International Journal on Digital Libraries (2021) 22:105–134
https://doi.org/10.1007/s00799-020-00296-2

Analyzing history-related posts in twitter

Yasunobu Sumikawa[1] · **Adam Jatowt**[2]

Received: 30 May 2019 / Revised: 17 August 2020 / Accepted: 23 September 2020 / Published online: 28 October 2020

Abstract
Microblogging platforms such as Twitter have been increasingly used nowadays to share information between users. They are also convenient means for propagating content related to history. Hence, from the research viewpoint they can offer opportunities to analyze the way in which users refer to the past, and how as well when such references appear and what purposes they serve. Such study could allow to quantify the interest degree and the mechanisms behind content dissemination. We report the results of a large scale exploratory analysis of history-oriented posts in microblogs based on a 28-month-long snapshot of Twitter data. The results can increase our understanding of the characteristics of history-focused content sharing in Twitter. They can also be used for guiding the design of content recommendation systems as well as time-aware search applications.

Keywords Social media analysis · History · Collective memory · Twitter

1 Introduction

History is regarded as knowledge that plays a special role in our society. This is because the comprehension of history is useful for multiple reasons. First, one can better understand the processes impacting the present world. Second, history forms the basis for the development of coherent national and local identities. Third, history offers support for decision making and provides guidance as for what can await us in the future [1,23]. Due to these and other reasons, history is one of the key subjects that are taught in elementary schools as well as in the subsequent stages of education.

Recently, social media and microblogs in particular have been often used as a convenient source for understanding public attitude towards entities or events (e.g., the US American elections [57]). Microblogs are also a platform useful for finding and sharing history-related content. Computational studies of references to the past in microblogs can then offer us novel perspectives for understanding the formation of collective memories and the pursuit of public history.

✉ Yasunobu Sumikawa
ysumikawa@acm.org

Adam Jatowt
jatowt@gmail.com

[1] Tokyo Metropolitan University, Hachioji, Tokyo, Japan

[2] Kyoto University, Kyoto, Japan

Collective memory analyses based on large-scale data and using computational methods have been already carried either on news article collections [5,16] or Wikipedia data [19,20,35,36]. However, when it comes to microblogs, little research has been done. One notable project is the analysis of the memories related to the First World War in Twitter [14] from the multi-cultural perspective. Our work also focuses on Twitter which constitutes a popular social media platform frequently utilized for a variety of studies in the computational social sciences and other domains. The analysis we perform has exploratory character aiming to offer broad investigation of practices of sharing history-related content in microblogging platforms.

The following questions are considered in our study:

1. How do users write about history in Twitter?
2. How does the time horizon of history-related references look like?
3. In what way are collective memories expressed in Twitter?
4. What are the key tweeted and re-tweeted past events and entities?
5. How different are collective memories expressed in tweets from the ones in re-tweets?

These and other related questions are investigated based on a compiled dataset of tweet messages which were issued from March 2016 to July 2018. We collect such posts by

searching for tweets which contain history-related hashtags. To increase the coverage, we apply a bootstrapping as an iterative process of collecting relevant hashtags starting from an initial set of seed history-focused hashtags. Thanks to this procedure, we collected the sufficient number of history-related hashtags, which allow us to gather over 2 million tweets which contain different kinds of references to history.

Based on the collected data, we then examine the characteristics of history-related tweets. We study their time horizons, mentioned entities, hashtag popularities as well as several other related aspects. Moreover, we describe our novel taxonomy of history-related hashtags and we analyze different hashtag categories. By this, we try to organize and provide structure to user activities related to referencing, evaluating and sharing history-related information in social networks like Twitter.

Besides answering specific research questions in this study, the results of our analysis can be useful for several practical applications. First, specialized content detection and recommendation systems can be better designed thanks to the results we report. Their objective would be to facilitate sharing of historical knowledge. Historical content recommendation in social media is an attractive and informal way for learning history. Building effective, dedicated recommendation systems could be supported based on understanding of the characteristics and types of popular history-related content in social media and the context in which this content is shared. Indeed, several existing projects already employ online social platforms like Twitter to stimulate interest in history and for teaching history.[1] An interesting idea is automatic content dissemination enabled by *history-focused chatbots* such as HistoChatbot.[2] Tweets, due to relatively short content and the simple yet effective methods for measuring their popularity (e.g., re-tweet counts and user response analysis), could constitute a useful source of data for such systems.

Naturally, some history-focused tweets are directly triggered by current events or current popular entities. Studying their formation and popularity could be useful for understanding the conditions and circumstances that would allow for "*historification*" of different types of documents. In practice, this would mean recommending relevant historical references and grounding for any present events and topics mentioned in these documents.

Besides providing answers to the research questions on history-related content dissemination in social media, our work may also offer clues about collection building for historians or other researchers who are interested in using tweet collections. The proposed categorization of history-related hashtags could be used for generating collections that contain content of special characteristics. In this context, we also discuss particular types of tools that can be used (temporal tagger, NER method) for effective analysis of collected datasets.

To sum up, we make the following contributions:

1. We study how users refer to history in social networks based on collected large scale data.
2. We perform tweet- and re-tweet-based analyses.
3. We provide novel findings which offer a better understanding of how collective memories are maintained and formed in microblogging.
4. We propose novel categorization of historical references in Twitter.
5. We outline novel research directions and potential applications that can utilize history-related content in microblogs.
6. We release our dataset of history-related tweets for further research.

This work is an extended version of the paper published at the JCDL 2018 conference [55]. We analyze here larger datasets (close to three years long span of data collection instead of 1 year as in our previous work). This allows us to undertake comparative analyses for different years (2016, 2017 and 2018). Besides the larger scale and comparative focus of this work, we also contrast the results obtained from tweets with those coming from re-tweets. This allows for pinpointing differences between active formulation of texts containing remembrances with their passive dissemination along with social networks. Finally, in comparison with the JCDL 2018 paper we analyze URLs included in tweets and show the results in this paper.

The remainder of this paper is structured as follows. We present related work in the next section. In Sect. 3, we detail the data collection and processing. Section 4 describes the findings of our analysis, while the next section introduces our novel categorization of hashtags and provides the results of the related analyses. We then provide discussions in Sect. 6. The last section concludes the paper and describes our future work.

2 Related work

In this section, we first start with the overview of temporal information retrieval studies and temporal text analysis (Sect. 2.1) and with surveying works on Twitter data analysis (Sect. 2.2) as our study uses temporal references in tweets. We then focus on broad studies of collective memory using computational approaches in Sect. 2.3.

[1] https://twitter.com/RealTimeWWII, https://twitter.com/civilwarwp, https://twitter.com/1948War, https://twitter.com/samuelpepys

[2] https://twitter.com/HistoChatbot

2.1 Temporal analysis

The current Web contains numerous digital archives including historical images, documents and so on due to intensive digitization efforts carried out over the last years. Due to the ever increasing amount of temporal data, analyzing temporal information has become an important process in information retrieval (IR) to improve satisfaction of users. Recently, several kinds of studies were undertaken in the broad area of Temporal IR (T-IR), for example, detecting temporal expressions or information [28], retrieving history-related images [13], organizing information by creating timelines [3,17,29], or future-related IR [7,34,50]. A detailed survey of T-IR is given in [11].

Similar to our study, several past-oriented temporal analyses have been performed. These could be roughly grouped into several sub-areas of T-IR: supporting users to perform retrieval of past specified data, extracting useful past data, and supporting or understanding historical sciences in general.

As for the supporting data search and retrieval, various methods and algorithms to assist users in finding past content were proposed [9,47,52]. For example, Singh *et al.* designed an IR framework to support historians in their searches [52]. According to the literature, if historians investigate an entity, they first try to see it from a big picture. Then, they further search for content on the entity according to some of its specific aspects. Thus, supporting historians' information seeking is useful to indicate not only important time information but also display several kinds of aspects. Bogaard *et al.* proposed a data-driven partitioning process to identify user interests and search behavior based on interactions with a historical newspaper collection spanning 400 years that is available from the National Library of the Netherlands [9]. They confirmed that their approach can detect user interests and observed that the related search behavior varies within the different parts of the collection. Abujabal and Berberich [2] proposed method to identify important past as well as future events based on frequent itemset mining and mutual information on sentences containing named entities and temporal expressions.

Works on finding analogical items over different temporal scopes are also related to our study. Zhang et al. proposed a framework for detecting counterparts of entities over time [61]. This framework bridges two different vector spaces that are created for different time-ranges such as [1900–1950] and [1960–2010] by applying an automatically learnt transformation matrix. The transformation matrix maps an entity in one vector space into the other one. The authors extend this approach to make use of hierarchical cluster structures [62]. In general, mining history-related knowledge is another popular direction of study. For example, several works try to find beneficial information from large amounts of data by evaluating the significance of historical entities [31], timestamping entities [32], analyzing trends [29], or trying to predict future from past events [33,49,50].

2.2 Twitter analysis

Twitter is one of the most popular social media platforms to share information. As a tweet can have at most 280 characters, this platform poses several challenges caused by the short content of messages. For example, there are studies extending traditional IR/NLP techniques designed for long documents such as news articles to fit short texts, e.g., identifying central topic model from tweet streams [48], summarizing tweets [22], retrieving opinions [21], detecting community [8], and building corpora [38,42,46]. In addition, Twitter contains not only texts but also unique features such as hashtags, followers and followees (i.e., Twitter users who follow or are followed by a particular user), and URLs. Using these features, past studies focused on (among others) automatic hashtag labeling by hashtag-based pooling tweets [43], analyzing factors affecting response [15], readability of crisis communications [56], language diversity [41], language and locations [59], detecting influencers in Twitter [60], classifying user's temporal intention when sharing resources [51], ranking users [58] or meme tracking in blogosphere [40].

As discussed above, many Twitter-related studies use unique Twitter's features, yet what these studies usually lack is a deep consideration of historical aspects.

2.3 Collective memory analysis

The concept of *collective memory* (or *social memory*) popularized by Halbwachs [25,26] describes the shared reflection of the past within social groups. Collective memory can be contrasted with *collective amnesia* defined by Jacoby [30] as forceful or unconscious suppressions of memories, especially those related to disgraceful or inconvenient events for a particular social group or nation. In a similar fashion to personal memory [18], social memory is known to thin out over time and to be subject to temporal variations following the occurrence of memory triggers such as sudden events or anniversaries [5,36,37].

Studies of collective memory can help us to understand the mechanisms of forgetting and remembering as well as explain the role of the history and the past in our lives. In addition, they have direct implications on the archival selection by memory institutions such as national or dedicated archives [37]. Traditionally, research on collective memory has been based on manual approaches and small-scale investigations of personal accounts and the activities of political and cultural institutions. There is still relatively little literature on the use of computational approaches for the quantification of the characteristics of social memory over large text datasets.

Cook et al. [16] investigated the decay of fame over time on the basis of the collection of news articles that span the twentieth century. Au Yeung and Jatowt [5] studied the way in which past year mentions appear in the datasets of recent news articles in order to understand which years are forgotten and which remain remembered, as well as the main topics associated with the remembering of past years.

Wikipedia has been quite often used as a reflection of collective memories and their formation processes. Ferron and Massa [19] and Kanhabua et al. [36] proposed to use Wikipedia as a global memory space. The latter work focused on memory triggers that cause forgotten or vaguely remembered events to be brought back into social attention. Anniversaries are natural examples of memory triggers. In another case, current events may also serve as triggers of the memories of similar, past events. García-Gavilanes et al. [20] revealed viewership statistics of Wikipedia articles on aircraft crashes and focused on memory triggering patterns. Miz et al. [44] proposed a new method that allows learning and remembering collective memories in an unsupervised manner by analyzing the Wikipedia Web network and hourly viewership history of its articles. The interests of Wikipedia visitors were also studied in [35] focusing on Wikipedia articles on historical persons. The authors have also investigated connectivity of Wikipedia articles about historical persons. Graus et al. [24] investigated about 80,000 entities emerging in online text streams before they got incorporated into Wikipedia analyzing in this way the processes behind collective memory formation.

Collective memories have been also researched in the context of particular items or objects. Strötgen et al. [53] performed large-scale worldwide analysis of street names with date references according to the intuition that temporal streets are frequently used to commemorate important events of different regions. Similarly, Nielek et al. [45] analysed street names distributions as a window to nation-level collective memory in Poland. Candia et al. [12] analyzed temporal decay of the attention received by cultural products such as academic articles, patents, songs, movies and biographies. The authors showed that the attention received by cultural products decays following a universal biexponential function and explained it by proposing a mathematical model based on communicative and cultural memory. The formation of collective memory has been also recently modeled by simulating opinion dynamics of collective agents including phenomena such as homophily [10]. Koutlis et al. [39] studied collective memory dynamics with regard to song recognition levels leveraging chart data, YouTube views, Spotify popularity and forgetting curve dynamics.

Despite the above-listed efforts, to the best of our knowledge, few researches focus on history-oriented studies in microblogging scenarios. Memory dynamics was investigated in Twitter data in [4] with regard to particular attributes of hurricanes. The authors tracked the use of ngrams involving hurricane name mentions and found that the most damaging and deadly storms of the 2010s generated the most attention and were remembered the longest. In another work, commemoration of the First World War was studied in relation to diverse countries [14]. In contrast to these works, we use relatively large size data (at least, for history-related studies), longer time spans, and we investigate multiple aspects ranging from the types of references, intensity of remembering, key entities, dates, temporal patterns and so on. Lastly, our analysis uses three temporal snapshots of data what allows comparison of collective memories in different years.

3 Data collection

This section describes the data collection and preprocessing procedures as well as general statistics of the dataset used for analysis. We also provide few basic statistics and example results of entity mention detection.

Collecting hashtags and tweets. We used the Twitter official search API[3] provided by Twitter to collect tweets. Note that three kinds of tweets are typically found in Twitter: tweets, re-tweets and quote tweets. A tweet is an original text issued as a post by a Twitter user. A re-tweet is a copy of an original tweet for the purpose of propagating the tweet content to more users (i.e., one's followers). Finally, a quote tweet copies the content of another tweet and allows also to add new content. A quote tweet is sometimes called a re-tweet with a comment. In this work, we simply treat all quote tweets as original tweets since they include additional information/text. There were, however, only 1,877 (0.2%) tweets recognized as quote tweets in the collected data.

To collect tweets that refer to the past and are related to collective memory of past events/entities, we performed hashtag based crawling together with a bootstrapping procedure. At the beginning, we gathered several historical hashtags selected by experts (e.g., `#HistoryTeacher`, `#history`, `#WmnHist`)[4]. In addition, we prepared several hashtags that are commonly used when referring to the past: `#onthisday`, `#thisdayinhistory`, `#throwbackthursday`, `#otd`. We then collected tweets that contain these hashtags by using Twitter's official search API. The procedure of the bootstrapping approach is shown in Procedure 1. $T1$ and $T2$ are conditions for collecting new seed hashtags and for stopping the tweet crawling, respectively.

These conditions depend on the data collection policy, such as $T2$ can be bound to the pre-determined tweet collec-

[3] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

[4] http://blog.historians.org/2013/08/history-hashtags-exploring-a-visual-network-of-twitterstorians/

Procedure 1 Collecting Tweets and Hashtags

Input: A set of hashtags *HTags*, a condition $T1$ to perform adding new seed hashtags and a condition $T2$ to end this procedure.
Output: *HTags*.

Function *Bootstrapping*(*HTags*)
  while $\neg T2$ **do**
    Tweets ← *Crawling*(*HTags*) // Collect tweets including hashtags stored in *HTags*.
    if $T1$
      HTags ← *CollectNewHashtags*(*HTags*, *Tweets*) // Add hashtags used for the crawling
    end if
  end while
  return *HTags*, *Tweets*
Function *CollectNewHashtags*(*HTags*, *Tweets*)
  CandNewSeedHashtag ← *ExtractHashtags*(*Tweets*) − *HTags* // Difference set.
  NewSeedHashtag ← *Inspection*(*CandNewSeedHashtag*)
  return *HTags* ∪ *NewSeedHashtag*

Table 1 Dataset statistics

Number of history-related hashtags	147
Period of timestamps	8 Mar. 2016 – 2 Jul. 2018
Number of tweet IDs	2,370,252
Number of users	910,355
Number of URLs	663,136
Number of tweets with URLs	588,847
Number of re-tweets with URLs	415,680

tion period and $T1$ may be implemented in order to perform manual checking by an expert. *CandNewSeedHashtag* serves as a difference set and is used to store newly found hashtags (while the ones already used for tweet crawling are removed) for their subsequent manual inspection.

The tweets we collected were issued from March 8, 2016, to July 2, 2018. Bootstrapping allowed us to search for other hashtags frequently used with the seed hashtags. The tweets tagged by such hashtags were then included into the seed set after the manual inspection of all the discovered hashtags as of their relation to the history, and filtering out unrelated ones. In total, we gathered 147 history-related hashtags which allowed us to collect 2,370,252 tweet IDs pointing to 882,977 tweets and 1,487,275 re-tweets[5]. Table 1 shows the key statistics of the collected data. Table 2 shows the number of tweets we collected in each year. We gathered on average approximately 77k tweets per month in 2016 and 2018 and, on average, 89k tweets per month in 2017. The complete list of the used hashtags is shown in Table 23.

Extracting time-references. In order to conduct temporal analysis, we extracted *time-references* from tweet content. In simple approaches, two categories of temporal references are typically distinguished [11]: *explicit* and *implicit* temporal expressions. The former one is a concrete time point or time period, such as "1945" or "1980s," while the latter is a relative temporal expression such as "yesterday" or

[5] The tweet IDs are available in https://doi.org/10.5281/zenodo.3904070

Table 2 Numbers of tweets and re-tweets and the time periods of their collection

Year	Tweets	Re-tweets	Duration
2016	271,547	489,236	Mar. ∼ Dec. (10 months)
2017	421,905	644,061	Jan. ∼ Dec. (12 months)
2018	189,525	353,978	Jan. ∼ Jul. (6 months)
Total	882,977	1,487,275	2016/3 ∼ 2018/7 (28 months)

"two years ago." We use both types of temporal references in our study. In particular, we convert all implicit (relative) temporal expressions to the explicit (absolute) ones. To extract both types of time references, we use HeidelTime [54], which is an effective and popular temporal tagger with a specialized option for tweet processing. HeidelTime outputs normalized temporal expressions according to the TIMEX3 annotation standard. For example, when applied to the following tweet: "*#OTD 22 FEB of 1965, @thebeatles fly to the Bahamas to film Help! #TheBeatles #JohnLennon #PaulMcCartney* https://t.co/GkZhz6xowg," the detected temporal expression is "1965-02-22." In another example, the output is "1850" for "*#ThisDay 1850 - The self-contained gas mask is patented by Benjamin J. Lane.*." In total, there were over 890,540 tweets containing temporal expressions, which represents 38% of the dataset[6]. Table 3 shows main statistics about the time references in our dataset.

Detecting entities and their types. In this work, we employ AIDA [27]—an annotation tool which is linking phrases in short text with their corresponding Wikipedia articles, thus detecting and disambiguating entities. We apply AIDA, and we first count the number of times a given entity is mentioned in tweets. Figure 1 lists the top frequent 30 entities overall. In this set, we notice that there are 20 countries, regions, or cities, 2 historical events (WWI and WWII), 3 persons (Adolf Hitler, Abraham Lincoln, and Donald Trump), and 5 other kinds of entities. We then perform the same analysis for re-tweets and show its results in Fig. 2. In the case of the re-tweets, there are 21 countries, regions, or cities, 2 historical events (WWI and WWII), 2 persons (Adolf Hitler and Abraham Lincoln), and 5 other kinds of entities. Location entity type tends to be then most frequently mentioned within the top common entities. This is because places are key constructs distinguishing different nations and countries, helping to locate the occurrence of events, indicating loca-

[6] Note that some tweets contain abbreviated temporal expressions (e.g., 6/11/16 or 3/19/88). As HeidelTime adds for such expressions "00" at the head of year information (e.g., 0016, 0088), we converted "00" to "19" or to "20" depending on whether the last two digits are less (conversion to 20) or more (conversion to 19) than the last two digits of the timestamp's year. To validate this simple approach, we randomly sampled 100 tweets after their conversion and have found 95 of them to be correct. The following tweet is an example of an incorrect conversion: "*#OTD 1863, Lincoln designated 4/30/63 as a day of national humiliation, fasting, & prayer.*". However, as the number of tweets that had to be converted is relatively small (558 which represents only 0.6% of all tweets we analyze), such potential errors should not affect the results too much.

Table 3 Statistics related to time references of the dataset

Tweet	Number of tweets	882,977
	Number of tweets with time references	262,234
Re-tweet	Number of re-tweets	1,487,275
	Number of unique re-tweet contents	454,947
	Number of re-tweets with time references	628,306
Total	Number of tweet IDs	2,370,252
	Number of tweet IDs with time references	890,540
	Period of time references	8156 BC – 2029

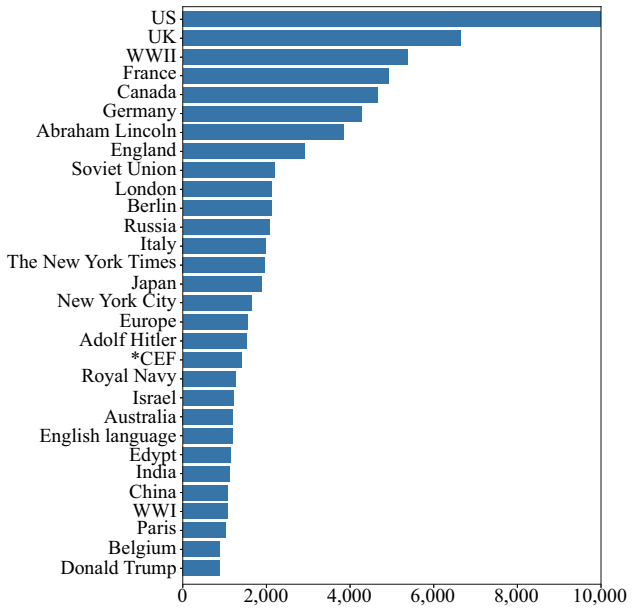

Fig. 1 Top 30 entities mentioned in tweets. "*" is used to denote abbreviations made for saving space (*CEF: Canadian Expeditionary Force)

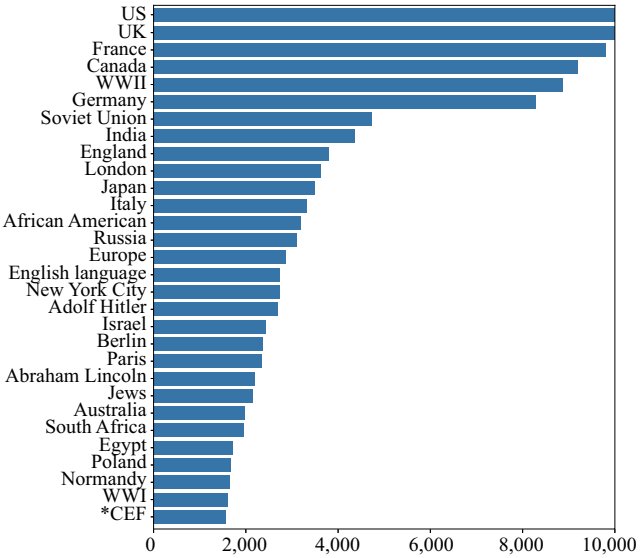

Fig. 2 Top 30 entities mentioned in re-tweets. "*" is used to denote abbreviations made for saving space (*CEF: Canadian Expeditionary Force)

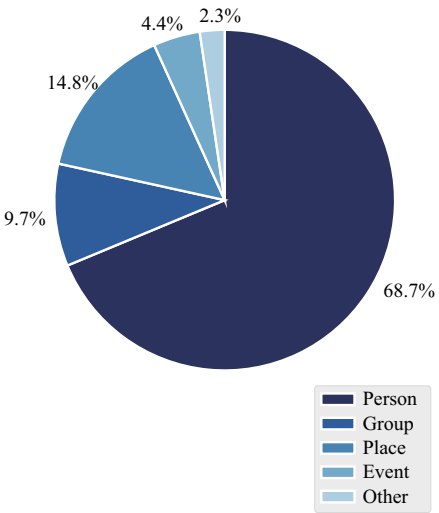

Fig. 3 Rate of different types of entities in tweets

tions of historical buildings or areas where famous people lived, as well as they often form a kind of a "bridge" between the past and the present by often "surviving" over time.

To thoroughly investigate entities and their types, we next automatically map all the entities into DBpedia [6] to obtain their type assignments. We then divide all the entities into five major types (Person, Group, Place, Event, and Others) and show their rates in Fig. 3. It can be noticed that *persons, places and groups tend to be frequently mentioned in history-focused tweets and the person category is especially common*. Note that while places were the most common entity type in the set of the top frequently mentioned entities as indicated in Figs. 1 and 2, they are actually less common than persons when all the entities in our dataset are concerned.

To give some examples of entities, Tables 4 and 5 list the top 10 entities for the Person, Group and Event types for tweets and re-tweets[7]. We can observe that the names

[7] Location examples can be seen in Tables 18, 19, 20 and 21.

Table 4 Top 10 entities of persons, groups and events in tweets

Person	Group	Event
Abraham Lincoln	Canadian Expeditionary Force	WW II
Adolf Hitler	Jews	Vietnam War
Donald Trump	US Congress	Omaha Beach
Napoleon	US Army	Battles of Saratoga
Sharon Corr	US Navy	Korean War
Barack Obama	Royal Air Force	Battle of Stalingrad
Alan Evans	Facebook	Battle of Gettysburg
George Washington	BBC	Gulf War
Bill Clinton	US House of Representatives	Battle of Verdun
Jerrard Tickell	Royal Navy	American Revol. War

Table 5 Top 10 entities of persons, groups and events in re-tweets. The abbreviated name of entity is for Battle of Merville Gun Battery (Battle of M. G. Battery)

Person	Group	Event
Adolf Hitler	Jews	WW II
Abraham Lincoln	US Congress	Vietnam War
Harriet Tubman	Central Intelligence Agency	Omaha Beach
Leonardo da Vinci	Canadian Expeditionary Force	Korean War
Yuri Gagarin	BBC	Battle of Verdun
David Bowie	US Army	Battle of Stalingrad
Charles Darwin	US House of Representatives	Battle of M. G. Battery
Donald Trump	NASA	Gulf War
George Washington	Royal Air Force	Gallipoli Campaign
Elizabeth II	US Marine Corps	English Civil War

of the 5 US presidents (Abraham Lincoln, Donald Trump, Barack Obama, George Washington and Bill Clinton) appear within the top 10 persons in tweets. Three of them (Abraham Lincoln, Donald Trump and George Washington) are also ranked as top-10 in re-tweets. As for the events, wars and battles are the prevailing type. Interestingly, groups include many military units (e.g., US Army, Royal Air force), *which suggests significant focus on wars and conflicts from the past*.

We manually analyzed tweets and re-tweets that include the mentions of the US presidents by performing random sampling of 50 tweets from both groups. Although only a few tweets and re-tweets explicitly include the name of elections (e.g., "*RT @JWilsonPenn: Forget #election2016 it's the 1916 election the #EmperorsofTime want to rig. #USHistory #TimeTravel #YA #IARTG #BYNR*"), presidential elections is one of the most popular topics; many Twitter users in our dataset mentioned the topic implicitly, for example, "*#ThisDayInHistory, 2008: Barack Obama is elected the first African American president of the United States.* https://t.co/83e2BAbVt7."

We also note that the dataset contains approximately 600 tweets mentioning term "Clinton." 93% of these tweets also contain the names "Bill" or "Hillary." We show few examples below:

- "*#TDiH: July 7, 2000, President Clinton declared the Cottage and 2.3 acres of land the President Lincoln & Soldiers' Home National Monument*"
- "*Early 1960s. A teenage Bill Clinton meets John FDI. Kennedy*"
- "*Laura Bush, Hillary Clinton push for women's history museum in DC* https://t.co/KQuGMVPmoN *#wmnhist #history*"
- "*RT @GameOnPatriots: #ThrowbackThursday Reopen the Clinton email case & investigate the conduct of Strzok, Page, Comey and others who may h...*"

4 General analysis

In this section, we investigate characteristic features of history-related tweets based on three data types: time expressions, entities and hashtags.

We first analyze which time periods appear commonly by investigating time references included in tweets. We map all the extracted temporal expressions on a timeline as shown in Fig. 4. We call the curve in Fig. 4, the *remembering curve* as it reflects the strength of the collective attention of users towards different time periods of history. To plot such a curve,

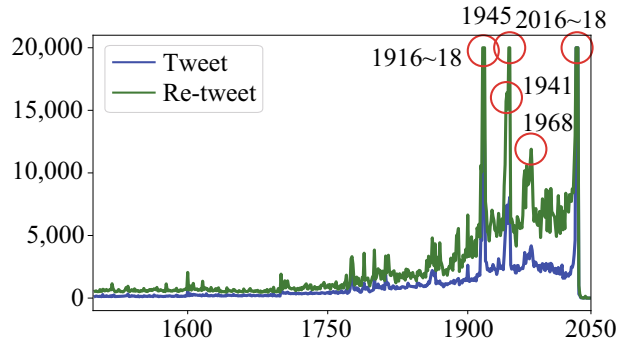

Fig. 4 Distribution of time references in all tweets of our dataset. The peaks are in years 1916, 1941, 1945 and 2016

we converted the extracted temporal references to probability distributions over their corresponding timespans using year level granularity. In other words, for a given time reference (e.g., 1960s) with t_s denoting its start year (1960) and t_e indicating its end year (1969) we set the probability distribution with zero values for $t < t_s$ and for $t > t_e$ (e.g., before 1960 and after 1969) and with nonzero values for $t_s \leq t \leq t_e$ that sum to 1 (e.g., 1/10 for each year from 1960 to 1969). We then combined for every year all the computed probability distributions based on all the tweets in our dataset. The calculated score for a year y is given in Eq. 1.

$$S(y) = \sum_{[t_s,t_e] \in T} \delta(y, [t_s, t_e]) * \frac{1}{t_e - t_s + 1} \quad (1)$$

where T includes all the extracted time references, and the function δ returns 1 if the first argument is included in the second argument; otherwise, it returns 0. The above equation assumes the uniform distribution as one that is most natural in this case.[8]

Looking at Fig. 4, we can see that the number of time references is usually rapidly increasing towards the present (neglecting short-term disturbances caused by key events that are to be discussed later). In general, *the recent past is referred to more than the distant past, and the memory decay is fastest in the recent years*. The memory decay in this context is understood as diminishing society interest and attention towards more distant past when compared to the recent past. Looking at the figure we can indeed observe that far away years in the past (e.g., ones in the nineteenth or eighteenth century) have on average less references than more recent years (e.g., ones in the twentieth century) if we do not take into account the memory peaks (highlighted in the figure). This is intuitive and correlates with the corresponding study conducted on news articles related to different countries [5].

[8] Other distributions, e.g., Gaussian distribution centered in a particular year such as a mid-year of a time period, could be alternatively used.

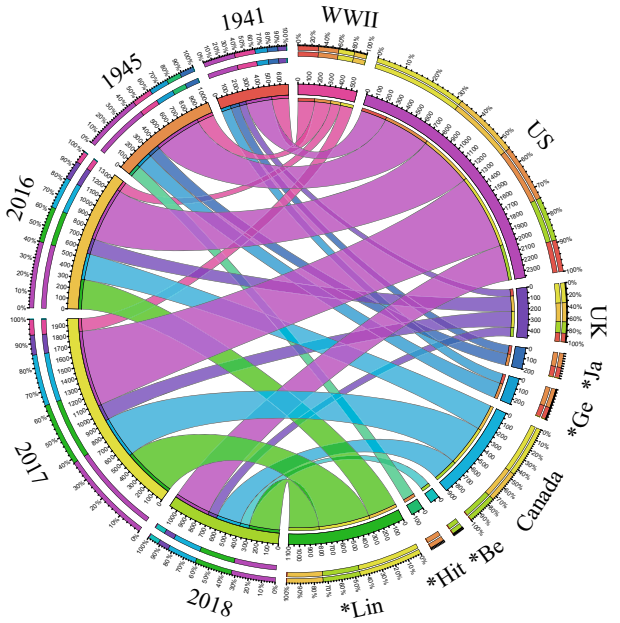

Fig. 5 Top entities associated with the four peaks of Fig. 4. "*" is used to denote abbreviations made for saving space (*Be: Berlin, *Hit: Adolf Hitler, *Lin: Abraham Lincoln, *Ge: Germany, *Ja: Japan)

Several significant peaks are visible in Fig. 4 *which represent two key events in the last century: WWI and WWII*, and years 2016∼2018. Indeed, after examining tweets posted in relation to the three peak years (1916, 1917, and 1918), we, respectively, found 193, 616, and 1,203 tweets that include the word WWI in our dataset. Two dates common for WWII are: 1941—the year of the Pearl Harbor attack and the subsequent start of the participation of USA in the war (for 1941 there are 184 tweets that include term WWII and 122 that mention Pearl Harbor attack), and 1945 which represents the end of the war (for 1945 there are 383 tweets that include the mention of WWII and 112 that mention Pearl Harbor attack). While WWII started in fact earlier with the Nazi invasion on Poland in 1939, many history-related tweets in our dataset originate from USA, UK and Canada due to the chosen English hashtags resulting in the focus on the North American involvement in the war. This can be confirmed when looking at Fig. 5 that shows the most common entities corresponding to the peaks and at Fig. 1 which lists the top entities in our dataset.

We next show in Fig. 6 the remembering curves for each year of the data collection (i.e., 2016, 2017 and 2018). It can be seen that for all the three years, WWI and WWII are commonly referred to. In addition, we can see that years falling exactly 100 years ago from the data collection year are also commonly referred to due to round anniversaries. Interestingly, for 2018, we cannot see the peak at 1941 even though it is present in 2016 and 2017. The reason is that this year is related to remembering the Pearl Harbor attack which

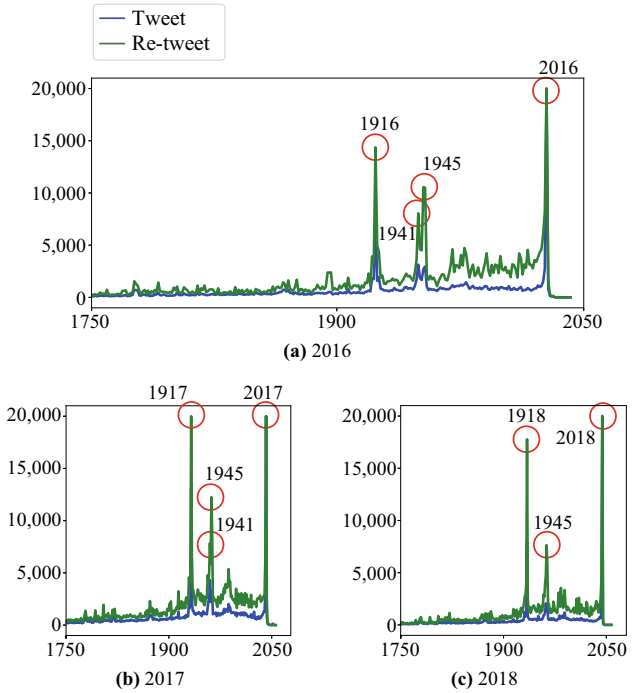

Fig. 6 Distribution of time references in tweets posted in each year

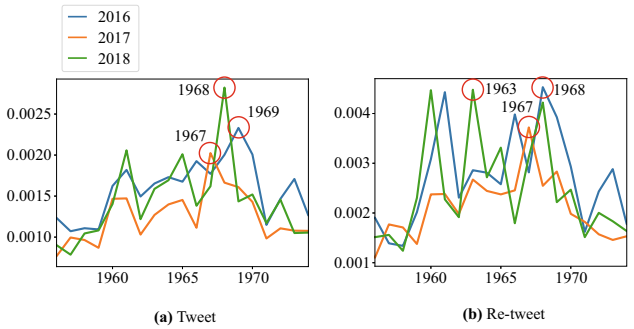

Fig. 7 Normalized distribution of time references of tweets about 1960s

occurred in December. Unfortunately, our dataset does not contain tweets posted on December in 2018.

Looking at Fig. 4 again, we can observe that there is a peak at 1968 exactly 50 years ago counting from 2018. As shown in Fig. 7 a peak on 1968 is for both tweets and re-tweets. Notably, the US presidential election was held in this year. Table 6 shows the entities and hashtags associated with 1968. As expected, the US and the candidate for the election (Robert F Kennedy) are within the top 5 entities. In addition, we can see entities that carry relation to Martin Luther King, Jr. (Memphis Tennessee and National Civil Rights Museum).

We next look into the extracted temporal expressions and the calculated remembering curves separately for each of the datasets we collected (i.e., for tweets gathered in 2016, 2017 and 2018). To perform a deeper analysis for identifying the differences among the three datasets, we

Table 6 Top 5 entities and hashtags of tweets in 1968

	Rank	Entity	Hashtag
tweet	1	US	`otd`
	2	Martin Luther King, Jr.	`thisdayinhistory`
	3	Vietnam	`vietnamwar`
	4	Robert F Kennedy	`thisday`
	5	UK	`tdih`
re-tweet	1	Martin Luther King, Jr.	`otd`
	2	Memphis Tennessee	`histoire`
	3	National Civil Rights Museum	`tdih`
	4	US	`onthisday`
	5	Vietnam	`cinema`

detected their peak years after normalizing the distributions as shown in Fig. 7. As a result, we found that the peaks for all the three datasets are around 1965 (1963, 67, 68 or 69). One common reason for these peaks is commemoration (e.g., "*referring to 50 years ago like Scars of war. Berlin, 50 years ago, in 1968. Leipzigstrasse (see the previous tweet).* `#Berlin #ColdWar` https://t.co/w327hkyzkw") including facet-focused hashtags such as the `#todayinblackhistory` ("*RT @Schomburg-Center: Wyomia Tyus becomes the first to win the gold medal in the 100m race in 2 Olympic Games (1968)* `#todayinblackhistory`"). Note that although the peak years are different for the three datasets, their main topics center around Cold War related events. For example, "*RT @SalfordUni_PCH:* `#OTD` *1967: President Lyndon B. Johnson meets with Soviet Premier Aleksei Kosygin in Glassboro, New Jersey.* `#ColdWarHisto`." Some tweets describe events of the Vietnam war ("`#OTD` *in 1969, President Nixon announces the Nixon Doctrine, laying the basis for Vietnamization in the* `#VietnamWar` https://t.co/0hFcTaSQKD") and Berlin Wall ("*RT @en_germany: "Ich bin ein* `#Berliner`" - `#OnT- hisDay` *in 1963, President John F.* `#Kennedy` *held his famous speech in front of the* `#BerlinWall`").

Figure 8 shows the most common hashtags used with content containing the peak years of Fig. 4. We can notice that 1941 and 1945 have strong connection with hashtag `#wwii` as the event was held during these respective years. Interestingly, Fig. 8 shows that there are many mentions of 2016 with `#ww1`. This is because 2016 marked the 100th anniversary of the Battle of Verdun which is especially remembered due to the exceptionally large number of casualties. In the same

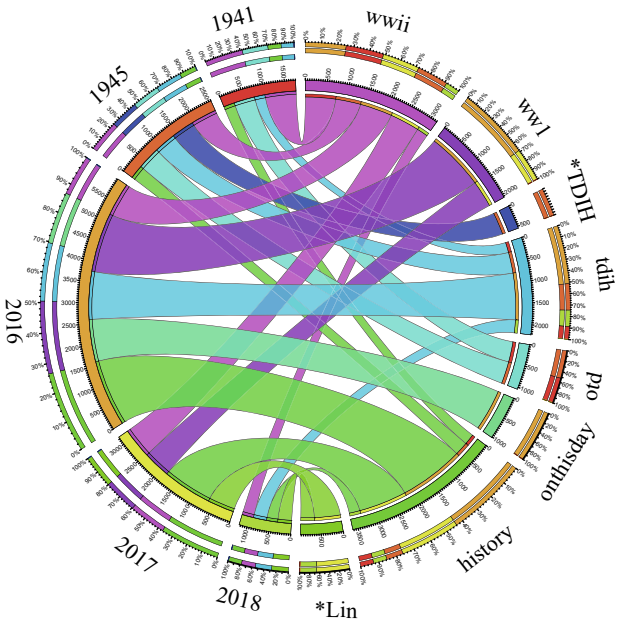

Fig. 8 Top hashtags associated with the four peaks of Fig. 4. "*" is used to indicate abbreviations made for saving space (*W1C: `#ww1centenary`, *TDIH: `#thisdayinhistory`, *Lin: `#AbrahamLincoln`)

year (1916), another remarkable event occurred—the British Army suffered its worst day with the loss of 19,240 men in the Battle of Somme (the hashtag `#somme100`). These events together with the Easter Rising of Irish republicans against British occupation in Dublin are likely causes for the strong remembrance of 1916. The round anniversary in 2016 especially amplified this commemoration; hence, the corresponding peak in Fig. 4 is even higher than the subsequent peaks associated with WWII.

Next, we notice that, as shown in Fig. 8, `#onthisday` and `#otd` are commonly used labels for indicating historical content that occurred on the same calendar day in the past. The reference to the same calendar day is in fact a popular way of recalling past events and is often used in newspapers (e.g., sections about events reported in "our newspaper on this day in the past"). `#otd` and `#onthisday` are then hashtag-based mechanisms for indicating this type of connection in Twitter.

We now analyze which hashtags and entities are commonly mentioned with years related to WWI and how different they are in each year. Tables 7 and 8 list top 5 hashtags for peak years related to WWI in tweets and re-tweets, respectively. Naturally, WWI related hashtags (`#wwi`, `#ww1`, `#wo1` and `#ww1centenary`) occupy the high ranks, and commemorative hashtags (`#otd`, `#thisdayinhistory` and so on) are also common. The `#wo1` hashtag seems to refer to WW1 in the Dutch language. We noticed that tweets including `#wo1` hashtag are often posted by a Twitter account from Belgium[9]. It seems that this account is managed by an organization located in Ypres, which is a Belgian municipality. The account uses hashtags in both English and Dutch to post WW1-related tweets, especially ones about many WW1 events related to Ypres, such as *"Today 1916 1CanArtBde, `#Zillebeke #SanctuaryWood`: Point 76 is seriously damaged by fire from our Howitzers `#ww1 #ypres #wo1 #ieper`"* (note that Zillebeke is a village that is now part of Ypres). Also it tweets on WW1-related events which occurred outside of Belgium, e.g., *"Today 1916 #French occupy #Florina, E of #Monastir (#Greece) #ww1 #ypres #wo1 #ieperToday"*. As expected, hashtags used for referring to events that happened 100 years ago are listed in the tables, too. For example, `#somme100` in re-tweets in 2016 and `#100-yearsago` in both tweets and re-tweets in 2018 are ranked within the top 5 hashtags.

Tweets using the `#silentfilm` hashtag touch on variety of topics including reviews, movie scenes, synopses or actors, for example:

- *"Near #Hereford today? Our author @cheshellen giving intro to `#silentfilm` Shoes (1916) @borderlines at 3pm on 3 March* https://t.co/CNbykY7E0q"
- *"Our older reviews now back on line. See what we thought of the 2016 `#silentfilm` movie scenes, synopses or actorsyear here* https://t.co/c3I8t3DWLy. https://t.co/85q54bXT7w"
- *"Ukrainian `#silentfilm` rarity Shkurnik (1929) screens tomorrow (30 May) at LSE London, Details at...* https://t.co/vZ7E1yl7eH."

Tables 9 and 10 list the top 5 entities for peak years related to WWI in tweets and re-tweets, respectively. As for WWI related entities, we can observe 5 of them in tweets: *WWI*, *Battle of Verdun* (its duration is from Feb. to Dec. 1916), *Mata Hari* who was a dancer and a spy for Germany during WW1, *Ralph Hamilton* who was a soldier taking part in epic battles at Loos, Ypres and Somme and who had a master degree from Belhaven, and *Woodrow Wilson* who issued his principles for the end to WWI. In contrast, there are 3 WWI related entities in re-tweets: *WWI*, *US Marine Corps*, and *Needham Roberts* who served in the US Army during WWI. Interestingly, places are the most popular entity type yet we can also see several persons in the tables that were born, died at these years, and events that occurred at that time. As shown in Fig. 5, peak years during WWII are often mentioned with place names.

Finally, we compute the correlation between the three datasets considering the same months and the same months in the same years. We use Jaccard coefficient and Mutual Information (MI) as follows:

[9] https://mobile.twitter.com/Frontline_Tours

Table 7 Top 5 hashtags of tweets

Year	Rank	1916	1917	1918
2016	1	ww1	ww1	ww1
	2	otd	wwi	wwi
	3	ww1centenary	otd	otd
	4	ypres	thisdayinhistory	thisdayinhistory
	5	wo1	fww	history
2017	1	ww1	ww1	ww1
	2	wwi	otd	wwi
	3	thisdayinhistory	wwi	otd
	4	silentfilm	fww	thisday
	5	otd	militaryhistory	fww
2018	1	thisdayinhistory	ww1	100yearsago
	2	otd	otd	ww1
	3	ww1	thisdayinhistory	otd
	4	wwi	wwi	thisdayinhistory
	5	thisday	botd	lestweforget

Table 8 Top 5 hashtags of re-tweets

Year	Rank	1916	1917	1918
2016	1	ww1	otd	ww1
	2	ww1centenary	ww1	otd
	3	otd	wwi	wwi
	4	onthisday	onthisday	onthisday
	5	somme100	fww	rememberthem
2017	1	ww1	ww1	ww1
	2	botd	fww	otd
	3	otd	onthisday	wwi
	4	wwi	ww1centenary	fww
	5	silentfilm	otd	histmed
2018	1	otd	ww1	100yearsago
	2	ww1	wwi	ww1
	3	onthisday	todayinblackhistory	otd
	4	todayinblackhistory	otd	wwi
	5	fww	germany	onthisday

$$Jaccard(A, B) = \frac{| T_A \cap T_B |}{| T_A \cup T_B |} \tag{2}$$

$$MI(A, B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \left(\frac{p(a, b)}{p(a)p(b)} \right) \tag{3}$$

where: $| \cdot |$ is the size of a set. T_A and T_B are the entity/hashtag sets included in tweets that are posted in Year A and B, respectively. The higher the score of the measurements, the more correlated they are. Note that we compute the scores for four months from March to June only as we have tweets posted during these months in all the three years.

Table 11 shows the correlation results for the same calendar months between different years. Looking at the values for hashtags, we see that all of the scores are over 0.05, especially the correlation score between 2017 and 2018 is over 0.1. This tendency can be seen for the entities too; all the scores are over 0.05 and the correlation scores between 2017 and 2018 are the highest in tweets and re-tweets collections. We then show the scores of MI between different years in Table 12. All the scores are approximately $0.4 \sim 0.6$, and the standard derivations are around 0.1 or 0.2. We observe the same tendency for entities; however, the scores of entities tend to be lower than the ones of hashtags. These results may indicate that Twitter users (at least in our dataset) tend to recall different past events over different years even when using the same hashtags, e.g., memorial days of WWI or WWII as it is shown that several commemorative hashtags are commonly used through all the years (Fig. 8) and based on the existence of common peak years in Fig. 6.

Table 9 Top 5 entities of tweets

Year	Rank	1916	1917	1918
2016	1	US	US	Ralph Hamilton
	2	France	Germany	Belhaven University
	3	Melbourne	Mata Hari	WWI
	4	Sydney	France	US
	5	UK	WWII	France
2017	1	US	The New York Times	US
	2	WWII	US	Woodrow Wilson
	3	UK	France	France
	4	France	WWII	WWII
	5	Germany	Belgium	WWI
2018	1	US	US	The New York Times
	2	UK	WWII	US
	3	France	The New York Times	WWII
	4	Battle of Verdun	France	France
	5	WWII	Germany	UK

Table 10 Top 5 entities of re-tweets

Year	Rank	1916	1917	1918
2016	1	US	US	Germany
	2	UK	US Marine Corps	France
	3	France	WWII	WWI
	4	WWII	Illinois	Needham Roberts
	5	Melbourne	East Saint Louis, Illinois	First Army (Bulgaria)
2017	1	Soviet Union	US	House of Romanov
	2	Lyudmila Pavlichenko	France	France
	3	UK	The New York Times	Ben Johnson (actor)
	4	WWII	WWII	WWII
	5	US	Germany	Canada
2018	1	US	Gwendolyn Brooks	The New York Times
	2	Austria–Hungary	Pulitzer Prize	France
	3	New York City	African American	US
	4	Emma Goldman	US	WWII
	5	Mabel Fairbanks	English language	UK

Table 11 Scores of Jaccard coefficients for the same calendar months between different years of data collection. ** indicates that p values of t-test comparing of tweets and re-tweets are less than 0.05 and 0.01, respectively

		Hashtags		Entities	
		Average	Standard Deviation	Average	Standard Deviation
tweet	2016–2017	0.064	0.037	0.077	0.045
	2016–2018	0.059	0.032	0.072	0.043
	2017–2018	0.101	0.029	0.116	0.050
re-tweet	2016–2017	0.061	0.038	0.054**	0.034
	2016–2018	0.057	0.035	0.057**	0.036
	2017–2018	0.104	0.032	0.072**	0.031

Table 12 Correlation measure using Mutual Information for the same calendar months between different data collection years

		Hashtags		Entities	
		Average	Standard Deviation	Average	Standard Deviation
tweet	2016–2017	0.437	0.173	0.441	0.184
	2016–2018	0.469	0.173	0.425	0.185
	2017–2018	0.546	0.143	0.463	0.125
re-tweet	2016–2017	0.496*	0.209	0.447	0.200
	2016–2018	0.486	0.215	0.438	0.222
	2017–2018	0.596*	0.160	0.542**	0.161

* and ** indicate that p values of t-test comparing of tweets and re-tweets are less than 0.05 and 0.01, respectively

Table 13 Scores of Jaccard coefficient for the same years between different months

		Hashtag		Entity	
		Average	Standard Deviation	Average	Standard Deviation
tweet	2016	0.046	0.032	0.034	0.025
	2017	0.115	0.028	0.072	0.022
	2018	0.116	0.031	0.066	0.025
re-tweet	2016	0.044	0.033	0.036	0.029
	2017	0.111*	0.028	0.056**	0.018
	2018	0.111*	0.033	0.062*	0.023

* and ** indicate that p values of t-test comparing of tweets and re-tweets are less than 0.05 and 0.01, respectively

Table 14 Correlation for the same years between different months

		Hashtag		Entity	
		Average	Standard Deviation	Average	Standard Deviation
tweet	2016	0.405	0.185	0.399	0.227
	2017	0.563	0.123	0.546	0.122
	2018	0.549	0.150	0.536	0.155
re-tweet	2016	0.417	0.227	0.389	0.247
	2017	0.613**	0.138	0.563*	0.153
	2018	0.589*	0.162	0.548	0.161

* and ** indicate that p values of t-test comparing of tweets and re-tweets are less than 0.05 and 0.01, respectively

Finally, Tables 13 and 14 show the results of the two correlation measures applied for different months in the same years. Overall, the two tables show the same tendency as Tables 11 and 12. Looking at the results of the Jaccard coefficient (Table 13), all the scores of hashtags are over 0.04 in tweets, whereas the scores of entities are approximately 0.05. The scores by MI for re-tweets in Table 14 are also similar to ones in Table 12.

4.1 Connection of past and present entities

We now look into entities referred to in tweets as, often, a particular entity such as a person or an event is what society remembers strongly from a particular period in the past. The basic question that we approach in this section is: Which past and present entities are compared or mentioned together?

The issue of connecting present and past entities is especially interesting as it relates to the notion of "usable history." Historical entities can be used for a variety of reasons, for example, for comparison with present entities or present context, for emphasizing analogy, making predictions and so on. To analyze the way in which past entities are utilized in connection to the present ones we first need to separate present and past entities. We apply a simple rule, such that an entity is regarded as a *past entity* if the end of its lifetime[10] (e.g., person's life, event duration, organization duration) falls within

[10] For currently valid entities such as alive persons the end of their lifetimes is set to the current year.

Table 15 Sizes of the past and present entity sets in our whole dataset

	Total	Person	Group	Place	Event	Other
Size of the past entity set	13,669	9,346	1,306	832	691	1,494
Size of the present entity set	20,542	14,178	4,901	339	787	337

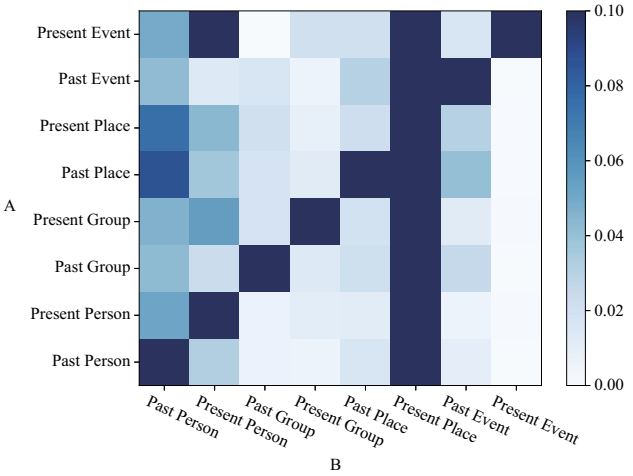

Fig. 9 Conditional probabilities $P(B|A)$ of entity type on x axis (in column) given the presence of entity type on y axis (in row)

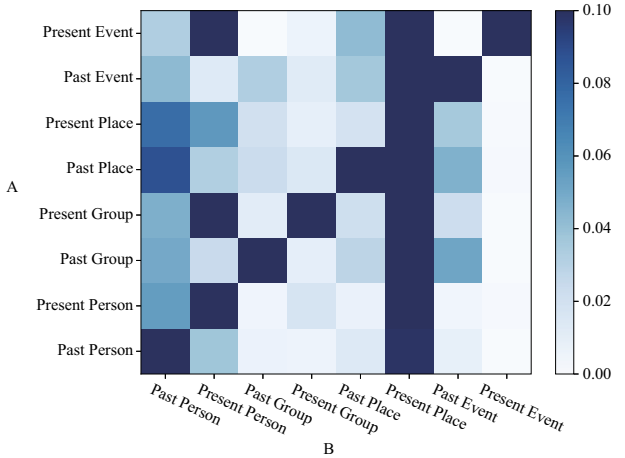

Fig. 10 Conditional probabilities P(B|A) of entity type on x axis (in column) given the presence of entity type on y axis (in row) in retweets

the last millennium. Entity lifetime data were collected from DBpedia[11].

First, in Table 15, we contrast the sizes of the past and present entity sets extracted from our dataset. We can observe that the number of unique past entities extracted by AIDA and typed by DBpedia is relatively large constituting roughly half of that of the present entities. This confirms that our dataset is specifically focused on history.

We then plot in Figs. 9 and 10 conditional probabilities based on the entity types that we study to analyze how often the different types appear given another entity of a particular type in tweets and re-tweets, respectively. We can notice that *present places*[12] *tend to co-occur with entities of any other type, and both the present and past persons also tend to co-occur with entities of any other type excluding past events*. Person probabilities, $P(Present\ Person|Past\ Event)$ are low in both tweets and re-tweets, as it is rare to mention present persons in the case where one recalls past wars—popular past events in our database as shown in Tables 16 and 17. In re-tweets, we can find that many probabilities of past entities are relatively high. In particular, the scores of conditional probabilities for past persons, places and events are high. This phenomena can be seen in tweets as well; albeit, the values of the conditional probabilities in re-tweets tend to be bit higher than ones in tweets. Finally, if a past place, such as Nazi Germany or Holy Roman Empire, is in a tweet, past events tend to occur as well. This is also intuitive. Interestingly, past places are often mentioned with present places as well (supposedly for emphasizing place continuity, spatial relations or for place-oriented comparisons). In addition, in re-tweets, past places tend to co-occur slightly more with past groups and present events than it is in the case of tweets. We manually checked the reason why present events can be mentioned with past places, and found that there are several annual events, e.g., motorsports held in East Germany, that have relatively long histories.

Next, in Tables 16 (tweets) and 17 (re-tweets) we take the top 5 common present persons (column "Entity" in the first part of the table), the top 5 common past persons (column "Entity" in the second part of the table), the top 5 common present events (column "Entity" in the third part of the table), and the top 5 common past events (column "Entity" in the last part of the table). We then output in columns "1," "2" and "3" their top 3 most often co-occurring entities from the opposite time frame. In particular, if the "Entity" column contains past entities, then in columns "1," "2" and "3" we show their top co-occurring present entities. Otherwise, we show past entities. When looking at past persons and past events, one can observe that indeed present places commonly co-occur with them. For example, for the top 5 past persons, their most common co-occurring entities contain 8 countries and 1 city, while for the past events the corresponding number

[11] We use "birthDate" and "deathDate" for person entities, "formationDate" and "dissolutionDate" for groups, and "foundingDate" and "dissolutionDate" for locations.

[12] Present places are those that do not have any end date or the end date is after 2000. Other places are considered as past places such as Nazi Germany or the Holy Roman Empire.

Table 16 Top 5 present and past persons and top 5 present and past events (given in "Entity" column) with their top 3 co-occurring past/present entities of any type in tweets

	Rank	Entity	1	2	3
Pres. person	1	Donald Trump	Adolf Hitler	Richard Nixon	Russia
	2	Ronald Reagan	Soviet Union	Russia	Richard Nixon
	3	Elizabeth II	Winston Churchill	James II of England	George VI
	4	Barack Obama	Russia	Adolf Hitler	M. Ali of Egypt
	5	Bill Clinton	NAFTA	Grover Cleveland	Richard Nixon
Past person	1	Abraham Lincoln	US	Donald Trump	Nationaal Archief
	2	Adolf Hitler	Germany	Donald Trump	US
	3	Napoleon	Italy	Russia	Egypt
	4	Richard Nixon	US	China	Gerald Ford
	5	George Washington	US	Philadelphia	Boston
Pres. event	1	Grand Ole Opry	Hank Williams	Roy Acuff	–
	2	Ind. M. Speedway	WWII	–	–
	3	The Simpsons	The T. U. Show	–	–
	4	American Idol	–	–	–
	5	War on Terror	WWII	–	–
Past event	1	WWII	US	UK	Japan
	2	Vietnam War	US	Ken Burns	CBS
	3	Battles of Saratoga	US	–	–
	4	Korean War	US	North Korea	South Korea
	5	Battle of Stalingrad	Germany	Ukraine	Italy

"-" denotes cases when no corresponding entity can be found. The abbreviated names of entities are for: Muhammad Ali of Egypt (M.Ali of Egypt), North American Free Trade Agreement (NAFTA), Indianapolis Motor Speedway (Ind. M. Speedway) and The Tracey Ullman Show (The T. U. Show)

of locations is 12. Apparently, when recalling past persons and past events users tend to also mention where the persons lived or where the events occurred, thus "grounding" them in spatial dimension. Naturally, it is rare to mention present events and past entities as shown in Figs. 9 and 10.

Finally, Tables 18 (tweets), 19 (tweets), 20 (re-tweets) and 21 (re-tweets) list present and past entities co-occurring with the top frequent locations, from where again the significance of WWII in relation to collective memories of many countries can be observed.

4.2 Hashtag analysis

In this section, we investigate the popularity patterns of hashtags. Hashtags are commonly used to indicate specific themes of tweets allowing others to find related tweets. First, Fig. 11 shows the top frequently used hashtags based on their tweet counts. We also list the top hashtags ranked by the re-tweet count in Fig. 12 and by the number of Twitter accounts from which the tweets originate in Fig. 13. From these data, we can observe that `#throwbackthursday`, which is representative for a trend among social media sites including Twitter and Facebook to post own past photographs (often from one's childhood), gains most attraction across all these dimensions of popularity. The tweets with hashtag `#throwbackthursday` and similar ones predominantly refer to personal experiences and these hashtags are used by large numbers of users. Another observation is that Fig. 11 *shows that* `#onthisday` *and* `#otd` *are present in a relatively large number of tweets and re-tweets, yet, they are used by fewer accounts.* Unlike `#throwbackthursday`, these hashtags tend to be used by specialists (often historians and scientists from related areas) who select and disseminate interesting content about the past. This content then triggers relatively *high engagement from other users as evidenced by the high re-tweet popularity of* `#onthisday`, `#otd` *in Fig.* 12. We note that the demographics analysis should provide in the future many more interesting insights regarding typical profiles of posting users, and is going to form a part of our future work.

4.3 URL analysis

The past is often recalled by the reference to diverse multimedia content or artifacts such as images, videos, objects or historical documents. Actually, many online services offer such kinds of data. We expect the Twitter users in our dataset to sometimes back their references to the past with online arti-

Table 17 Top 5 present and past persons and top 5 present and past events (given in "Entity" column) with their top 3 co-occurring past/present entities of any type in re-tweets

	Rank	Entity	1	2	3
Pres. person	1	Ronald Reagan	Soviet Union	Marilyn Monroe	Russia
	2	Paul Mccartney	Linda Mccartney	John Lennon	Dahomey
	3	Donald Trump	Richard Nixon	Adolf Hitler	Andrew Johnson
	4	Elizabeth II	Cecil Beaton	Winston Churchill	George VI
	5	David Bowie	Elvis Presley	The Manhattan Transfer	Simon Vouet
Past person	1	Adolf Hitler	Germany	Elie Wiesel	US
	2	Abraham Lincoln	US	Nationaal Archief	Lewis Lehrman
	3	Napoleon	Russia	Egypt	Italy
	4	George Washington	US	Philadelphia	Saudi Arabia
	5	Marilyn Monroe	Jerry Lewis	Arthur Miller	Lauren Bacall
Pres. event	1	Ind. M. Speedway	–	–	–
	2	Sesame Street	–	–	–
	3	2006 Lebanon War	–	–	–
	4	Bagpuss	–	–	–
	5	Operation Enduring Freedom	-	–	–
Past event	1	WWII	US	UK	Canada
	2	Korean War	US	North Korea	Canada
	3	Vietnam War	US	Ken Burns	UK
	4	Omaha Beach	US	Germany	NATO
	5	Battle of Stalingrad	Germany	Italy	China

"-" denotes cases when no corresponding entity can be found. The abbreviated name of entity is for Indianapolis Motor Speedway (Ind. M. Speedway)

Table 18 Top 10 present locations and their top 3 co-occurring past entities of any type in tweets

Rank	Present locations	1	2	3
1	US	WWII	Soviet Union	Battles of Saratoga
2	UK	WWII	James II of England	Russia
3	Canada	Canadian Exped. Force	Seaforth Highlanders	42nd Regiment of Foot
4	Germany	Adolf Hitler	Soviet Union	WWII
5	Russia	WWII	Catherine the Great	Nicholas II of Russia
6	Italy	Benito Mussolini	WWII	Battle of Monte Cassino
7	Japan	WWII	H. F. Mears	Soviet Union
8	Israel	WWII	Nazi Germany	Adolf Hitler
9	Australia	La Trobe University	WWII	Amy Johnson
10	Egypt	Anwar Sadat	Gamal Abdel Nasser	WWII

The abbreviated names of entities are for: Canadian Expeditionary Force (Canadian Exped. Force) and Helen Farnsworth Mears (H. F. Mears)

facts, although such enhancement might not be very common according to the intuition that finding textual description of a historical event or an entity is likely easier than finding relevant multimedia. As shown in Table 1, approximately half of our tweets include links to some kind of web services. We then analyze the types of external data that users refer to when they send history-related tweets. Figures 14 and 15 list the top 20 websites mentioned in tweets and re-tweets, respectively, in our dataset[13]. Note that URLs linked in tweets are automatically replaced with shortened URLs (t.co links). For obtaining the original URLs, we used the twitter-text-python library[14].

As it can be observed, *users often enrich tweets with images (Instagram) or videos (YouTube and vine)*. After

[13] Note that we have removed http://twitter.com focusing on external websites.

[14] https://github.com/edburnett/twitter-text-python

Table 19 Top 10 past places and their top 3 co-occurring present entities of any type in tweets

Rank	Past places	1	2	3
1	Soviet Union	US	Germany	Mikhail Gorbachev
2	Nazi Germany	Poland	US	Israel
3	Ottoman Empire	UK	Greece	US
4	West Germany	US	NATO	Germany
5	South Vietnam	NATO	Laos	173rd ABC Team
6	Russian Empire	UK	Ukraine	Afghanistan
7	Czechoslovakia	US	Poland	US
8	Empire of Japan	US	Australia	Philippines
9	Roman Empire	Marcus Aurelius	Titus	Egypt
10	North Vietnam	US	D.R. Congo	Henry Kissinger

The abbreviated names of entities are: for 173rd Airborne Brigade Combat Team (173rd ABC Team) and Democratic Republic of The Congo (D.R. Congo)

Table 20 Top 10 present locations and their top 3 co-occurring past entities of any type in re-tweets

Rank	Present locations	1	2	3
1	US	WWII	Soviet Union	Vietnam War
2	UK	WWII	Linda Mccartney	James II of England
3	Canada	Canadian Exped. Force	WWII	Battle of Vimy Ridge
4	Germany	Adolf Hitler	Soviet Union	WWII
5	India	Royal Flying Corps	Indira Gandhi	Mahatma Gandhi
6	Japan	WWII	Battle of Okinawa	Bombing of Darwin
7	Italy	Benito Mussolini	G. B. Donati	Battle of Ortona
8	Russia	WWII	Catherine the Great	Napoleon
9	Israel	Adolf Eichmann	Chiune Sugihara	Albert Einstein
10	Australia	Amy Johnson	Battle of Kapyong	WWII

The abbreviated names of entities are for: Canadian Expeditionary Force (Canadian Exped. Force) and Giovanni Battista Donati (G. B. Donati)

Table 21 Top 10 past places and their top 3 co-occurring present entities of any type in re-tweets

Rank	Past places	1	2	3
1	Soviet Union	US	Germany	Poland
2	Nazi Germany	Israel	US	Poland
3	Ottoman Empire	Syria	UK	Greece
4	Russian Empire	UK	US	San Francisco
5	Empire of Japan	US	Singapore	Wake Island
6	West Germany	US	Germany	East Germany
7	Roman Empire	Marcus Aurelius	Titus	Israel
8	Lithuania	Romania	Slovenia	NATO
9	Francia	Guillermo Francella	Portugal	-
10	Czechoslovakia	Warsaw Pact	Nicholas Winton	UK

"-" denotes cases when no corresponding entity can be found

manual inspection we found that roughly two kinds of images are linked from tweets: personal past and general historical images. The former one tends to be used with `#throwbackthursday` or `#tbt` hashtags. The latter one means any past pictures and occurs with diverse hashtags.

In addition, users tend to sometimes link to shopping sites (e.g., Amazon, eBay) using past-related hashtags[15]. Closer investigation of such tweets revealed that the posted links refer to history related items such as movies about historical events or entities.

[15] E.g., https://twitter.com/MilitaryBios/status/832723671269642240

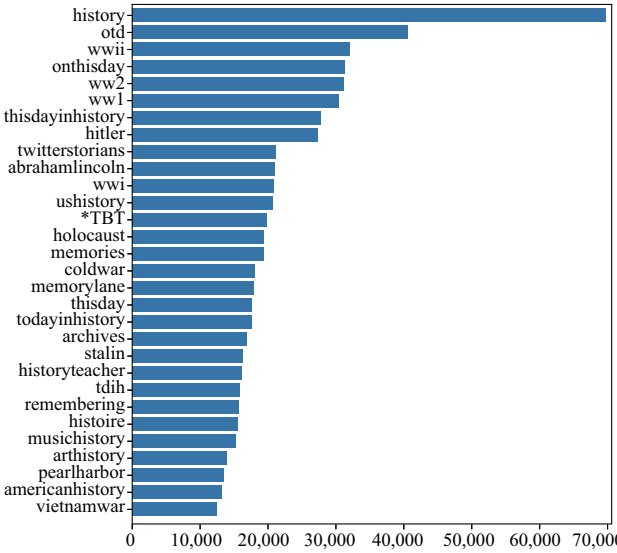

Fig. 11 Top 30 popular hashtags by counts of tweets. "*" is used to indicate abbreviations made for saving space (*TBT: throwbackthursday)

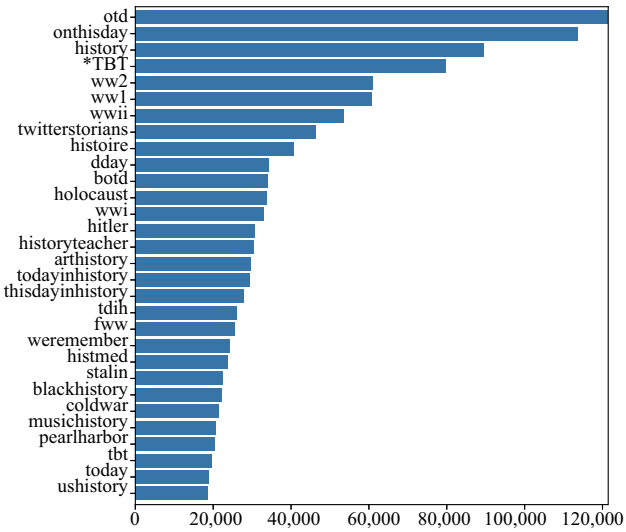

Fig. 12 Top 30 popular hashtags by counts of re-tweets. "*" is used to indicate abbreviations made for saving space (*TBT: throwbackthursday)

We finally list in Table 22 the main types of websites linked from tweets. To compute these statistics, we have randomly sampled 32k tweets and 36k re-tweets that contain URLs and we classified the linked websites into different types using website-category-api[16]. The API assigns to each URL multiple labels from 404 predefined categories based on IAB Tech Lab Content Taxonomy[17]. The table shows the top 15 categories and their ratios. Note that the same tendencies were observed in re-tweets and hence their results are not shown.

[16] https://www.webshrinker.com/website-category-api/

[17] https://www.iab.com/guidelines/iab-tech-lab-content-taxonomy/

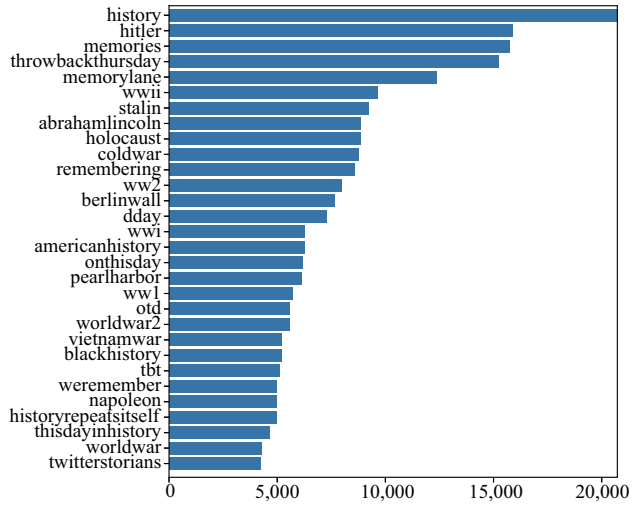

Fig. 13 Top 30 popular hashtags by counts of users

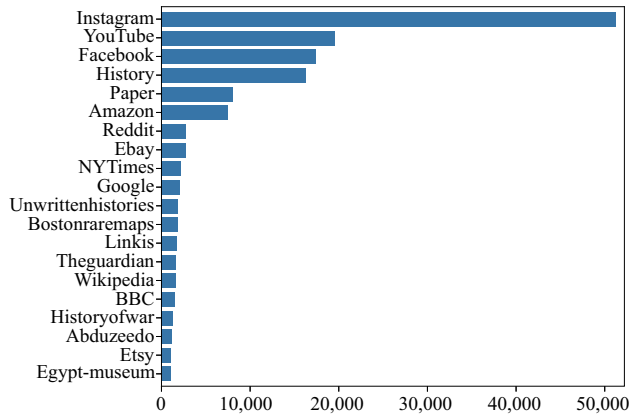

Fig. 14 Top 20 websites referred in tweets

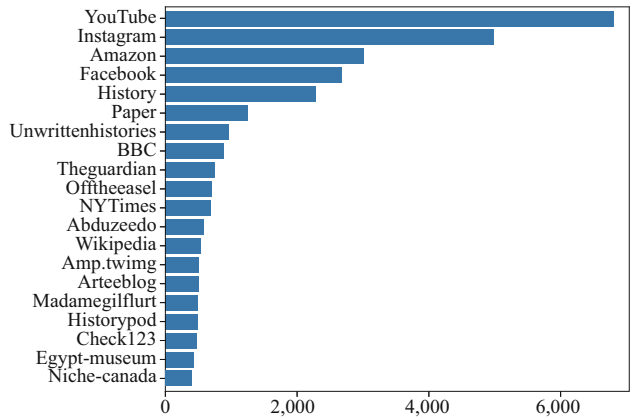

Fig. 15 Top 20 websites referred in re-tweets

Table 22 Top 15 website categories referenced

Rank	Category	Percentage (%)
1	News/Weather/Information	23.4
2	Technology & Computing	11.5
3	Arts & Entertainment	9.0
4	Education	8.7
5	Hobbies & Interests	7.9
6	Uncategorized	6.5
7	File Sharing	4.9
7	Travel	4.9
9	Law, Government, & Politics	4.8
10	Streaming Media	3.9
10	Society	3.9
12	Sports	2.9
12	Social Networking	2.9
14	Books & Literature	2.5
15	Health & Fitness	2.3

The top 5 categories are: *News/Weather/Information* (e.g., newsweek.com, nytimes.com), *Technology & Computing* (e.g., giphy.com, snapshotsofthepast.com), *Arts & Entertainment* (e.g., youtube.com, instagram.com, vine.co), *Education* (e.g., britishmuseum.org, en.wikipedia.org, litencyc.com, history.com) and *Hobbies & Interests* (e.g., collections.mcny.org, classicwarbirds.co.uk, coinworld.com).

We can also see from the table that the News type is the most common category for both tweets and re-tweets. In the top-5, the Education type is included, too. This provides some support to the conclusion on the educational role of the history-related content on Twitter. Interestingly, history and memories are also sometimes connected to commercial and arts or to entertainment activities as in *Sports* (e.g., fifa.com, motorsportmagazine.com, sports.abs-cbn.com), *Society* (e.g., soundcloud.com, www.europeana.eu, hmd.org.uk), and *Books & Literature* (e.g., thehistorypress.co.uk, amazon.com, askdavid.com) categories.

5 Category-based analysis

5.1 Definitions

In this section, we describe our categorization scheme of hashtags related to history. The objective is to determine key types of history references. Based on the proposed categories, automatic classifiers could be built to allocate tweets into different classes. Automatically labeling tweets could be then used for improving content retrieval, recommendation or for further analysis that would lead to better understanding of history-related interest and content sharing.

Based on manual investigation of a large sample of tweets, we distinguish the following categories:

1. **General History** hashtags used in general to broadly identify history-related tweets that do not fall into any specific type (e.g., `#history`, `#histo ryfacts`).
2. **National or Regional History** hashtags which relate to national or regional histories, for example, `#ushistory` or `#can adianhistory` inclu -ding also past names of locations (e.g., `#anci entgreece`).
3. **Facet-focused History** hashtags which relate to particular thematic facets of history (e.g., `#sporthistory`, `#arthistory`).
4. **General Commemoration** hashtags that serve for commemorating or recalling a certain day or period (often somehow related to the day of tweet posting), or unspecified entities, such as `#todaywe remember`, `#otd`, `#onthisday`, `#4ye arsago` and `#remem berthem`.
5. **Historical Events** hashtags related to particular events in the past (e.g., `#wwi, #seven yearswar`).
6. **Historical Entities** hashtags denoting references to specific entities such as persons, organizations or objects (e.g., `#stalin`, `#napoleon`).

Table 23 shows our manual assignment of all the history-related hashtags found within the dataset into the above-described categories. Note that hashtags concerning particular dates or time periods such as `#june61944` could be considered as a separate category or could be made a part of **General Commemoration**. We decided, however, to place them under the **Historical Events** category as they tend to be used to refer to particular events by their dates (e.g., `#june61944` referring to the Normandy landings).

We next show in Figs. 16 and 17 the rate of each category based on the counts of tweets and re-tweets. In tweets, **General History** appears to be quite common followed by **General Commemoration** and **Historical Events**. On the other hand, in re-tweets, **General Commemoration** is the most common category followed by **General History** and **Historical Events**. These results suggest *commemorating past events appears to be a common activity.* It seems that users tend to be quite interested in historical events, especially, in relation to their anniversaries.

5.2 Inter-category similarity

To better understand characteristics of the proposed categories, we now investigate inter-category affinity by measuring the co-occurrence between the hashtag categories. For

Table 23 Collected hashtags and their categories

Category	Hashtags
General History	history, historyfacts, oldpicture, historyteacher, memorylane, histoire, twitterstorians, historicalcontext, colorization, memories, oldphoto, earlymodern, historicalevent, worldhistory, twitterstorian, historynerd, histedchat, historyfeed, archives, historymatters
National or Regional History	canadianhistory, ushistory, histoireducanada, jewishhistory, nazigermany, ottoman, cdnhistory, dchistory, cdnhist, thirdreich, tohistory, mdhistory, bchist, abhistory, vthistory, britishhistory, ancientchina, ancientegypt, ancientgreece, americanhistory, thisiscanadashistory, ottomanempire, ontariohistory, earlyamhistory, japanhistory, japanesehistory, chinesehistory, localhistory
Facet-focused History	blackfacts, histoiremiliterre, wmnshist, arthistory, sporthistory, womenshistory, navalhistory, presidentialhistory, musichistory, militaryhistory, blackhistory, envhist, histmed, wmnhist, todayintennishistory, todayinblackhistory, ibhistory, u2history, historythroughcoins, histSTM, silentfilm, historyscience, histsci, digitalhistory, foodhistory, histmonast, histnursing, histgender, histtech
General Commemoration	onthisday, otd, otdh, thisdayin, thisdayinhistory, todayinHistory, tdih, onthisdayinhistory, otdih, 100yearsago, thisday, lessthan100yearsago, todayweremember, titanicremembranceday, weremember, 100yearsago, remembering, wewillrememberthem, rememberthem, remembranceday, historyrepeatsitself, throwbackthursday, tbt
Historical Events	1ww, gulfwar, ColdWar, ww2, ww1, worldwar, worldwarii, vietnamwar, worldwar2, worldwartwo, veday, worldwarone, greatwar, battleofmidway, holocaust, frenchrevolutionarywar, wwii, wwi, sevenyearswar, firstworldwar, coldwarhist, gulfwar, battleofokinawa, dday, berlinwall, ddayoverlord, operationoverlord, fww, pearlharbor, americanrevolution, 6juin44, sww, june61944, victoryineuropeday, dday72, neverforget84, warof1812, ww1politics, ww1centenary,ww1economy, cw150
Historical Entities	stalin, hitler, abrahamlincoln, rudolfhess, napoleon

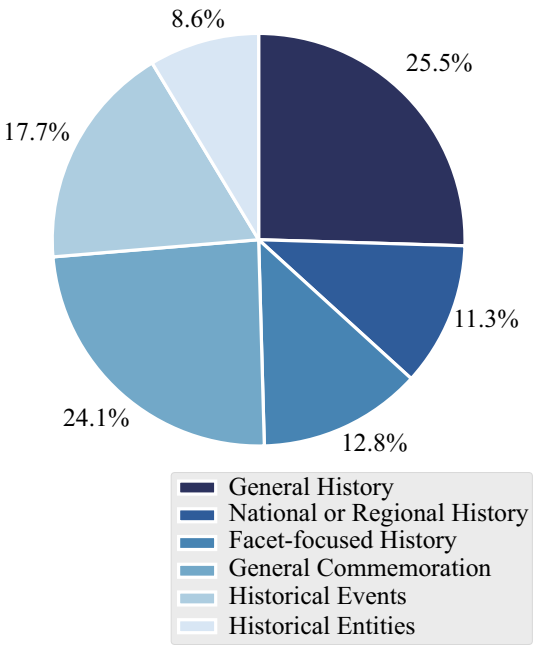

Fig. 16 Distribution of hashtag categories in tweets

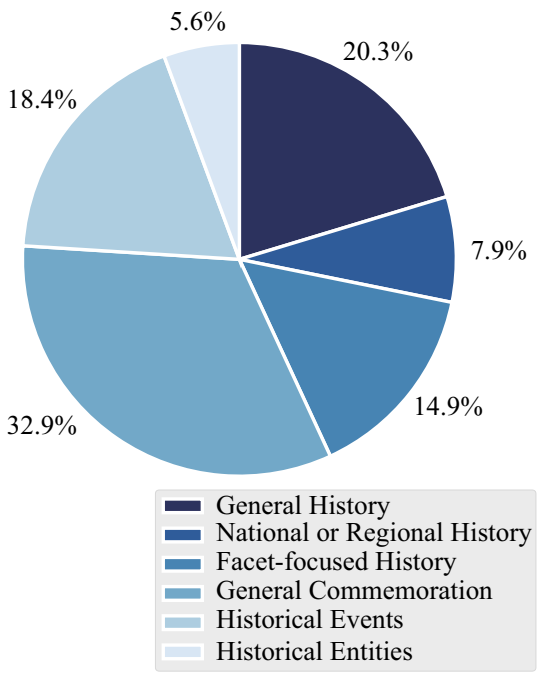

Fig. 17 Distribution of hashtag categories in re-tweets

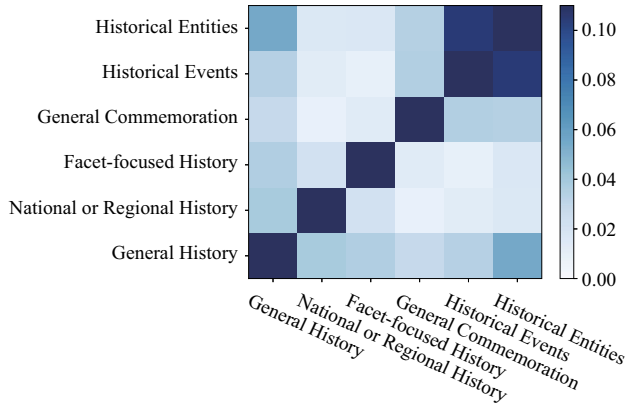

Fig. 18 Hashtag category co-occurrence in tweets

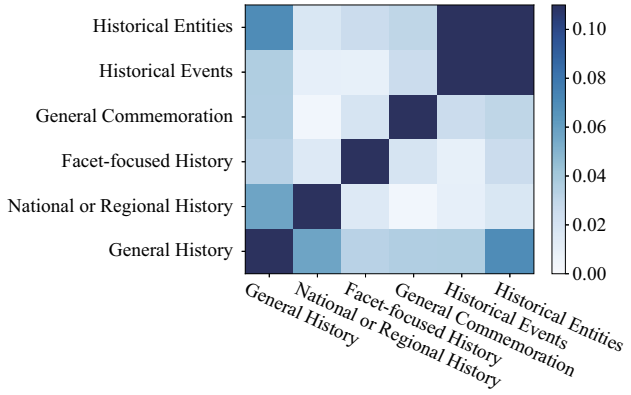

Fig. 19 Category co-occurrence in re-tweets

calculating the inter-category correlation, we apply Eq. 2 and we use the numbers of tweets that include hashtags classified into category A and B as T_A and T_B, respectively.

Figures 18 and 19 plot the calculated co-occurrence values between the categories. We can see that the *category* ***General History*** *is truly "general" since the hashtags in this category tend to highly co-occur with hashtags in the other categories*. This is the only category with relatively high similarity to every other category. Another observation is that the co-occurrence value between **Historical Events** and **Historical Entities** is quite high. This is because many famous entities in our dataset were involved in key events in the past (e.g., Stalin, Hitler in WWII). Similarly, when users refer to the past as a general commemoration they tend to focus on particular well-known past events and key persons. Hence, in tweets, **Historical Events/Historical Entities** and **General Commemoration** hashtags are sometimes used together. This is a unique phenomenon of tweets compared to re-tweets. On the other hand, **National or Regional History** and **Facet-focused History** category hashtags rarely co-occur with hashtags of other categories (except ones from **General History**). This indicates that hashtags under these two categories tend to be assigned to relatively unique and specialized content.

5.3 Temporal category analysis

We now investigate temporal references in tweets in relation to their categories. Our interest is in understanding how similar are time references included within tweets annotated with the hashtags of the same category (or in other words, whether tweets under the same category tend to mention similar or rather different years).

We compute such temporal coherence for each category by comparing the vectors of hashtags in a given category. These are built based on temporal expressions associated with the hashtags. In particular, for each hashtag, we construct a vector representing year scores derived from temporal references within tweets labeled by this hashtag. We first map all the temporal references to the year level granularity and then compute year scores using Eq. 1. Such vectors reflect commonly mentioned years for each hashtag. Pairwise similarities of hashtags falling into the same category are then computed by using the cosine similarity measure and are averaged to give the final scores displayed in Table 24. The way to compute the cosine similarity measure is as follows:

$$CosSim(A, B) = \frac{A \cdot B}{|| A || || B ||} \tag{4}$$

where: A and B are vectors corresponding to two different categories and $\cdot$ and $|| A ||$ are a dot product and magnitude, respectively.

Table 24 shows that the time-based similarities of hashtags in **General History**, **Historical Entities** and **National or Regional History** are relatively high (when compared to the average value for the entire data shown in the last row). However, one should keep in mind that many tweets under these hashtags lack any time references as indicated by their lower than average ratio values (see the 4th column). Nevertheless, the hashtags under these three categories tend to be relatively similar to each other in terms of focused time periods. This actually is not surprising as in the **Historical Entities** category several hashtags represent entities with overlapping lifetimes (at least this is the case in our dataset). Yet for **General History** and **National or Regional History** it would mean that there is a good level of agreement in the question of the most important historical periods and events (e.g., WWI, WWII, 11th Sep. for U.S. and 11th Mar. for Japan).

We then perform the same analysis this time for the re-tweets and show the results in Table 25. Interestingly, all of the similarity scores are lower than ones of tweets, especially, the score for **General History** decreased about 0.22 and the decrease for the average score is about 0.17. In con-

Table 24 Average cosine similarity for each category based on years in tweets (2nd column), standard deviation of the similarities (3rd column) and the rate of tweets including time references (4th column)

Categories	Similarity	Std. dev.	Ratio
General History	0.71	1690.71	0.16
National or Regional History	0.62	756.84	0.17
Facet-focused History	0.59	949.40	0.28
General Commemoration	0.53	2066.37	0.59
Historical Events	0.42	2288.96	0.18
Historical Entities	0.65	375.36	0.10
All	*0.59*	*1354.61*	*0.25*

Table 25 Average cosine similarity for each category based on years in re-tweets (2nd column), standard deviation of the similarities (3rd column) and the rate of tweets including time references (4th column)

Categories	Similarity	Std. dev.	Ratio
General History	0.49	1966.00	0.29
National or Regional History	0.43	792.66	0.22
Facet-focused History	0.44	1967.30	0.40
General Commemoration	0.39	4599.80	0.63
Historical Events	0.34	4847.68	0.26
Historical Entities	0.41	451.43	0.14
All	*0.42*	*2437.48*	*0.32*

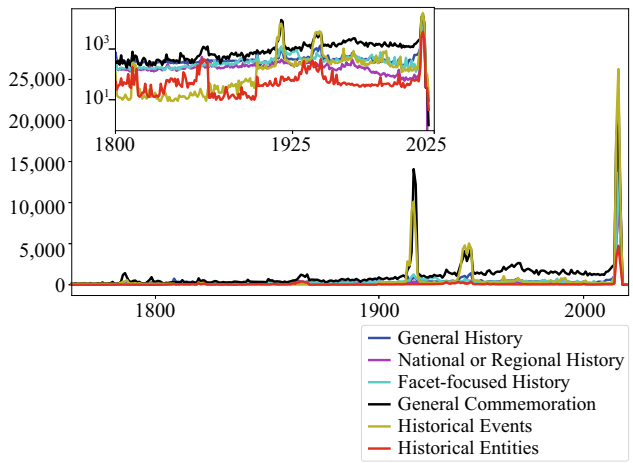

Fig. 20 Distributions of time references in the categories (small inner graph shows the plots in log scale)

trast, the values of the standard deviations increased over all categories. On average, they are twice as high as ones for tweets.

We next plot in Fig. 20 the distributions of time references extracted from tweets in each category. Naturally, all the categories have strong relation to the present (e.g., to the current events or present entities) since years 2016 ∼ 2018 are characterized by a high peak for all the plots. **Historical Events** category is strongly focused on the two dramatic events of the last century: WWI and WWII. While **General Commemoration** has also strong focus on the two wars, it features actually a very different time plot when compared to the ones from other categories. Since the common reason for commemorating events is their anniversaries (e.g., `#otd`) rather than external triggers such as ongoing events, the *tweets under **General Commemoration** relate to many diverse years in the past. The pattern of tweeting under this category reflects thus higher diversity of the collective attention towards time periods of history*.

5.4 Entity-focused category analyses

Next, we look into entity distributions to investigate them in each category. In a similar way to the above temporal analysis, we first count how many times each entity is mentioned with a given hashtag in order to create a hashtag-based entity vector. We then calculate pairwise similarities between hashtags in each category by comparing their vectors and we average these similarities to compute the final score per each category.

In Table 26 per each category, we show the following data: the average similarities, the standard deviations of similarities, the rates of all the entities, the rates of the past entities and the rates of present entities. The entity-focused similarities are in general relatively low, indicating that hashtags in the same category tend to refer to different entities. Interestingly, when looking at Table 26 we can observe that the average similarities of **General History**, **Facet-focused History** and **General Commemoration** are higher than the average values for the entire dataset (see the last row). In particular, over one-third of tweets under **General Commemoration** category contain entities (see the 4th column in Table 26), while more than half of its tweets contain time references as shown in Table 24. Thus, *compared with other categories, users tend to include more named entity names and more temporal expressions in tweets which are tagged with the hashtags from **General Commemoration** category*. This may be a distinguishing characteristic of commemorating activity. Similarly, *tweets using hashtags of **Facet-focused History** are characterized by the time references and entities whose similarities are higher than the average values*. Looking at similarities of the two categories in Tables 24 and 26, both the scores for **Facet-focused History** are higher than ones for the **General Commemoration** category. This result indicates a particular characteristic of **Facet-focused History**; users tend to tweet about common entities and dates using hashtags of this category. Moreover, when looking at **National or Regional History**, we observe that its similarity is quite low even though locations are common entities as we have observed in Figs. 1, 2, 3 and 5 and Tables 9,10, 16 and 17. Looking at Table 29 that is discussed in the last part

Table 26 Average cosine similarity for each category based on entities in tweets (2nd column), its standard deviation (3rd column) and the rates of tweets including: entities, past entities and present entities, displayed, respectively, in the last three columns

Categories	Sim.	Standard deviation	All	Past	Present
General History	0.34	3331.34	0.21	0.09	0.12
National or Regional History	0.07	1419.50	0.32	0.08	0.23
Facet-focused History	0.28	2489.54	0.28	0.11	0.17
General Commemoration	0.23	7566.31	0.44	0.19	0.25
Historical Events	0.15	1938.28	0.25	0.10	0.15
Historical Entities	0.11	676.07	0.22	0.14	0.08
All	*0.20*	*2903.51*	*0.28*	*0.12*	*0.17*

Table 27 Average cosine similarity for each category based on entities in re-tweets (2nd column), its standard deviation (3rd column) and the rates of re-tweets including: entities, past entities and present entities, displayed, respectively, in the last three columns

Categories	Sim.	Std. dev.	All	Past	Present
General History	0.13	3045.30	0.29	0.13	0.16
National or Regional History	0.07	920.10	0.31	0.09	0.21
Facet-focused History	0.14	2287.34	0.32	0.15	0.17
General Commemoration	0.13	8023.04	0.43	0.19	0.24
Historical Events	0.13	1957.08	0.31	0.12	0.19
Historical Entities	0.16	418.88	0.20	0.10	0.11
All	*0.13*	*2775.29*	*0.31*	*0.13*	*0.18*

Table 28 Top five years (2nd row marked as "Y"), entities (3rd row marked as "E") and hashtags (4th row marked as "H") of **General History** for tweets (T) and re-tweets (R). "*" is used to denote abbreviations made for saving space (*`histteacher`: `historyteacher`, *`twstorians`: `twitterstorians`)

		1	2	3	4	5
Y	T	1945	1944	1942	2015	1914
	R	1944	1945	1942	1941	1943
E	T	US	UK	France	Canada	WWII
	R	US	UK	France	Germany	Canada
H	T	`history`	*`twstorians`	`memories`	*`histteacher`	`archives`
	R	`history`	*`twstorians`	`histoire`	*`histteacher`	`earlymodern`

of this sub-section, there are past places (Ancient Egypt and Ottoman Empire) and the nineteenth century years as top entities and years. Also, looking back at Tables 24 and 25, we notice that the rate of tweets including past entities is the lowest of the 6 hashtag categories. These results indicate that authors of tweets in our dataset may mention several kinds of past entities and it is difficult for NER tools to extract entities from the texts.

In Table 27, we show results of the same analysis but this time done on re-tweets. First, we can observe that both the average scores of similarity and standard deviation are now lower than ones for tweets, while ratios of included entities are almost the same. Interestingly, we can observe that the score of the similarity for **General Commemoration** becomes lower than one for tweets, while standard deviation becomes higher. Looking at the ratios of tweets including all, past and present entities (the 4th, 5th and 6th columns of the two tables, respectively), the scores are almost the same; the past's score is 0.19 for both the tweets and re-tweets and the differences for All and Present are only 0.01 point between the tweets and re-tweets. Thus, re-tweets have more kinds of entities than tweets.

To analyze the four hashtag categories (**General History**, **National or Regional History**, **Facet-focused History** and **General Commemoration**) in detail, we show top 5 years, entities and hashtags in Tables 28, 29, 30 and 31. In these tables, we remove three years from 2016 to 2018 because these years should occupy top-3 as Fig. 20 showed that there are high peaks on the three years for all the categories. Similar to the above, for **General History** and **General Commemoration** categories, we observe WWI/WWII related years and entities related to these events. Looking at the **National or Regional History** category, the two years in the 19th century (1861 and 1870) are ranked as one of the most popular years in the category for tweets and re-tweets. American Civil War is the main topic for 1861 since this year was the start year of the war. On the other hand, there are several different topics for 1870 in re-tweets; for example, some users refer

Table 29 Top 5 years (2nd row marked as "Y"), entities (3rd row marked as "E") and hashtags (4th row marked as "H") of **National or Regional History** for tweets (T) and re-tweets (R). "*" is used to denote abbreviations made for saving space (*ushist: ushistory, *localhist: localhistory, *americanhist: americanhistory, *cdnhist: cdnhistory)

		1	2	3	4	5
Y	T	1861	1900	1934	1967	1917
	R	1870	1917	1910	1900	1911
E	T	Canada	US	Egypt	Greece	Ancient Egypt
	R	Canada	US	Egypt	Ottoman Empire	Adolf Hitler
H	T	*ushist	cdnhist	ottoman	*americanhist	*localhist
	R	cdnhist	*ushist	thirdreich	*localhist	*cdnhist

Table 30 Top 5 years (2nd row marked as "Y"), entities (3rd row marked as "E") and hashtags (4th row marked as "H") of **Facet-focused History** for tweets (T) and re-tweets (R). "*" is used to denote abbreviations made for saving space (*A. A.: African American, *arthist: arthistory, *tbh: todayinblackhistory)

		1	2	3	4	5
Y	T	1917	1916	1920	1928	1927
	R	1918	1917	1890	1888	1927
E	T	US	UK	Canada	France	Sydney
	R	US	*A. A.	UK	Ida B. Wells	Canada
H	T	*arthist	envhist	silentfilm	histSTM	histmed
	R	*arthist	histmed	envhist	histSTM	*tbh

Table 31 Top 5 years (2nd row marked as "Y"), entities (3rd row marked as "E") and hashtags (4th row marked as "H") of **General Commemoration** for tweets (T) and re-tweets (R). "*" is used to denote abbreviations made for saving space (*NYT: The New York Times, *tbt: throwbackthursday)

		1	2	3	4	5
Y	T	1917	1918	1945	1942	1916
	R	1917	1918	1945	1944	1916
E	T	US	*NYT	France	UK	WWII
	R	US	France	UK	*NYT	WWII
H	T	otd	tdih	remembering	*tbt	WeRemember
	R	onthisday	WeRemember	botd	otd	*tbt

to Canadian history, while others share photographs about Shanghai or the Ottoman empire. Furthermore, in Table 30 #arthistory is ranked as one of the most popular hashtags in the category of **Facet-focused History**. After manual check, we found that some users mentioned that Van Gogh moved to Auvers–sur–Oise in 1890. Twitter authors as found in our dataset sometimes discuss art-related history as shown in Table 22 and Fig. 23. In this kind of tweets, the users not only talk about birth, death or various activities of famous artists, but also show their art works[18].

[18] E.g., https://twitter.com/ImadSalamoun/status/1000575925229178880 and https://twitter.com/ImadSalamoun/status/1005281316550606849

5.5 Analysis of entity and time reference dispersions

Finally, we study the dispersions of entities and time expressions for each category. To do this, we first calculate the entropy measures for each hashtag in terms of named entities, time references, users, and we also compute the frequency of tweets that include given hashtags. The entropy is defined as follows:

$$Entropy(h_e) = -\sum_{e \in HElm(h_e)} p(e) \log p(e) \quad (5)$$

where e represents an element being either a named entity, time reference, or a user depending on the computation objective, while h_e, and $p(e)$ are a hashtag associated with the element, and its probability for e, respectively. Also,

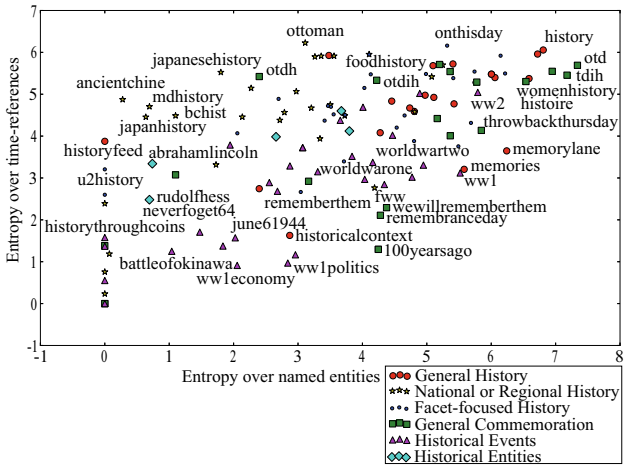

Fig. 21 Named entities (x axis) vs. time references (y axis)

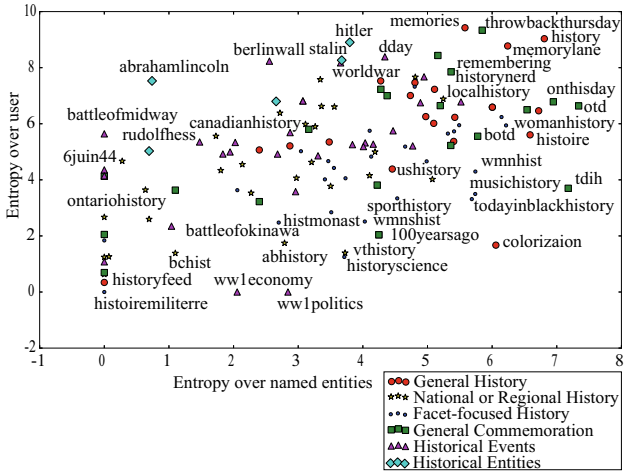

Fig. 22 Named entities (x axis) vs. users (y axis)

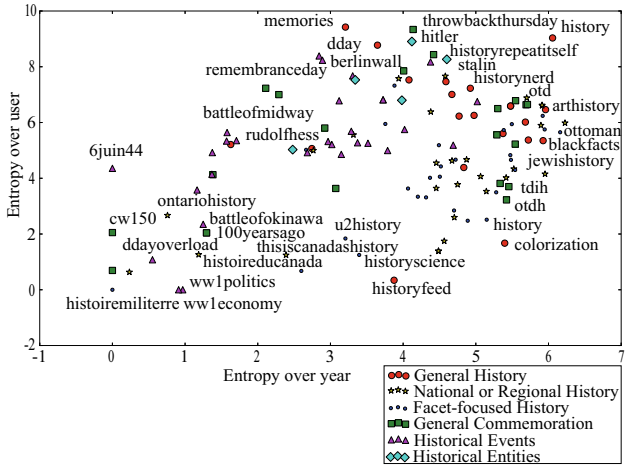

Fig. 23 Time references (x axis) vs. users (y axis)

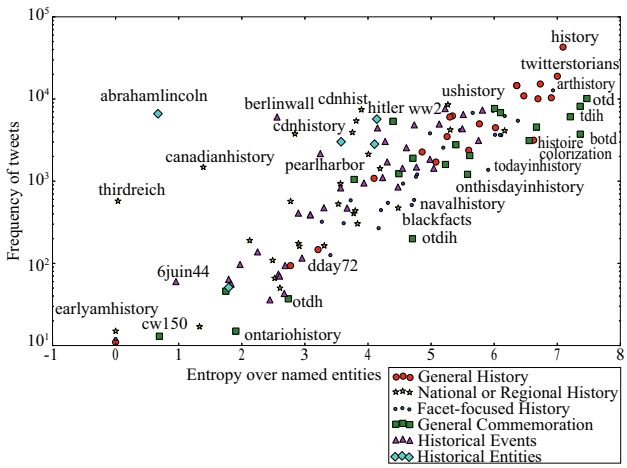

Fig. 24 Entropy over named entities (x axis) vs. hashtag frequency (y axis)

$HElm(h_e)$ is a set of e where every e is related to h. The probabilities of e are calculated using the numbers of tweets that include h used with e in the case of the named entities and time references, or that are tweeted by e in the case of users.

In Figs. 21, 22, 23, and 24, the hashtags are positioned according to their entropy scores. To facilitate observing the tendency of different categories, we assign colors to the hashtags depending on their categories.

Figure 21 places the hashtags according to their entropy values calculated over the distributions of contained entities and mentioned temporal expressions. We can observe that **General Commemoration** hashtags (e.g., `#otdih`, `#otd`) and **General History** hashtags (e.g., `#history`) are often characterized by high values of the two entropies. This is expected due to the general character of these kinds of hashtags. On the other hand, the **Historical Events** and **Historical Entities** categories tend to have low values of the entropy over temporal references, which is understandable given that they focus on relatively short-lasting events (e.g., `#june61944`, `#ww1politics` and `#ww1economy`) or particular entities. The lower entropy values over time references and entities is also noticeable for the **Historical Events** hashtags that refer to narrower topics such as `#ww1politics`, unlike ones indicating more general events like `#ww1` and `#ww2`. For the same reason, these categories also achieve low values of entity-based entropy without `#worldwarone`, `#ww1`, `#worldwartwo` and `#ww2`. Furthermore, **National or Regional History** hashtags (e.g., `#japanhistory`) tend to have less varying entities which is due to their less broad topical scopes. On the other hand, the entropies of their time references are on average high. In general, the broader the theme and the longer the related time scope of a particular hashtag, the higher are the entropies of the hashtag.

Next, in the remaining two figures (Figs. 22 and 23) we analyze the relation between the entropy of user distributions and the entropy of entity distributions as well as the relation between the entropy of user distributions and the entropy of temporal references.

The right-hand side corner of Fig. 22 is occupied by both ***General Commemoration*** *and* ***General History*** *hashtags suggesting that many different users tweet with these hashtags and the users tend to include references to many different entities*. **Historical Entities** in our dataset like `#stalin` and `#hitler` tend to be referred to by large numbers of users, and interestingly, they also have relatively high values of the entity entropy. This is likely because the hashtags of this category refer to rather famous historical persons. In Fig. 23, **General Commemoration** tweets contain many diverse dates and are issued by many users. In contrast, we can see that user bases of hashtags of **National or Regional History** (e.g., `#jewishhisotry` and `#thisiscanadashistory`) and **Facet-focused History** (e.g., `#historyscience` and `#u2history`) are rather low.

Lastly, in Fig. 24, we show the entropy over entities vs. hashtag frequency. We can observe the positive correlation between the two measures. Naturally, general hashtags (e.g., `#history` and `#otd`) embrace the use of many different entities and are quite frequent and, as we can see, they occupy the right-hand upper side of the figure in contrast to more focused hashtags (e.g., `#dday72`), which are used with relatively few entities and with less frequency.

6 Discussion

6.1 Summary of main findings

- Based on the rates of entity types collected from DBpedia (Fig. 3), we conclude that persons, places and groups tend to be frequently mentioned in history-focused tweets and the person category is especially common in these types of tweets. In particular, tweets including groups focus on wars and conflicts from the past as the top entities in the group include many military units, e.g., US Army, Royal Air Force (Tables 4 and 5).
- In general, the recent past is referred to more than the distant past, and the memory decay is fastest in the recent years (Fig. 4), which is quite expected. There are, however, several periods that are against this tendency. Users tend to focus more on years which represent two key events (WWI and WWII) in the last century and years falling exactly 50 or 100 years ago from the data collection year. Recalling events that occurred 100 years ago is also observed in Fig. 8 (e.g., `#ww1` is one of the top hashtags for 2016) and in Tables 7 and 8 (e.g., `#somme100` and `#100yearsago` are top hashtags there).
- According to the analysis of co-occurrences of the different types of hashtags (Figs. 9 and 10), the entities in the category of present places tend to co-occur with the entities of any other type. Both the present and past persons also tend to co-occur with entities of any other types excluding past events.
- Commemorative hashtags like `#onthisday`, `#otd`, and `#throwbackthursday` are used in a relatively large number of tweets and re-tweets (Figs. 11, 12, 13, 16 and 17).
- Users often enrich tweets with images (e.g., Instagram) or videos (e.g., YouTube) (as shown in Figs. 14 and 15) as the past is often recalled by making use of diverse artifacts like images, videos, objects or historical documents. Indeed, News, Education, Arts, and Entertainment are the top 5 common URL categories (Table 22).
- We have proposed 6 categories of history-related hashtags: **General History**, **National or Regional History**, **Facet-focused History**, **General Commemoration**, **Historical Events** and **Historical Entities** (Table 23). According to the analysis of inter-category similarity, the hashtags in the **General History** category tend to co-occur with hashtags in all the other categories (Figs. 18 and 19). Also, the similarity score between **Historical Events** and **Historical Entities** is quite high as many famous entities in our dataset were involved in key events (e.g., Stalin, Hitler in WWII).
- Temporal category analysis (Fig. 20) demonstrated that the tweets under the **General Commemoration** category relate to many diverse years in the past. The pattern of tweeting under this category reflects thus higher diversity of the collective attention towards time periods of history. In addition, users tend to include more entity names into tweets tagged with the hashtags from this category (Table 26).
- Looking at the results of the relation between entropy of entity distributions and entropy of user distributions (Fig. 22), we found that many different users tweet with either **General Commemoration** or **General History** hashtags and they tend to refer to many diverse entities.
- These results of comparison between tweets and re-tweets reveal many shared tendencies; for example, Place and Person types of entities are popular, the recent past is referred to more than the distant past, and there are significant attention peaks on timelines representing key events (WWI and WWII). In contrast, the category-based analyses suggest several differences between tweets and re-tweets. Although the three categories: **General History**, **General Commemoration** and **Historical Events** are most commonly used, their order is different (Figs. 16 and 17). The **General History** is the most popular in

tweets, whereas the **General Commemoration** is the most common in the case of re-tweets. As another difference, in tweets, **Historical Entities**, **Historical Events** and **General Commemoration** hashtags are sometimes used together. However, in re-tweets, **General Commemoration** hashtags are not used with **Historical Entities** or **Historical Events** compared with tweets (Figs. 18 and 19).

6.2 Limitations

Data Collection. We note that the data collection method that we relied on naturally misses a certain portion of tweets. There are three reasons for this. First, our method cannot collect tweets which are not tagged by any history-related hashtags as it relies on hashtags for retrieving relevant tweets. Second, it is impossible to collect all tweets tagged by history-related hashtags because the official Twitter's Search API is known to not return all the tweets using a given hashtag. Third, we may have missed tweets that include history-related hashtags if they were posted before the hashtags got identified as history-related during our data collection process. This is because the Search API only provides the most recent 7-10 days of data (tweets).

We list here three other potential approaches that could be used to collect history-related data: (1) *collecting content with temporal expressions pointing to the past*, (2) *collecting content that contains past entities*, and (3) *collecting tweets by inputting history-related words or hashtags in the "Search Twitter" window of Web UI*. Every approach is, however, not without its shortcomings. The first method was used in [5] for extracting past references in news articles and relied on the presence of temporal expressions in text. This, however, is not always guaranteed for history-related tweets. Indeed, as can be seen in Table 3 the rate of tweets with temporal expressions is 40%; hence, about 4 tweets out of 10 contain tweets with any time reference. Thus, this approach would miss many relevant tweets resulting in rather low recall. Similarly, history-related tweets may not mention any past entity. Indeed, from Table 26 (see the column "Past" and the last row) we can notice that the rate of tweets containing at least one past entity, which can be recognized by the state-of-the-art tools, is only 0.12 for our dataset. Furthermore, some entities, especially more obscure ones may not be present in any knowledge base or may not be detectable using standard tools. We thus assumed in this work an approach that relies on extracting explicit history-focused hashtags and on subjecting them to the manual analysis of tagged content. While such a choice is likely characterized by a high precision, it may obviously suffer from lowered recall as discussed before. Yet, in the view of the reported statistics, we still believe it is superior to collecting tweets based on contained dates or historical entities. Future work should nevertheless explore more refined approaches for extracting implicit past-related content, ideally ones making use of the combination of all the signals (hashtags, past entities, temporal expressions, etc.). We also note that while we put great effort into the manual verification of used hashtags as for their relevance to the history, there is always a chance that some might not be fully devoted to the past, or, in general, their selection may cause certain biases.

Different Languages. In the current study, we have mainly focused on English tweets. Further exploration should involve different languages (e.g., French, Japanese) as well as the cross-comparison of the obtained results. Our current aim is, however, not to look into particular aspects and specificities related to different countries but to rather uncover general tendencies. The focus on English was a natural decision for this initial study due to the international role and ubiquity of this language.

Extracting Original Tweets from Quote Tweets. In this study, we treat quote tweets as original tweets. Due to relatively small number of quote tweets in our dataset (0.02%), this choice should not impact the results. Nevertheless, analyzing quote tweets may lead to discovering supplementary content added by others that may correct, extend or comment on the original tweets (e.g., content that incorporates key missing information or one that provides novel aspects of past events & entities).

Analyzing Re-tweets as Original Tweets. Some contents in re-tweets may not appear in our original tweet dataset; thus, re-tweets may not represent the actual re-tweeted content of the original tweets that we collected. We could circumvent this by extracting content from re-tweets in order to incorporate it into our tweet dataset in case such content does not appear there. However, we decided to separate tweets and re-tweets for performing simple comparative analysis between these two. Otherwise, the results of the analysis could be regarded as potentially biased.

Fine-grained Analysis. The present study is mainly quantitative aiming to provide first glimpse into the issue of historical references in social network services and to conduct broad exploratory analysis. Deeper qualitative exploration should be later conducted for obtaining fine-grained comprehension of the way in which users refer to the past. For example, the future analysis could examine why some entities are (or are not) popular as well as could apply sentiment analysis to identify tendencies in polarity towards particular past events or entities. Another research direction could be the detailed analysis of contexts in which the past entities or past years are mentioned. These would necessarily require some sort of manual and qualitative exploration of tweet content.

User-focused Analysis. Our study has exploratory character and focuses on the shared content first. However, a very interesting question is about the type of users who share or

are interested in historical content in social network services. Due to the time and space constraints, we have left, however, the user-focused analysis for future work.

6.3 Potential applications

Finally, we list here several example applications that could be potentially constructed based on the history-related content shared in Twitter:

1. Recommending past-related content for readers interested in studying history. This could be, for example, popular and interesting content or the content that matches particular user interests such as tweets under hashtags of the **Facet-focused History** category that a user is interested in. Through this analysis and other forthcoming ones, we could better understand what kind of history-related content at what time periods is becoming attractive to many users.
2. Creating history-focused chatbots such as HistoChatbot[19] for disseminating historical knowledge and for entertaining users.
3. Finding, summarizing and explaining past entities which are mentioned in relation with the popular present entities to provide analogy and a novel, potentially interesting context for the latter.
4. Automatically summarizing and comparing history-related opinions and popular topics across different regions.
5. Automatically suggesting hashtags for tweets based on included entities, years and based on the predicted hashtag categories.

7 Conclusions and future work

In this paper, we have studied how users refer to the history in microblogging and in which contexts such references occur. As mentioned before, history-focused content recommendation should offer opportunities for the dissemination of interesting and trendy recollections by pushing them to users.

Our analysis is broad and is meant to lay the groundwork in establishing how microbloggers conceive of, share and refer to history-related content such as one on past events and persons. Through this exploratory study, we hope to shed more light on the way in which history-related content is used and shared in microblogging, and by this to encourage subsequent research and development of systems aiming at educating history. We perform basic study on a coarse level, providing initial observations, identifying several interesting research directions and suggesting potential applications. Our analysis is nevertheless conducted from multiple perspectives.

Future work will (a) *identify differences between general and personal histories as well as will look into what makes tweets about personal history appear interesting*. Social media provides novel opportunities to create such personal connections which can help raise an interest in the significance of historical knowledge beyond the personal experience. Future work will also include (b) *geographical analysis of tweets which may point to different cultural practices with regard to references to the past*. As currently most of the tweets in our dataset originate from the English-speaking part of the world, we should contrast these results with ones obtained on data collected in different languages. Next, we plan also to (c) *explore the interdependence between present-day events, their function as triggers for references to history and the latter's effect on the interpretation of the present*. Furthermore, as mentioned before, (d) *we plan to study in detail the characteristics of users sharing history-related content in Twitter* such as their demographics, characteristics of their followers and followees, and their interaction patterns.

Finally, (e) we will *build classifiers for determining the categories of tweet content that were introduced in this study*. As we discussed in the limitations, the current method for tweet collection relies on the proper selection of hashtags. To increase the coverage, we will define tweets that are not history-related and we are planning to train a binary classifier for detecting history-related tweets.

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

References

1. Abelson, R.P., Levi, A.: Decision Making and Decision Theory, Handbook of Social Psychology, pp. 231–309. Random House, New York (1985)
2. Abujabal, A., Berberich, K.: Important events in the past, present, and future. In: WWW'15, pp. 1315–1320. ACM, New York (2015)

[19] https://twitter.com/HistoChatbot

3. Althoff, T., Dong, X.L., Murphy, K., Alai, S., Dang, V., Zhang, W.: Timemachine: Timeline generation for knowledge-base entities. In: KDD'15, pp. 19–28. ACM, New York (2015)
4. Arnold, M.V., Dewhurst, D.R., Alshaabi, T., Minot, J.R., Adams, J.L., Danforth, C.M., Dodds, P.S.: Hurricanes and hashtags: Characterizing online collective attention for natural disasters (2020)
5. Au Yeung, C.m., Jatowt, A.: Studying how the past is remembered: towards computational history through large scale text mining. In: CIKM'11, pp. 1231–1240. Glasgow, Scotland, UK (2011)
6. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: ISWC'07/ASWC'07, pp. 722–735. Busan, Korea (2007)
7. Baeza-Yates, R.: Searching the future. SIGIR Workshop MF/IR'05, ACM (2005)
8. Bates, A., Kalita, J.: Counting clusters in twitter posts. In: ICTCS'16, pp. 85:1–85:9. ACM, New York, NY, USA (2016)
9. Bogaard, T., Hollink, L., Wielemaker, J., Hardman, L., van Ossenbruggen, J.: Searching for old news: User interests and behavior within a national collection. In: CHIIR'19, pp. 113–121. ACM, New York, NY, USA (2019)
10. Boschi, G., Cammarota, C., Kühn, R.: Opinion dynamics with emergent collective memory: a society shaped by its own past. Phys. A Statist. Mech. Appl. **558**, 124909 (2020)
11. Campos, R., Dias, G., Jorge, A.M., Jatowt, A.: Survey of temporal information retrieval and related applications. ACM Comput. Surv. CSUR **47**(2), 15 (2015)
12. Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.L., Hidalgo, C.A.: The universal decay of collective memory and attention. Nat.e Human Behav. **3**(1), 82–91 (2019)
13. Chew, M.M., Bhowmick, S.S., Jatowt, A.: Ranking without learning: Towards historical relevance-based ranking of social images. In: SIGIR'18, pp. 1133–1136. ACM, New York (2018)
14. Clavert, F., Majerus, B., Beaupré, N.: #ww1. twitter, the centenary of the first world war and the historian. Twitter for Research (2015)
15. Comarela, G., Crovella, M., Almeida, V., Benevenuto, F.: Understanding factors that affect response rates in twitter. In: HT'12, pp. 123–132. ACM, New York (2012)
16. Cook, J., Sarma, A.D., Fabrikant, A., Tomkins, A.: Your two weeks of fame and your grandmother's. In: WWW'12, pp. 919–928. Lyon, France (2012)
17. Do, Q.X., Lu, W., Roth, D.: Joint inference for event timeline construction. In: EMNLP-CoNLL'12, pp. 677–687. ACL, Stroudsburg (2012)
18. Ebbinghaus, H.: Memory: A Contribution to Experimental Psychology (reprint). Martino Fine Books (2011)
19. Ferron, M., Massa, P.: Collective memory building in wikipedia: The case of north african uprisings. In: WikiSym'11, pp. 114–123. Mountain View, California (2011)
20. G.-Gavilanes, R., Mollgaard, A., Tsvetkova, M., Yasseri, T.: The memory remains: understanding collective memory in the digital age. Sci. Adv. **3**(4) (2017)
21. Giachanou, A., Crestani, F.: Opinion retrieval in twitter: is proximity effective?. In: SAC'16, pp. 1146–1151. ACM, New York (2016)
22. Gillani, M., Ilyas, M.U., Saleh, S., Alowibdi, J.S., Aljohani, N., Alotaibi, F.S.: Post summarization of microblogs of sporting events. In: WWW'17 Companion, pp. 59–68. Republic and Canton of Geneva, Switzerland (2017)
23. Gilovich, T.: Seeing the past in the present: the effect of associations to familiar events on judgments and decisions. J. Personal. Soc. Psychol. **40**(5), 797 (1981)
24. Graus, D., Odijk, D., de Rijke, M.: The birth of collective memories: analyzing emerging entities in text streams. J. Assoc. Inf. Sci. Technol. **69**(6), 773–786 (2018)
25. Halbwachs, M.: La Memoire Collective. Les Presses universitaires de France, (in French) (1950)
26. Hoerl, C., McCormack, T.: Time and Memory: Issues in Philosophy and Psychology. Oxford University Press, Oxford (2001)
27. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: EMNLP'11, pp. 782–792 (2011)
28. Holzmann, H., Risse, T.: Named entity evolution analysis on wikipedia. In: WebSci'14, pp. 241–242. ACM, New York (2014)
29. Huet, T., Biega, J., Suchanek, F.M.: Mining history with le monde. In: AKBC'13, pp. 49–54. ACM, New York (2013)
30. Jacoby, R.: Social Amnesia: A Critique of Contemporary Psychology. Transaction Publishers, Piscataway (1997)
31. Jatowt, A., Kawai, D., Tanaka, K.: Predicting importance of historical persons using wikipedia. In: CIKM'16, pp. 1909–1912. ACM, New York (2016)
32. Jatowt, A., Kawai, D., Tanaka, K.: Timestamping entities using contextual information. In: SIGIR'17, pp. 1205–1208. ACM, New York (2017)
33. Jatowt, A., Kawai, H., Kanazawa, K., Tanaka, K., Kunieda, K., Yamada, K.: Multi-lingual analysis of future-related information on the web. In: Culture and Computing'13, pp. 27–32 (2013)
34. Jatowt, A., Kanazawa, K., Oyama, S., Tanaka, K.: Supporting analysis of future-related information in news archives and the web. In: JCDL'09, pp. 115–124. ACM, New York (2009)
35. Jatowt, A., Kawai, D., Tanaka, K.: Time-focused analysis of connectivity and popularity of historical persons in wikipedia. Int. J. Dig. Libr. **20**(4), 287–305 (2019)
36. Kanhabua, N., Nguyen, T.N., Niederée, C.: What triggers human remembering of events?: A large-scale analysis of catalysts for collective memory in wikipedia. In: JCDL'14, pp. 341–350. London, United Kingdom (2014)
37. Kanhabua, N., Niederée, C., Siberski, W.: Towards concise preservation by managed forgetting: Research issues and case study. In: iPres'13 (2013)
38. Kiprov, Y., Gencheva, P., Koychev, I.: Generating labeled datasets of twitter users. In: UMAP'17, pp. 191–196. ACM, New York (2017)
39. Koutlis, C., Schinas, M., Gkatziaki, V., Papadopoulos, S., Kompatsiaris, Y.: Data-driven song recognition estimation using collective memory dynamics models. In: ISMIR'19, pp. 368–375 (2019)
40. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. pp. 497–506. In: SIGKDD'09, ACM, New York (2009)
41. Magdy, A., Ghanem, T.M., Musleh, M., Mokbel, M.F.: Understanding language diversity in local twitter communities. In: HT'16, pp. 331–332. ACM, New York (2016)
42. McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., McCullough, D.: On building a reusable twitter corpus. In: SIGIR'12, pp. 1113–1114. ACM, New York (2012)
43. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: SIGIR'13, pp. 889–892. ACM, New York (2013)
44. Miz, V., Benzi, K., Ricaud, B., Vandergheynst, P.: Wikipedia graph mining: dynamic structure of collective memory. arXiv preprint arXiv:1710.00398 (2017)
45. Nielek, R., Wawer, A., Wierzbicki, A.: Collective memory in poland: A reflection in street names. SocInfo Workshops QMC'13, vol. 8359, pp. 134–142. Springer (2013)
46. Nwala, A.C., Weigle, M.C., Nelson, M.L.: Bootstrapping web archive collections from social media. In: HT'18, pp. 64–72. ACM, New York (2018)
47. Odijk, D., de Rooij, O., Peetz, M.H., Pieters, T., de Rijke, M., Snelders, S.: Semantic document selection. In: TPDL'12, pp. 215–221. Springer Berlin (2012)

48. Peng, M., Zhu, J., Li, X., Huang, J., Wang, H., Zhang, Y.: Central topic model for event-oriented topics mining in microblog stream. In: CIKM'15, pp. 1611–1620. ACM, New York, NY, USA (2015)
49. Radinsky, K., Davidovich, S., Markovitch, S.: Learning causality for news events prediction. In: WWW'12, pp. 909–918. ACM, New York (2012)
50. Radinsky, K., Horvitz, E.: Mining the web to predict future events. In: WSDM'13, pp. 255–264. ACM, New York (2013)
51. SalahEldeen, H.M., Nelson, M.L.: Predicting temporal intention in resource sharing. In: JCDL'15, pp. 205–214. ACM, New York (2015)
52. Singh, J., Nejdl, W., Anand, A.: History by diversity: helping historians search news archives. In: CHIIR'16, pp. 183–192. ACM, New York (2016)
53. Strötgen, J., Andrade, R., Gupta, D.: Putting dates on the map: harvesting and analyzing street names with date mentions and their explanations. In: JCDL'18, pp. 79–88. ACM, New York (2018)
54. Strötgen, J., Gertz, M.: Temporal tagging on different domains: Challenges, strategies, and gold standards. In: LREC'12, pp. 3746–3753. ELRA, Istanbul, Turkey (2012)
55. Sumikawa, Y., Jatowt, A., Düring, M.: Digital history meets microblogging: Analyzing collective memories in twitter. In: JCDL'18, pp. 213–222. ACM, New York (2018)
56. Temnikova, I., Vieweg, S., Castillo, C.: The case for readability of crisis communications in social media. In: WWW'15 Companion, pp. 1245–1250. ACM, New York (2015)
57. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM'10 (2010)
58. Wagle, N., Jasani, S., Gawand, S., Tilekar, S., Patil, P.: Twitter userrank using hadoop mapreduce. In: WIR'16, pp. 150–153. ACM, New York (2016)
59. Wang, Y., Mohd Pozi, M.S., Siriaraya, P., Kawai, Y., Jatowt, A.: Locations & languages: Towards multilingual user movement analysis in social media. In: WebSci'18, pp. 261–270. ACM, New York (2018)
60. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: Finding topic-sensitive influential twitterers. In: WSDM'10, pp. 261–270. ACM, New York (2010)
61. Zhang, Y., Jatowt, A., Bhowmick, S., Tanaka, K.: Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In: ACL/IJCNLP'15, pp. 645–655. ACL (2015)
62. Zhang, Y., Jatowt, A., Tanaka, K.: Temporal analog retrieval using transformation over dual hierarchical structures. In: CIKM'17, pp. 717–726. ACM, New York (2017)