* streams [1, 3, 6, 9, 10, 15], although none of these had narrative extraction in mind [11]. Recently, Campos et al. [2] have proposed the Tweet2Story

framework¹, which performs the automatic narrative extraction from a bundle of *tweets*. However, this framework doesn't work in real time, requiring users to previously collect and process the *tweets* that will be given as input. In this paper, we present TweetStream2Story², an extension of Tweet2Story that fills this gap, by incorporating the real-time collection of *tweets* on a given topic, as well as the automatic extraction of narratives from these *tweets*. As a further contribution to the research community, we make the source code of our project available, thus challenging researchers to use and expand it³.

2 TweetStream2Story

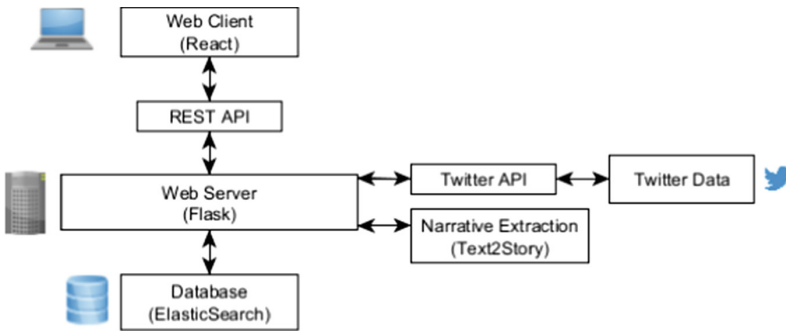


Fig. 1. Architecture overview of TweetStream2Story

Figure 1 depicts the architecture of the TweetStream2Story framework. The first step of this pipeline is to issue a topic, a query (e.g. Denmark Shooting) in the user interface (the web client) to search for related *tweets*. The user must also provide the time period for the collection of *tweets* (e.g. July 4 2022, 4:30 pm to July 5 2022, 12:30 am), which will be divided into time windows of a specified duration (e.g. 2 h). The narrative will be generated in two modes: in the global mode, each time window uses *tweets* since the start of the topic ([4:30 pm–6:30 pm], [4:30 pm - 8:30 pm], and on and so forth). In the interval mode, instead, each time window uses *tweets* posted strictly during that time window ([4:30 pm–6:30 pm], [6:30 pm–08:30 pm], and so on and so forth). Once this information is defined, we proceed by obtaining the collection of related *tweets* using either the Twitter API's Filtered stream, in case the user wants to follow up *tweets* posted in real-time, or the Full-archive search, to look for events in the past. For every collected *tweet*, a preprocessing stage, involving hashtags removal, hyperlinks and emojis is applied. Similar tweets, with a term-frequency cosine similarity higher

¹ <http://tweet2story.inesctec.pt/>.

² <http://tweetstream2story.inesctec.pt/>.

³ <https://github.com/LIAAD/tweetstream2story>.

than 80% are also removed. The resulting set is then stored in Elasticsearch, a flexible document-oriented database. To reduce the amount of *tweets*, we then proceed with a summarization-like step where only the most relevant tweets are taken into account. To do this, we use, as in Rishab S. et al. [14], the Okapi BM25 function [8] as our IR model, a function that estimates the relevance of a document to a given search query, and by that, retrieve the top- X most relevant *tweets* belonging to a particular time window, where X equals 50 (a trade-off between the number of *tweets* and their Precision). Following, we proceed to use these *tweets* as input to the Text2Story narrative extraction pipeline. In the coming section, we demonstrate how such pipeline is used to create a visual representation of the topics narrative.

3 Demo

In this section, we describe the main features of this [demo](#). Its live version can be used by anyone who wishes to extract the narrative of a specified topic from *tweets* posted either in real time or in the past. The first step for generating a narrative requires the user to input a topic of their interest. After typing in a topic and clicking on the Extract Narrative button, a modal opens where the user can specify parameters such as the desired language, the duration of each time window (e.g. 2h), and the mode for collecting tweets (e.g. streaming, past *tweets*). Currently, the only languages supported are English and Portuguese. Although the focus of this work is the retrieval of *tweets* posted in real time, our framework also allows retrieving past *tweets*. In this case, however, the user must provide their Twitter API credentials, which will not be stored, but discarded as soon as they're used. Topics are automatically added to a private list of topics, owned by the user, allowing them to keep track of its status, visualize the corresponding narrative or perform actions such as stopping the retrieval of tweets or deleting the topic from the list. [Figure 2](#) shows the interface for the list of topics therein presented.









#	Topic	Language	Start Date	End Date	Status	Options
1	seleção portugal espanha	 PT	2022-06-02 19:45	2022-06-02 22:15	 Completed	 Open 
2	Queen Elizabeth death	 EN	2022-09-10 15:27	2022-09-11 22:00	 In progress...	 Open  Stop

Fig. 2. Topics list

By clicking on a topic, users are offered the chance to visualize its narrative through a timeline, as shown in [Fig. 3](#). Added, they can choose between the two modes mentioned before: global view or interval view. In each mode, the timeline is divided into time windows with the duration previously specified by the user, where each one shows its respective narrative and information. The default

visualization of a narrative is the knowledge graph, which shows actors as nodes and semantic relationships as the edges between the actors. It also highlights in yellow the nodes that weren't present in the previous time window, as a way for the user to quickly see new information. Other modes of visualization include the list of tweets that were used to generate the narrative, as well as the list of actors, as can be seen in Fig. 4. Further advanced analysis can also be performed by viewing and downloading the information in formal representations, as is the case of DRS annotations [5] or the Text2Story annotation [13].

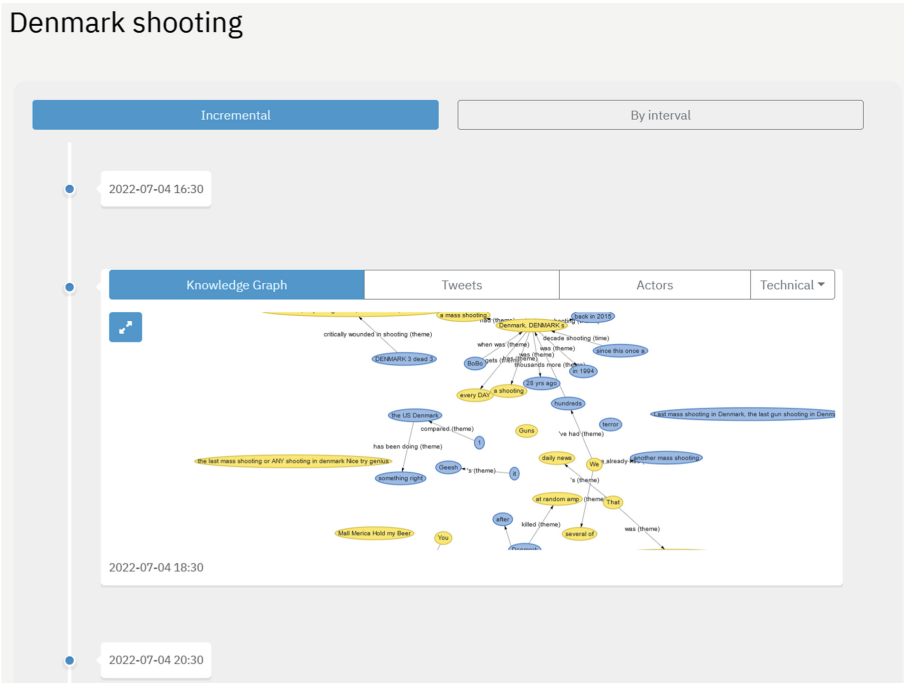


Fig. 3. Timeline representation of a topic

As a means to demonstrate Twitter’s potential for narrative extraction, some examples of topics in both Portuguese and English, are pre-loaded in the interface. Figure 4 shows a visual representation of the topic *Denmark Shooting*, an event that occurred in Copenhagen in 2022. This knowledge graph is able to capture information about the number of deaths, critically wounded people, and previous shootings in the country. These examples are able to demonstrate Twitter’s usefulness as a news source, as the information contained in some of the extracted actors and relations is able to complement a news article.

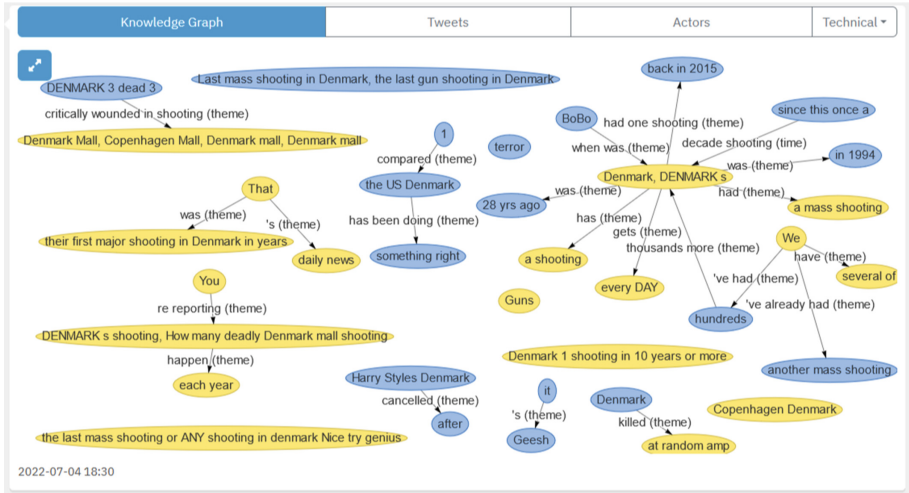


Fig. 4. Narrative representation of a time window

4 Conclusions and Future Work

In this paper, we have presented a framework that allows the automatic collection of *tweets* and extraction of their narrative elements, TweetStream2Story. This tool can be beneficial not only for journalists, but also for users interested in an ongoing event. Some of its limitations are the requirement for a user to enter their Twitter API credentials when generating narratives from events in the past, and the long computational time to extract the narrative. In the future, we would like to improve the quality of the results by incorporating techniques such as irony detection and offensive speech, as a way to filter out some *tweets*. We also plan on improving the user-system interactions, as well as implementing an abstractive summarization approach, in order to use original content as the source of the narratives.

Acknowledgements. This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. The authors Alípio Jorge and Ricardo Campos are financed by the project Text2Story, financed by the ERDF - European Regional Development Fund through the Norte Portugal Regional Operational Programme - NORTE 2020 under the Portugal 2020 Partnership Agreement and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within project Text2Story, with reference PTDC/CCI-COM/31857/2017 (NORTE-01-0145-FEDER-031857).

References

1. Alsaedi, N., Burnap, P., Rana, O.: Automatic summarization of real world events using twitter. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 10, no. 1, pp. 511–514 (2021). <https://ojs.aaai.org/index.php/ICWSM/article/view/14766>
2. Campos, V., Campos, R., Mota, P., Jorge, A.: Tweet2Story: a web app to extract narratives from twitter. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 270–275. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_32
3. Chellal, A., Boughanem, M.: Optimization framework model for retrospective tweet summarization. In: Haddad, H.M., Wainwright, R.L., Chbeir, R. (eds.) Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC 2018, Pau, France, 09–13 April 2018, pp. 704–711. ACM (2018). <https://doi.org/10.1145/3167132.3167210>
4. Jurkowitz, M., Gottfried, J.: Twitter is the go-to social media site for U.S. journalists, but not for the public (2022). <https://pewrsr.ch/3yqfRP>. Accessed 31 Aug 2022
5. Kamp, H., Reyle, U.: From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Springer, Dordrecht (1993). <https://doi.org/10.1007/978-94-017-1616-1>
6. Li, Q., Zhang, Q.: Twitter event summarization by exploiting semantic terms and graph network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 17, pp. 15347–15354 (2021). <https://ojs.aaai.org/index.php/AAAI/article/view/17802>
7. MuckRack: The state of journalism 2022 (2022). <https://muckrack.com/blog/2022/05/18/2022-state-of-journalism-on-twitter>. Accessed 31 Aug 2022
8. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.* **3**, 333–389 (2009). <https://doi.org/10.1561/1500000019>
9. Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., Ghosh, S.: Extracting situational information from microblogs during disaster events: a classification-summarization approach. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, 19–23 October 2015, pp. 583–592. ACM (2015). <https://doi.org/10.1145/2806416.2806485>
10. Rudra, K., Goyal, P., Ganguly, N., Imran, M., Mitra, P.: Summarizing situational tweets in crisis scenarios: an extractive-abstractive approach. *IEEE Trans. Comput. Social Syst.* **6**(5), 981–993 (2019). <https://doi.org/10.1109/TCSS.2019.2937899>
11. Santana, B., Campos, R., Amorim, E., Jorge, A., Silvano, P., Nunes, S.: A survey on narrative extraction from textual data. *Artif. Intell. Rev.* (2023). <https://doi.org/10.1007/s10462-022-10338-7>
12. Sayce, D.: The number of tweets per day in 2022 (2022). <https://www.dsayce.com/social-media/tweets-day/>. Accessed 25 Sept 2022
13. Silvano, P., Leal, A., Cantante, I., Oliveira, F., Mario Jorge, A.: Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In: Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pp. 1–13. Association for Computational Linguistics, Groningen, The Netherlands (online) (2021). <https://aclanthology.org/2021.isa-1.1>

14. Singla, R., Modha, S., Majumder, P., Mandalia, C.: Information extraction from microblog for disaster related event. In: Proceedings of the First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness co-located with European Conference on Information Retrieval, SMERP@ECIR 2017, Aberdeen, UK, 9 April 2017. CEUR Workshop Proceedings, vol. 1832, pp. 85–92. CEUR-WS.org (2017). <http://ceur-ws.org/Vol-1832/SMERP-2017-DC-DAICT-IR-LAB-Retrieval.pdf>
15. Wang, Z., Shou, L., Chen, K., Chen, G., Mehrotra, S.: On summarization and timeline generation for evolutionary tweet streams. *IEEE Trans. Knowl. Data Eng.* **27**(5), 1301–1315 (2015). <https://doi.org/10.1109/TKDE.2014.2345379>