Event Timeline Generation for Tokyo 2020 Olympics

December 24, 2024

1 Introduction

In the era of user-generated content (UGC), billions of people are generating, retrieving, and disseminating trending news through social media. The proliferation of social media websites has led to the rapid growth of the users and UGC across the world. The contents generated by social media users are overwhelming in terms of their "velocity", "volume", and "variety", though the "veracity" and "value" of UGC is often questioned by the official presses. The real-time feature and collective wisdom nature of UGC on social media have attracted the attention of both researchers and practitioners to analyze its contents to gain timely knowledge and insights about the real-time states of domestic events and international affairs. Specifically, Twitter, as one of the most popular and leading micro-blogging and social media service providers, has been a recent focus and primary source for people to gather news on domestic events and international affairs happening in real time.

An "event" is an occurrence of interest in the real world that might result in heated discussions on the related topics by the public, either soon after the occurrence or even in anticipation of its occurrence (Li et al., 2012). The occurrence of an event is usually characterized by its entities, including time, location, person, and action. An occurrence of a breaking event is usually accompanied by a temporarily increased volume and velocity of UGC generated and disseminated on its related topics through social media in a specific time window. Due to the demand for retrieving and analyzing newsworthy events from UGC on social media, event detection and summarization system is widely studied by natural language processing (NLP) and social media researchers to generate the corresponding event timeline. Event timeline generation systems are often used to deliver the causes, evolution, and outcomes of a trending event to the public. Event timeline generation consists of two major tasks: event detection and event summarization (Guo et al., 2017). Event detection is used to detect and extract newsworthy trending events from real-time large-scale crowdsourced data. The events detected often correspond to a high

volume of unorganized texts containing valuable information about the event. As the texts are often unorganized, event summarization is used to effectively comprehend the core and essence of the events. Besides event detection and event summarization, event timeline visualization, as an value-adding task, can enhance the public awareness of ongoing events.

In this group project, we propose to extract the newsworthy events from streaming Twitter data and then generate and visualize the event timeline of the results of the Tokyo 2020 Summer Olympics. Tokyo 2020 Summer Olympics was an international multi-sport event held from 23 July to 8 August 2021 in Tokyo, Japan. It held multi-sports events in a limited time window and had precise timing cutoffs for sports event results. These characteristics make it an ideal source of event timeline generation tasks. Event timeline generation tasks can usually be classified into different types by three criteria (Hasan et al., 2018): (1) the type of event being detected by an event detection system (i.e., specified events versus unspecified events), (2) the detection task (i.e., new event detection versus retrospective event detection), and (3) the event detection methods (i.e., supervised, unsupervised or hybrid). The event timeline generation task of our project, therefore, falls into the specific category of (1) specified event – the sports event results of the Tokyo 2020 Summer Olympics, (2) retrospective event detection – it happened in the summer of 2021, and (3) unsupervised method – a clustering algorithm is used.

2 System Architecture

Our event timeline generation system consists of six components: (1) data acquisition, (2) data pre-processing, (3) event detection, (4) event summarization, (5) event ranking, and (6) event timeline visualization. Figure 1 in the Appendix displays the framework of our system. In this section, we only briefly introduce the functionality of each component, more technical details of our methodologies and algorithms used are discussed in the Methodologies and Algorithms section.

2.1 Data acquisition

The popular corpus collection methods for Twitter event detection include streaming data with a filter, queries using search API, crawlers used to obtain the dataset, and secondary data sources. Due to the limited time resources for the group project, we use secondary data of tweets from Kaggle as our primary data source to complete the task of event timeline generation. We also design customized crawlers to crawl the information we need for information extraction.

2.2 Data pre-processing

The data pre-processing stage of event timeline generation receives tweets and processes them to be useful inputs in subsequent tasks. The popular methods to be performed on tweets in the literature include part-of-speech (POS) tagging, slang-words conversion, and named entity recognition (NER). Some studies will also use tweet filtering methods based on particular criteria, such as excluding retweets and non-target language tweets from the corpus. Moreover, most studies will remove stop words, URLs, punctuation, and user-name mentions from tweets.

2.3 Event detection

Event detection is used to detect and extract newsworthy trending events from real-time large-scale crowd-sourced data. The events detected often correspond to a high volume of unorganized texts, containing valuable information about the event. In this project, we extract the five attributes of Tokyo 2020 sports events from Tweets text: (1) the name of the medalist, (2) the country the medalist belongs to, (3) the sports name, (4) the event name the medalist won (note that one sports can have multiple events, for example, the sports "athletics" has 48 events, and "badminton" has 5 events), and (5) the medal type that the medalist won (i.e., gold medal, silver medal, or bronze medal). The final important attribute of Tokyo 2020 sports events is (6) the timing of winning the medal, which is determined in event summarization.

2.4 Event summarization

The relevant information on events is extracted from the preprocessed data which contains a lot of unstructured events. Most of the tweets point to the same actual event. The following information extracted will be used in the event summarization: (1) **medalist** name, (2) **country** the medalist belongs to, (3) **sports** name, (4) **event** name, (5) **medal type** that the medalist won, (6) **timestamp** of the tweet. We want to group those related events that point to the same single medal event into one and remove the noise from the data at the same time. After event summarization, we should have a list of summarized events for the timeline generation task.

2.5 Event ranking

Event ranking is then applied to the summarized events. As the summarized events are supposed to contain the information required for the timeline generation but not all of them contain all the information. At the same time, as an advantage of data analytics, we believe that in data analysis, most used terms in different information [medalist, country, sport, event, medal type,

timestamp] should be highly related to the correct information we need. Therefore, we need to join all the information from the extracted information among all the summarized events. Then we can rank the frequency of wording used in each information required so that the higher frequency indicates it is more relevant to the event. Finally, we can gather the information for timeline generation.

2.6 Event timeline visualization

Visualizing temporal data is a critical step in demonstrating our event timeline generation system to enhance the public awareness of the real-time states of the on-going or past events. A timeline chart is a chart that graphically depicts how a series of chronological events happened over time (Hasan et al., 2013). Visualizing the event timeline with timeline charts is the last stage of our event timeline generation system to visually deliver the outputs.

3 Methodologies and Algorithms

3.1 Data acquisition

Due to the limited time resources for the group project, we use secondary data of tweets as our primary data source to complete the task of event timeline generation. The dataset author collected continuously using a script that collects a small number of recent tweets using Twitter API and *tweepy*, waits for a predefined time of two minutes, and restarts the process. The new dataset obtained at each sampling time step is merged with the previously collected dataset. This secondary data source of Tokyo 2020 tweets has 160,549 tweets and includes the information of the tweet's date, time, text, hashtags, user information, number of retweets, number of favorites, etc. All of the tweets have the hashtag ["Tokyo2020"], suggesting its relevance to the Tokyo 2020 events. The time span of the dataset is from 7/24/2021 to 7/27/2021, the first four days of Tokyo 2020. The number of tweets and long time span ensure that we have enough samples to complete the event extraction and summarization tasks with limited computing resources. We also need external data to assist with the task of event extraction. We use the technique of **web crawling** to obtain external data. Specifically, we implement customized web crawlers based on the HTML parser *BeautifulSoup* to crawl data about the list of Tokyo 2020 medalists from the Wikipedia page and the list of Tokyo 2020 sports (and events) from https://www.edudwar.com/.

3.2 Data pre-processing

We use data pre-processing techniques of NLP tasks to preprocess the text of Tokyo 2020 tweets. Specifically, we apply **noise removal, case folding, word tokenization, part-of-speech (POS) tagging, and named-entity recognition (NER)** (Nadeau and Sekine, 2007) on tweets. First, we apply a noise remover on tweets, including English stop words (e.g., "a", "the", "is", "are"), URLs, emojis, punctuation (e.g., comma, period), and special symbols (e.g., #, @, &). Second, we apply case folding to tweets to change uppercase letters to lowercase letters. Third, we use word tokenization on tweets to split the text into words. Fourth, after word tokenization, part-of-speech (POS) tagging is used to mark up a word in a text as corresponding to a particular part of speech, based on its definition and context. POS tagging is used to assign specific tokens (parts of speech) to each word. For example, if the input sentence is: "Everything to bother us", POS tagging output is: [('Everything', NN),('to', TO), ('bother', VB), ('us', PRP)]. Finally, named-entity recognition (NER) is used to locate and classify named entities mentioned in tweets into pre-defined categories such as person names, organizations, locations, countries, etc. NER tags will be used in information extraction of person names and country names in downstream tasks. For example, if we input the sentence "Jack Ma is the founder of Alibaba, a company from China", NER can identify three types of entities: "Person": Jack Ma, "Company": Alibaba, "Location": China. We decide to not add steps of **lemmatization or stemming** during the data pre-processing, because we find that they would improve the output quality of downstream tasks.

3.3 Information extraction

We then extract the four types of information about Tokyo 2020 sports events from Tweets text: (1) the name of the medalist, (2) the country the medalist belongs to, (3) the sports name, (4) the event name the medalist won, and (5) the medal type that the medalist won (i.e., gold medal, silver medal, or bronze medal). Specifically, we detect the name of medalist from tweets using two methods: **NER tagging and string matching**. For NER tagging, if the NER tag is "Person", then we extract the person name from tweets. For string matching, if there is a pre-processed person name in tweets matched with the crawled medalist name list, then we extract the person name from tweets. We detect the country the medalist belongs to by string matching. If there is a pre-processed country name in tweets matched with the pre-processed country name from the *pycountry* country list, then we extract the country name from tweets. We extract the sports name (and its event name) by string matching. If there is a pre-processed sports name (or event name) in tweets matched with the pre-processed sports

name (or event name) from the crawled list, then we extract the sports name (or event name) from tweets. Finally, we extract the medal type by keyword string matching. If there is pre-processed keywords in tweets, , then we identify the medal type. For example, if there is "gold", "champion", "1st place", or "first place" in tweets, then we identify this medalist as the gold medalist.

3.4 Text vectorization and timestamp normalization

For event summarization, **TF-IDF vectorization** (Ramos et al., 2003) (see its formula in the Appendix) is applied to all tweets which have information extracted in order to have a first layer filtering that can limit the noise in tweet classification. We found it helpful to filter out all the tweets without detecting medal type. Because it is the most necessary information for timeline generation and most accurately detected information among all the information. On the other hand, we will also include the timestamp of the filtered tweets in the event summarization. In order to do this, we normalize the timestamp of the tweets into the range [0, 1] using **min-max normalization** (see its formula in the Appendix).

3.5 Clustering

For event summarization, after all the information from tweets is transformed into a matrix, we will apply clustering to group related tweets into a single event in the timeline. **K-means clustering** (Hartigan and Wong, 1979) is the method that we use to find the clusters. We use an unsupervised learning method. Due to the nature of the timeline generation task, we cannot label all the source events manually. However, we predict how many sports events will be held in Tokyo 2020 during our sample period as the pre-specified parameter of the number of clusters k. In this case, we applied k-means to help us calculate the clusters (see its formula of calculating k-means cluster centroids in the Appendix). One advantage of our retrospective timeline generation system on the specific topic of Tokyo 2020 is that, we know the exact number of "real clusters" in k-means retrospectively.

3.6 Event label extraction

The summarized events will then be processed as event ranking. Although the clustering process can help us summarize the tweets into event clusters, we still need to extract the labels used in timeline generation. We also need to identify which means extract what the event is from the summarized events. With the similar idea of text similarity, the higher occurrence of information in the cluster means the event is highly related to that information. Unlike TF-IDF, using **count**

vectorization without tokenizing the existing label is more effective. Because we can rank the occurrence word in different labels [medalist, country, sport, event, medal type]. The word with the highest rank will be chosen in each category to become the label of the summarized event for timeline generation. On the other hand, timestamp extraction should also be related to the occurrence. However, unlike country name or medalist name in the tweets, the timestamps of the same event usually are not the same between tweets. Therefore, we will extract it by the average of all the timestamps in the summarized tweets.

3.7 Visualization

From the previous stages of event summarization and event ranking, we finally generate a data file containing attributes [medalist], [sports], [medal], [country], [datetime], and [tweets] as our outputs. To graphically demonstrate the timeline, we use the **timeline visualization** module of *matplotlib* to generate timeline charts. Due to the high density of time ticks in our result, the data is separated by the timing of the medal type. Figures **??-??** in the Appendix visualizes the timeline of the events in Tokyo 2020 (7/24/2021 to 7/27/2021) generated by our event timeline generation system.

4 Experiments

Among different clustering methods, we conduct experiments to compare OPTICS and k-means on the extracted data to see the difference in results. We choose these two clustering methods because the key parameter of OPTICS clustering is the minimum number of samples in one cluster which can be defined by how many tweets or above can become a valid event. On the other hand, the key parameter of k-means clustering is the number of clusters which can be calculated by the how many sports events in Tokyo 2020. We have compared the OPTICS with the minimum number of samples per cluster of 50 and k-means with the number of clusters of 210. Because the total valid tweet extracted is around 2,000 and around 100 sports events were played during the 4-day sample period, so for a minimum number of samples per cluster of 50, we gave some buffer because the tweets of each event must be unevenly distributed. In addition, some sports events had more than one medal received together with a buffer of unevenly distributed events. As a result, we found that k-means is more accurate in this case because it is not practical to predict the minimum number of samples to be used in OPTICS. OPTICS clustering usually classifies many tweets into un-clustered and the summarized events are too noisy for subsequent event label extraction.

5 Performance Evaluation

To evaluate the performance of our event timeline generation system, we also collect the official data from the Olympic Game official website as the verification data. Specifically, the information we collect includes [sports name], [event time], [gold: medalist], [gold: country], [silver: medalist], [silver: country], [bronze: medalist], [bronze: country]. From the official data we obtained, there are exactly 113 sports events (with at least one medal medalist is produced) happened in our sample period (23 July 2021 - 27 July 2021), and a total of 212 medals were awarded to the medalists (including 65 gold medals, 65 silver medals, and 82 bronze medals).

5.1 Qualitative evaluation

The same method of timeline visualization was applied to the official data for qualitative comparison. In this comparison, we generated eight timeline graphs for the result data and the verification data, respectively. Figures **??-??** in the Appendix visualize the timeline of the events in Tokyo 2020 (7/24/2021 to 7/27/2021) based on verification data. To compare the difference in each day, each medal type, and graph of the whole time period. For comparison of graphs in different medals (gold, silver, brown lines), we can see that compared with the verification data (roughly balanced in each day), our result has high density on gold medals and slightly low density on silver and bronze medals than the verification data. This might be an inaccurate aspect, which shows that our tweet-based system tends to extract gold medals more often than other medals. For comparison of graphs in different days (green lines: note that on graph of verification data, one record represent 1 to 3 medals while on the result graph, one record only represent 1 medal). The number of results in format [Day1, Day2, Day3, Day4] in our result is [23,61,87,39], while in verification, it is [38,49,64,60], which trend can be roughly matched. For comparison of the whole-time range (red lines), we can conclude that, for the real sports events, the awards always happened during day time, but for these twitter users, they may not post their comments immediately, so there are still many results occurred during the night time.

5.2 Quantitative evaluation

To quantitatively evaluate the effectiveness of our event timeline generation system, we check if each true medal event in Tokyo 2020 (23 July 2021 - 27 July 2021) is extracted by our tweet-based system or not. The results of performance evaluations are reported in Table 1 in the Appendix. Among 65 gold medal events, our system accurately extracts 27 of them (41.50%). Among 65 silver medal events, our system accurately extracts 6 of them (9.20%). Among

bronze medal events, our system accurately extracts 12 of them (14.60%). It is noteworthy that our system is more sensitive to gold medals but not silver or bronze medals. We propose two explanations. First, the social media users are more likely to post congratulatory tweets for gold medalists but not for silver and bronze medalists, this makes the gold medal-related clusters more easy to detect. Second, gold medalist and silver medalist always come out together (65 gold medal events and 65 silver medal events), therefore, gold medal events and silver medal events tend to come together in one tweet. It is more challenging to extract both events from one tweet, and it seems our system biased towards gold medalist events.

6 Discussions

6.1 Fragmentation and overlapping

Fragmentation in event detection refers to the situation where the same event is detected multiple times as different events. Overlapping refers to the situation where two seemingly similar events take place in almost the same time window. We should propose a more effective clustering method to make the event clusters more separable. Moreover, using k-means clustering which requires the number of clusters k to be pre-specified before clustering might not be effective when a greater variety of topics are covered in the time window. It is usually difficult to predict the total number of expected event clusters in advance (Hasan et al., 2018).

6.2 Credibility and misinformation

Misinformation and disinformation is a notable problem on social media. We should propose effective countermeasures to filter out trivial, spam, and fake news tweets. Identifying the credibility of tweets can also benefit the event ranking process. To identify the credibility of the related tweets of a specific event, tweet credibility detection and misinformation exclusion can be incorporated into the data pre-processing, or even done at the post-processing stage.

6.3 Ground truth

The availability of ground truth labels for a qualitative performance evaluation is a major advantage for our event timeline generation task for Tokyo 2020. We can conduct quantitative performance evaluations by calculating performance metrics, and compare our methodology and baseline methods in future work. For many other event topics that are not clearly defined, if the ground truth is unavailable, we cannot have quantitative or even qualitative performance evaluation.

6.4 External validity

Our event timeline generation methodology is retrospectively applied to a specific group of events, i.e., sports event results of Tokyo 2020. Therefore, we cannot evaluate our system on another dataset from different context. It is imperative that our methodology is applied and evaluated on more public datasets with greater variety of events to qualitatively and quantitatively evaluate its effectiveness and efficiency.

6.5 Scalability

Our project selectively uses the Twitter streaming data of the first four days of Tokyo 2020. However, we notice the issue of scalability of our event timeline generation system as some event detection and summarization methods are not so efficient when applied to high volume, high velocity, or high dimensional data.

6.6 Exceptions

In event extraction stage, it is noteworthy that the relevant information of the following three entities are difficult to extract: Russian Olympic Committee (ROC), United Kingdom, and Chinese Taipei. Twitter users usually call them Russia, Great Britain (or UK), and Taiwan in informal communications. These exceptions make the event extraction stage more challenging which requires improvements in future work.

References

[1] Guo, B., Ouyang, Y., Zhang, C., Zhang, J., Yu, Z., Wu, D., and Wang, Y. (2017). Crowdstory: Fine-grained event storyline generation by fusion of multi-modal crowdsourced data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–19.

[2] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):100–108.

[3] Hasan, K. T., Abdullah, S. S., Ahmed, R., and Giunchiglia, F. (2013). The history of temporal data visualization and a proposed event centric timeline visualization model. *International Journal of Computer Applications*, 70(27).

[4] Hasan, M., Orgun, M. A., and Schwitter, R. (2018). A survey on real-time event detection from the twitter data stream. *Journal of Information Science*, 44(4):443–463.

[5] Li, C., Sun, A., and Datta, A. (2012). Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 155–164.

[6] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[7] Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, volume 242, pages 29–48. New Jersey, USA.

A Appendix

A.1 Tables

Table 1: Performance evaluation of timeline generation results

Medal type	Gold	Silver	Bronze
Total no. of true medalists in Tokyo 2020 (23 July 2021 - 27 July 2021)	65	65	82
Total no. of true medalists extracted by tweet-based timeline system	27	6	12
Percent of true medalists extracted by tweet-based timeline system	41.50%	9.20%	14.60%

A.2 Figures

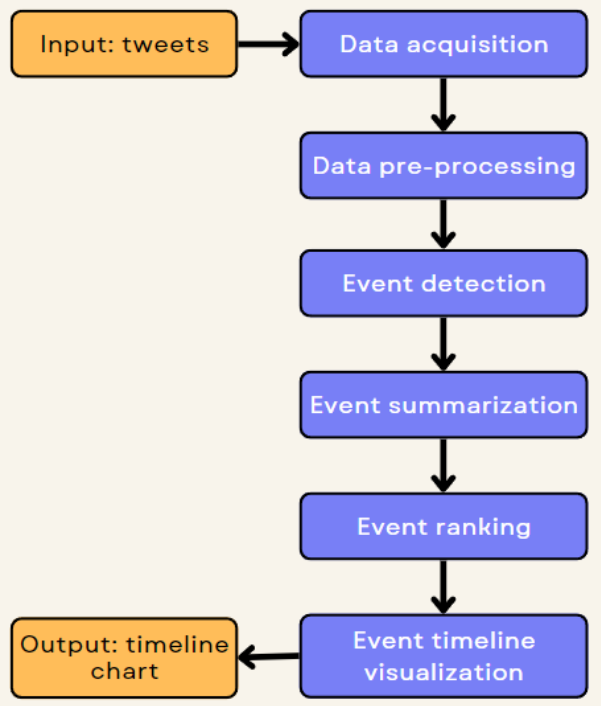

Figure 1: Tweet-based Tokyo 2020 event timeline generation system framework