Data in Brief 38 (2021) 107344

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Annotated dataset of history-related tweets

Yasunobu Sumikawa[a,1,*], Adam Jatowt[b,1]

[a] Takushoku University, Japan
[b] University of Innsbruck, Austria

ARTICLE INFO

Article history:
Received 5 April 2021
Revised 1 September 2021
Accepted 2 September 2021
Available online 4 September 2021

Keywords:
Digital history
Tweets
Hashtags
Hashtag categories
Temporal analysis

ABSTRACT

In this article, we present a dataset containing history-related content obtained from social media. It contains hashtags and tweets that include these hashtags, as well as the results of third party tools applied to the tweets that include extracted entities, years, and url categories, and the categories for the history-related hashtags we used to crawl the tweets. We collected the tweets from Twitter official API using hashtag-based crawling. The crawling process had been performed from March 2016 to July 2018. During the crawling, we applied a bootstrapping approach which is an iterative process of collecting tweets using a small set of seed hashtags, and a manual inspection of newly acquired hashtags that co-occur with the seed hashtags to include those they are related to history. Finally, we collected 147 history-related hashtags and 2,370,252 tweets. We then defined 6 categories for the collected hashtags after their manual investigation. The presented dataset could be useful for further analysis on how people refer to history in Twitter, for collecting new history-related tweets, for training classifiers to detect history-related tweets, or for further investigations of the proposed hashtag categories.

* Corresponding author.
E-mail addresses: ysumikaw@cs.takushoku-u.ac.jp (Y. Sumikawa), adam.jatowt@uibk.ac.at (A. Jatowt).
[1] https://mobile.twitter.com/HistoChatbot

https://doi.org/10.1016/j.dib.2021.107344

Specifications Table

Subject	History
Specific subject area	Digital history and collective memory analysis
Type of data	Table (csv files)
How data were acquired	We implemented a python program to use Twitter's official API.
Data format	Raw Filtered
Parameters for data collection	To use Twitter API, we utilized history-related hashtags as parameters. The initial hashtags that are listed on the url shown below were collected based on historian's selections. The historian's selection was done by a manual inspection of Twitter crawling and crowdsourcing. Finally, 60 commonly used history hashtags were collected. https://www.historians.org/publications-and-directories/perspectives-on-history/summer-2013/history-hashtags-exploring-a-visual-network-of-twitterstorians
Description of data collection	The crawling process was performed from March 2016 to July 2018. During the crawling, we applied a bootstrapping approach which is an iterative process of collecting a large number of tweets using a small set of seed hashtags, and a manual inspection of newly acquired hashtags that co-occur with the seed hashtags to judge whether they are related to history or not.
Data source location	The Web The initial hashtags: https://www.historians.org/publications-and-directories/perspectives-on-history/summer-2013/history-hashtags-exploring-a-visual-network-of-twitterstorians URL categories: https://www.webshrinker.com/website-category-api/
Data accessibility	Public repository Repository name: Zenodo Data identification number: zenodo.4657223 Direct URL to data: https://doi.org/10.5281/zenodo.4657223
Related research article	Sumikawa, Y., Jatowt, A.: Analyzing History-related Posts in Twitter, Int. J. Digit. Libr. 22, 105–134 (2021). https://doi.org/10.1007/s00799-020-00296-2 Sumikawa, Y., Jatowt, A., M., Düring: Digital History meets Microblogging: Analyzing Collective Memories in Twitter, In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18). Association for Computing Machinery, New York, NY, USA, 213–222 (2018). https://doi.org/10.1145/3197026.3197057

Value of the Data

- The dataset provides seed data which could be used to collect history-related contents from microblogging datasets, to collect hashtags that would be used to increase the number of history-related contents using this dataset, to analyze the process of how public history is shaped in social media and how the present influences the remembrance of history in our society, and to train machine learning models for not only detecting history-related tweets in Twitter but also for estimating their categories and other characteristics.
- Historians and social scientists working in a broad field of digital humanities and computer scientists can benefit from this dataset by either using it directly as the basis of their analysis or by utilizing its data for further crawls such as using the found history-related hashtags that we provide as seed hashtags, or for training classifiers to detect history-related tweets.
- There are several potential applications using our dataset: 1) Analyzing and visualizing what aspects of history the Twitter users are interested in, and 2) Creating history-focused chatbots such as HistoChatbot[1] for disseminating historical knowledge and for entertaining users.

[1] https://mobile.twitter.com/HistoChatbot

Table 1
Database files.

File name	Content	Columns
History_hashtag_categories.csv	Hashtag categories and their hashtags.	Categories: Name of a category of a history-related hashtag Hashtags: Hashtags belong to the category of the same row. All hashtags in the same category are comma-separated.
Tweet_type_hashtags.csv	Tweet IDs and history-related hashtags included in each tweet ID.	Tweet ID: Tweet ids whose hashtags are related to history. Type: This indicates a tweet or retweet. The retweet is identified by the text that starts with "RT". Hashtags: History-related hashtags can be used to crawl the tweet ID in the same row. All hashtags included in the same tweet ID are comma-separated.
Entity.csv	Entities of tweets.	Tweet ID: IDs of tweets whose hashtags are related to history. Entities: Entities of the Tweet ID listed in the same row. NULL means that no entities were extracted by AIDA.
Time_references.csv	Time references of tweets	Tweet ID: Tweet ids whose hashtags are related to history. TReferences: Explicit time references of the Tweet ID listed in the same row. All the references are identified by HeidelTime. NULL means that no temporal references were extracted by HeidelTime tagger.
WebCat.csv	Web categories of tweets	Tweet ID: Tweet ids whose hashtags are related to history. WebCat: Results of web shrinker API. If these categories are NULL it means that no categories were extracted.

- This dataset includes not only IDs of collected history-related microblogging data but also the results of applying third party annotation tools for extracting entities, url categories and temporal references from the microblogging data.

1. Data Description

The published dataset (see metadata in Table 1) consists of five types of data. The first type includes 147 history-related hashtags and 6 categories for the hashtags in the History_hashtag_categories.csv file.

The second type includes tweet IDs and history-related hashtags used to crawl each tweet in Tweet_type_hashtags.csv. As we must respect Twitter's Terms of Service licensing agreement[2] and copyright, any raw data including tweet text, timestamp of the created tweets, url, and hashtag texts of the collected tweets have been removed.

As for the third type, the Entity.csv file includes entities identified by the named entity recognition and disambiguation tool, called AIDA [1]. In this file, each line includes a tweet ID and entities included in the corresponding tweet or NULL. NULL indicates that AIDA has not found any entities in the textual content of a tweet ID.

The fourth type are time references identified by HeidelTime [2] in Time_references.csv. Typically, there are 2 types of temporal expressions: explicit and implicit temporal expressions.

[2] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

Table 2
Statistics of the whole dataset published in this paper.

Number of hashtag categories	6
Number of history-related hashtags	147
Total Number of tweet IDs	2,370,252
Number of tweets	882,977
Number of retweets	1,487,275
Size of entity set	76,585
Size of Web category set	335
Period of timestamps	8 Mar. 2016 ~ 2 Jul. 2018
Period of time references	8156BC ~ 2029

The first expression type is a concrete timepoint or time period, such as “1900” or “1890s”, while the second one is relative such as “yesterday” or “two years ago”. This file includes only explicit temporal expressions; all found implicit (relative) temporal expressions were converted to the explicit (absolute) ones by using tweet timestamps as a reference.

Lastly, the fifth type includes the categories of web sites whose URLs are included in tweets. The categories were obtained from Webshrink API (website-category-api)[3] in WebCat.csv. As the API outputs the name of category, score, and the confident field, we store the category names and their scores if the confident flag is True.

2. Experimental Design, Materials, and Methods

Collecting history-related hashtags and tweets. To collect tweets we used Twitter’s official search API[4]. In Twitter there are three kinds of tweets: tweets, retweets and quote tweets. Tweet is an original text a Twitter user issues as a post. A retweet is a copy of an original tweet. Finally, a quote tweet adds new texts to a copy of the content of another tweet. In this dataset there are 1,877 (0.2%) quote tweets. This dataset treats all quote tweets as original tweets since the former include additional information/text compared to the content of the quoted tweets.

To make sure that we collected tweets which refer to the past or are related to collective memory of past events/entities, we performed a bootstrapping procedure including hashtag based crawl. We first defined seed hashtags by using historical hashtags selected by experts (e.g. **#history, #WmnHist, #HistoryTeacher**) and ones used commonly when referring to the past: **#historicalevent, #otd, #onthisday, #throwbackthursday, #thisdayinhistory**. We then used these hashtags as queries for Twitter’s official API to collect tweets including these hashtags. This allowed us to find other history-related hashtags. To do this, we first extracted all hashtags from the collected tweets. We then performed a manual investigation to check whether the collected hashtags are history-related or not excluding the seed hashtags. During this investigation, three persons manually checked if collected new hashtags are history-related or not. The investigation process was as follows: checking if the numbers of tweets using the history-related hashtag candidates were more than 5 tweets, reading the tweet texts to check whether they are related to history content. After all the three investigators agreed that the collected tweet texts were related to history, we added the collected hashtags into the seed hashtags in the bootstrapping process. Finally, we collected the 147 history-related hashtags.

Table 2 shows the statistics of the collected tweets. In this table we distinguish tweets and retweets by whether their texts start with RT or not. If a tweet text starts without RT, we consider the tweet ID as a tweet; otherwise, we treat the tweet as a retweet. As Twitter’s official API allocates unique tweet IDs for all tweets and retweets, the “Total Number of tweet IDs” represents both tweets and retweets.

[3] https://www.webshrinker.com/website-category-api/
[4] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

Entity collection. We employ AIDA that detects and disambiguates entities by linking phrases in text with their corresponding Wikipedia articles. Before applying AIDA we evaluated its performance on a random sample of 50 tweets selected from our dataset. Based on the manual inspection of the results, the obtained precision and recall were 96.0% and 77.4%, respectively.

Web category collection. We use WebShrinker API to find Web categories. The output of the API includes a confident flag, label, and score. The confident flag indicates whether the majority of the analyzed content relates to this category; thus, it is a flag. The label is the name of the selected category. The score is a floating point number representing how much confidence is given to the category selected as the label. We store the labels and their scores if their confident flag is True. The API assigns to each URL multiple labels from 404 predefined categories based on IAB Tech Lab Content Taxonomy. We applied this API to randomly sampled 32k tweets and 36k retweets that contain URLs, and obtained the linked Web sites classifying the tweets into different types.

Time-reference collection. We use HeidelTime, which is an effective temporal tagger with a specialized option for tweet processing to extract temporal references. HeidelTime outputs normalized temporal expressions according to the TIMEX3 annotation standard. We evaluated its accuracy on our dataset by randomly sampling 100 tweets that contain temporal references. We have confirmed that this tool obtained 88.9% of precision and 98.1% recall.

Hashtag Category Definition. The hashtag categories are defined from manual investigation of a large sample of collected tweets [3,4].

Ethics Statement

To respect Twitter's Developer Policy[5], this dataset excludes all raw tweet contents without tweet IDs. Other data are collected from manual definition or third-party tools. Personal identifying information was removed from all tweet contents and metadata.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

CRediT Author Statement

Yasunobu Sumikawa: Data curation, Writing – original draft; **Adam Jatowt:** Writing - review & editing.

Acknowledgments

This work was supported in part by the MEXT Grant-in-Aids (#17H01828 and #19K20631).

References

[1] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, Robust disambiguation of named entities in text, in: EMNLP'11, 2011, pp. 782–792.
[2] J. Strötgen, M. Gertz, Temporal tagging on different domains: Challenges, strategies, and gold standards, in: LREC'12, ELRA, Istanbul, Turkey, 2012, pp. 3746–3753.
[3] Y. Sumikawa, A. Jatowt, Analyzing history-related posts in twitter, Int. J. Digit. Libr. 22 (2021) 105–134, doi:10.1007/s00799-020-00296-2.
[4] Y. Sumikawa, A., M. Jatowt, Düring: digital history meets microblogging: analyzing collective memories in twitter, in: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18), Association for Computing Machinery, New York, NY, USA, 2018, pp. 213–222, doi:10.1145/3197026.3197057.

[5] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases