Topic Detection and Tracking in Social Media Platforms

Riccardo Cantini(✉) and Fabrizio Marozzo

DIMES Department, University of Calabria, Rende, Italy
{rcantini,fmarozzo}@dimes.unical.it

Abstract. The large amount of information available on the Web can be effectively exploited in several domains, ranging from opinion mining to the analysis of human dynamics and behaviors. Specifically, it can be leveraged to keep up with the latest news around the world, although traditional keyword-based techniques make it difficult to understand what has been happening over an extended period of time. In fact, they do not provide any organization of the extracted information, which hinders the general understanding of a topic of interest. This issue can be overcome by leveraging a Topic Detection and Tracking (TDT) system, which allows detecting a set of topics of interest, following their evolution through time. This work proposes a TDT methodology, namely *length-weighted topic chain*, assessing its effectiveness over two real-world case studies, related to the 2016 United States presidential election and the Covid19 pandemic. Experimental results show the quality and meaningfulness of the identified chains, confirming the ability of our methodology to represent well the main topics underlying social media conversation as well as the relationships among them and their evolution through time.

Keywords: Topic Detection · Topic Tracking · Latent Dirichlet Allocation · Covid19 · USA Presidential Election · Social Media

1 Introduction

Every day, a huge amount of digital data are produced on the Web, effectively exploitable to extract information in several application domains, including opinion mining [5], emotion analysis [3], news gathering [18], information diffusion [2], and influence maximization [9]. For this purpose, frameworks and tools for efficient computing in distributed and high-performance infrastructures are used [4], as well as special analysis techniques, which in the case of social media are often based on hashtags, particular keywords with high semantic content [8].

Information extracted on social media can be leveraged to keep up with the latest news, even if it is difficult to understand what has been happening over an extended period of time [11]. Indeed, traditional keyword-based search techniques do not organize retrieved information, hindering the global understanding of a topic of interest. To address this problem, Topic Detection and

C. Comito and D. Talia (Eds.): PerSOM 2022, LNICST 494, pp. 41–56, 2023.
https://doi.org/10.1007/978-3-031-31469-8_3

Tracking (TDT) systems were developed, which provide automated techniques for organizing large amounts of news streams in a way that helps users quickly interpret and analyze relevant information over time [1].

This paper describes a Topic Detection and Tracking methodology, namely *length-weighted topic chain*, aimed at finding the main topics of discussion in a given corpus, tracking their evolution over time and detecting the relationships among them. It is based on the *topic chain* model proposed by Kim et al. [11], by introducing several changes and improvements. In particular, our methodology overcomes a resolution-based issue of the original approach, related to the joint use of static connection probabilities and a sliding time window of fixed size. Specifically, we removed the time window, introducing an exponential decay mechanism applied to the probability of topic connection, which is computed with respect to the length of the chain. This ensures that the connection probabilities do not go to zero instantaneously as the window size is exceeded, but they smoothly decrease as the length of the chain increases, allowing greater flexibility in the creation of the chains.

An extensive experimental evaluation was carried out on two real-world case studies, related to the 2016 United States presidential election and the Covid19 pandemic. Identified chains in both cases are quite meaningful and coherent, with no conflicting topics within the same chain. This last aspect is more evident in the first case study, characterized by the rivalry between Donald Trump and Hillary Clinton, in which almost all identified topics have a neat political polarization. The main contributions of this research are the following:

- A novel length-weighted topic chain model is proposed, aimed at finding the main topics of discussion, effectively tracking their evolution over time by discovering high-quality chains with low noise and high coherence.
- It addresses the resolution-based issue of topic chains by dynamically adjusting topic connection probability, which allows the identification of links at different probability levels, between topics at an arbitrary temporal distance.
- A precise and in-depth investigation is conducted on the main topics underlying the Twitter conversation about the 2016 US presidential election and the Covid19 pandemic.

The remainder of the paper is organized as follows. Section 2 discusses related work and the main concepts of TDT systems. Section 3 describes the topic chain model. Section 4 describes in detail the proposed methodology. Section 5 presents the case studies and Sect. 6 concludes the paper.

2 Related Work

The main objective of Topic Detection and Tracking (TDT) is to extract information about the topics of discussion and their evolution over time automatically (i.e., without human intervention), starting from flows of news in different formats (e.g., text and audio). The key concepts of TDT are the following [1]:

- *Event*: it is something that occurs at a precise time and place, triggered by a set of causes and followed by a set of consequences.
- *Activity*: represents groups of events with the same purpose, which occur at certain times and places.
- *Topic*: it is defined as a seminal event or activity, together with all closely related events and activities.
- *Story*: refers to a multimodal source of information, such as a newspaper article, radio, or television broadcast.

The main tasks of a TDT process are discussed below, together with the main approaches present in the state of the art for the implementation of such systems.

2.1 Main Tasks of a TDT Process

Story Segmentation. It represents the process by which a multimodal stream of input data is broken down into stories. The input can be either in the form of audio (*Broadcast Story Segmentation*) or text (*Text Story Segmentation*). Story segmentation represents one of the main tasks of the TDT process, as a correct splitting is crucial to identify and follow the different topics of discussion.

First Story Detection. The goal of this task is to recognize, within a flow of chronologically ordered stories, the first story that deals with a specific topic. In particular, when a new story is detected, the system decides whether it deals with a previously encountered topic or is inherent to a new topic. This task is therefore a form of online clustering, where a cluster is created if the news is not sufficiently similar to any other already seen by the system.

Topic Tracking. The goal of Topic Tracking is the identification of stories related to a specific topic, given a stream of input stories. In *Traditional Topic Tracking* a statistical approach is exploited in order to analyze the association between stories and connect them through knowledge of the specific domain. *Adaptive Topic Tracking*, instead, is a more refined technique that uses a probabilistic approach to determine the correlation between topics and stories, by progressively adapting the topic model [14].

Topic Detection. It aims to identify the topics discovered through first story detection, tracing them through topic tracking techniques. In particular, clusters composed of stories relating to the same topic are identified. The topic detection task can be divided into two sub-categories. The first *Retrospective topic detection* deals with the search for topics occurs retrospectively on a collection of stories, which are grouped into clusters on the basis of the treated topic. The second *On-line topic detection* calculates the clustering structure progressively, sequentially processing a stream of stories.

Story Link Detection. Its purpose is to determine whether two given documents deal with the same topic or not. In particular, this task is framed as a binary classification on the presence of a link between the two documents, and is based on pair-wise document similarity/divergence measures.

2.2 Main Approaches to the Realization of TDT Systems

This section describes the main techniques used to build TDT systems. The main approaches in the literature, described in this section, are based on the use of clustering techniques, semantic classes, and vector spaces. A further approach, based on topic modeling and the use of particular structures called topic chains, will be explored in Sect. 4 together with the proposed methodology.

Clustering-based. TDT techniques following this approach aim to identify a topic-based clustering structure that represents a significant grouping of the processed documents, generally represented within a vector space. In the following, we discuss the main clustering-based techniques present in the literature.

Agglomerative Hierarchical Clustering. The approach proposed by Trieschnigg et al. [15] aims to solve two typical problems of topic-based clustering structures. Firstly, a hierarchical approach allows capturing topics at different levels of granularity, obtaining different fine-grained topics in the same macro-topic. Secondly, in a hierarchical representation where document clusters are defined at different levels of granularity, stories can be assigned to multiple clusters. The relationships between clusters are expressed through an n-ary tree in which, as the depth increases, there are nodes relating to increasingly specific sub-topics, up to the leaves that represent the topics with the greatest granularity.

Single-Pass Clustering. This is a widely used approach, due to its simplicity, high efficiency, and low cost, which makes it suitable when processing a large amount of data on a large scale [13]. It is based on a clustering structure built incrementally, based on the processed documents. In particular, given a document, it is compared with all the clusters present and assigned to one of them if the similarity exceeds a certain threshold. Otherwise, it will locate within a new cluster, representative of a new topic. The main disadvantage of this approach is that the temporal distance between the document candidate to be part of a cluster and those already assigned to it is not considered in any way. This can lead to the *topic drifting* issue, which is the deviation from the original topic caused by a lowering of the purity of the identified clusters. To overcome this problem, an improved version of the Single Pass has been developed by Zhe et al. [17]. It exploits the concept of sliding time window and a double similarity threshold $(\theta_{class}, \theta_{cand})$, with $\theta_{class} > \theta_{cand}$. Specifically, if the similarity is greater than θ_{class}, the story is assigned to a specific cluster. Otherwise, a check is made on θ_{cand}, to determine weaker links to candidate topics, which are validated later within the time window. If by the end of the time window a strong similarity (greater than θ_{class}) between the story and the candidate topic is not measured, this story will constitute a new cluster.

Semantic-Based. This approach, proposed by Makkonen et al. [12], is based on the use of semantic classes, i.e. classes of terms with similar meaning (places, names, temporal expressions). In particular, the semantic content of a document is represented through the use of four classes described in the following.

- *Names*: express the subjects involved in an event.
- *Terms*: express the occurrence of an event (nouns, verbs, and adjectives).
- *Time expressions*: represent points mapped on a time axis.
- *Places*: indicate the places involved in the carrying out of an event.

To be inherent to the same event, two documents do not necessarily have to coincide in all four classes. As an example, if two documents coincide in the class of time expressions and places, they are likely discussing the same event. In addition, if we consider large geographical areas, such as continents, the similarity will be weaker than that calculated on more specific areas. Therefore, this approach allows performing class-based comparisons, through the use of three different techniques described below.

General Term Weight. This technique is based on the intuition that in short online news the event being talked about is immediately identifiable. This is exploited by introducing weights for each term calculated on the basis of its occurrences and their position within the document.

Temporal Similarity. This technique is based on the fact that news related to new events tend to be published in bursts. In particular, there is usually an initial news story followed by a group of news published shortly after. Thus, these techniques exploit the temporal information to weigh the value obtained from the calculation of similarity. Particularly, a decay factor is introduced which is proportional to the temporal distance between a news and the original one.

Spatial Similarity. This technique exploits a geographical ontology in order to measure the similarity of the spatial references present in the documents. In particular, a hierarchical structure is used, comprised of continents, regions, nations, regions and cities, which can be represented as an n-ary tree. To measure the similarity between two places, the paths that go from the root of the tree to those places are identified. Then, the ratio between the length of the common path and the sum of the total lengths of these paths is calculated.

3 A Topic Modeling Based Approach: Topic Chain

Besides the main algorithms for Topic Detection and Tracking, based on clustering and semantic classes, described in Sect. 2, there is a further approach that relies on probabilistic topic modeling [6] and a particular structure called *topic chain* [11]. A topic chain consists of a temporal organization of similar topics that appear within a specified interval of time, represented as a sliding window on a global time axis. There are different elements that make up a topic chain:

- *Long-term topic*: it consists of a general topic, present in social media conversation or online news over a long period of time, such as the discussion about Covid19 pandemic or the war in Afghanistan.
- *Temporary issue*: it is a specific topic, that is talked about for a short period of time. It can be related to sporadic events or a part of a broader topic. An example can be a manifestation, which is generally a sporadic event, often connected to a long-term topic or to a specific trend on social media.
- *Focus shift*: it is the change over time of the particular aspect of a long-term topic on which news about that topic is focused. As an example contagion-prevention rules and vaccination within the general topic of Covid19.

The methodology is based on the creation of topic chains, aimed at understanding how topics and issues emerge, evolve, and disappear within the analyzed news corpus. This is achieved by observing long- and short-term topics, along with the different topic shifts occurring in the corpus. In particular, the methodology is comprised of three main steps, described in the following.

3.1 Topic Discovery

In this step, the analyzed corpus of news is divided into several time slices and the main topics in each time slide are found. The topic discovery step is performed by using the Latent Dirichlet Allocation (LDA) [7], a widely used algorithm for probabilistic topic modeling. LDA models each document as a random mixture of latent topics, while each topic is a distribution of terms over a fixed-sized vocabulary. Specifically, given K latent topics underlying a corpus composed of M documents of N words each, the generative process works as follows:

- For each document $d_i, i = 1, \ldots, M$, a multinomial distribution θ_i over the K latent topics is randomly sampled from a Dirichlet distribution with parameter α.
- For each topic $z_k, k = 1, \ldots, K$, a multinomial distribution ϕ_k over the N words is randomly sampled from a Dirichlet distribution with parameter β.
- For each word position $j = 1, \ldots, N$ of the document d_i, a topic $z_{i,j}$ is sampled from θ_i.
- The j-th word of d_i (i.e. $w_{i,j}$) is generated by random sampling from the multinomial distribution $\phi_{z_{i,j}}$.

Summing up, the total probability of the model is obtained as:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{k=1}^{K} P(\phi_k; \beta) \prod_{i=1}^{M} P(\theta_i; \alpha) \prod_{j=1}^{N} P(z_{i,j}|\theta_i) P(w_{i,j}|\phi_{z_{i,j}}) \quad (1)$$

The various distributions, such as the specific mixture of each document in the corpus, are not known a priori and have to be learned via statistical inference. In particular, different approaches were proposed to deal with this task, by approximating the posterior distribution, such as Variational Bayes [7] and Monte Carlo Markov Chain (MCMC) algorithms like Gibbs Sampling [10].

3.2 Choice of the Topic Similarity Measure

A topic can be treated as a multinomial distribution over the words of the corpus vocabulary, as a ranked list of words, or as a vector in which each word of the vocabulary is associated with that topic with a certain probability. This allows the use of several metrics to compute the similarity between topics, introduced in the following.

- *Cosine similarity*: it measures the similarity between two given n-dimensional vectors as the cosine of the angle between them.
- *Jaccard coefficient*: it defines the similarity of two sets of items as the cardinality of the intersection divided by the cardinality of the union.
- *Kendall's τ coefficient*: it is a non-parametric measure of the rank correlation between two sets of items.
- *Discounted cumulative gain (DCG)*: it measures the overall normalized relevance of a set of ranked items by penalizing highly relevant documents appearing at a low position in the ranking.
- *Kullback-Leibler divergence (KL)*: it measures the dissimilarity, in probabilistic terms, between two given distributions.
- *Jensen-Shannon divergence (JS)*: it is the symmetric version of the KL divergence between two given distributions, obtained as the average divergence from their mixture distribution.

The choice of the most suitable topic similarity measure is done by finding the measure that identifies the best associations between topics of two consecutive time windows. In particular, the LDA model is trained for the time windows t and $t+1$, finding two distinct sets of topics $\phi^t = \phi_1^t, \ldots, \phi_k^t$ and $\phi^{t+1} = \phi_1^{t+1}, \ldots, \phi_k^{t+1}$. Then the topic-wise similarity is computed for each pair ϕ_i^t, ϕ_j^{t+1}, finding the top five most similar pairs. Finally, for each pair in the top five, the topic ϕ_i^t is substituted to ϕ_j^{t+1} and the log-likelihood of the data at time t is computed. This process is repeated for each similarity measure, by finding the one that minimizes the negative log-likelihood. This measure will be the selected one, as it leads to the most meaningful substitutions thus obtaining the set of topics that best explain the corpus.

3.3 Topic Chains Construction

In this step, topic chains are built by identifying sequences of similar topics through time. The corpus is divided into a sequence of time slices by using a sliding window and the similarity between topics is computed by leveraging the similarity measure identified in the previous step. In particular, denoted as $\phi^t = \phi_1^t, \ldots, \phi_k^t$ the topic distribution at time t, the construction process proceeds as follows:

1. The topic distribution at the previous time $t-1$, i.e. $\phi^{t-1} = \phi_1^{t-1}, \ldots, \phi_k^{t-1}$ is found and the similarity between ϕ_i^t and each topic ϕ_j^{t-1} is computed.

2. For each pair ϕ_i^t, ϕ_j^{t-1} such that their similarity is greater than a threshold, a link between them is created and the process moves to the next topic ϕ_{i+1}^t.
3. If no link was created by comparing ϕ_i^t with the topics ϕ^{t-1}, a comparison is made with topics in ϕ^{t-2}.
4. This process is iterated backward, until at least one connection is found or the size of the time window is exceeded.

Note that if a divergence measure is directly used (see step 2), such as KL or JS divergence, the measured value between the two considered topics must be less than the threshold for a link to be created.

3.4 Chains Analysis and Interpretation

As a last step, the obtained chains are analyzed in order to find long-term topics such as politics and economics, and temporary issues related to specific events and topics that do not last for a long period of time. Furthermore, the different focus shifts in long-term topics are identified, by analyzing the use of named entities along the chain. Specifically, a named entity is a real-world object, such as a person, location, organization, or product, characterized by a proper name.

4 Proposed Methodology

The topic chain methodology, described in Sect. 3, is characterized by a tuning phase in which the value of the threshold and the dimension of the sliding time window are selected against a wide set of possible values. However, the joint optimization of these hyper-parameters can lead to some issues. In the following we provide an in-depth description of these issues, together with the weighted variant we propose to overcome them.

4.1 Main Limitations of the Original Approach

Let's assume that a divergence measure is used for computing topic similarity so that a connection between topics is created when the divergence is below the threshold. By increasing the threshold value, the connection probability also increases, as this allows the association of topics with greater divergence. The increase in connection probability generally leads to the creation of longer chains, but can also negatively impact the meaningfulness of the obtained chains. In particular, as the size of the chain increases, it can incorporate other small chains and singlet topics, with the risk of introducing noise, i.e. non relevant or even contradictory links. This negative effect, which causes a general decrease of coherence, can be partially avoided by finding a suitable size for the sliding time window. Specifically, this parameter limits the number of possible links that can be created, by allowing the comparison only with a small number of previous time slices. However, the main limitation of this approach is that a time window of fixed size does not allow the emergence of links between time slices

at a distance greater than that imposed by the window size. In fact, in order to identify distant connections it is necessary to increase the window size, which again can lead to the presence of noisy connections. Another way to reduce the presence of noisy links could be lowering the threshold value, but this would lead to the loss of chains with a lower connection probability.

By summing up, the main limitation of the original methodology is that it is not able to detect links at different probability levels and at an arbitrary distance in time, due to the joint action of the threshold and window size. In particular:

- Connections between time slices distant from each other need a wide time window. Consequently, they may not be isolated but included in broader chains along with other noisy connections.
- Lowering the threshold to remove noisy connections can cause the loss of weaker links, which does not allow to find chains at a lower probability level.

This is a resolution-based issue that is also present in other application domains. As an example, a similar drawback characterizes the DBSCAN clustering algorithm, which is not able to detect a global structure composed of clusters at different density levels.

4.2 Proposed Solution: Length-Weighted Topic Chain

In order to overcome the issues discussed in Sect. 4.1, we propose a variant of the topic chain methodology, namely *length-weighted topic chain*, which introduces an exponential decay of connection probabilities. Exponential decay is a widely used mechanism, exploited in several application domains, especially in modeling natural phenomena, such as radioactive decay, the variation of atmospheric pressure, and enzyme-catalyzed reactions. For what concerns TDT techniques in the literature, an example is provided by Xu et al. [16], that used an exponential decay in computing the similarity between two topics, which decreases as the temporal distance between them increases.

In our solution, the decay is computed with respect to the length of the topic chain. In particular, we removed the limitations imposed by the time window size, potentially allowing connections between topics on the whole temporal axis. In this way, the connection probability does not go to zero instantaneously, when the fixed size of the window is exceeded, but decreases smoothly as the length of the chain increases. This effect is obtained by dynamically modifying the threshold in relation to the current length of the chain. This threshold specifies the cutting value for topic divergence: specifically, two topics are linked to each other within the chain if their divergence is not greater than the threshold. The exponential decrease of the threshold is controlled by the decay factor λ, which also affects the length of the chain: larger values of this constant cause a more rapid decrease of the threshold, which results in lower connection probabilities and shorter chains. Formally, let $\phi_i^t, \phi_j^{t'}$ be a pair of topics detected in two different time slices t and t' with $t > t'$, and let th_0 be the initial value of the threshold used to test if a link can be created between two given topics. This threshold undergoes an

exponential decay based on the current length of the chain L, defined as the number of links present in the chain up to $\phi_j^{t'}$. Therefore, the current value of the threshold is computed as follows:

$$th_L = th_0 \cdot e^{-\lambda L} \tag{2}$$

Afterwards, a link between ϕ_i^t and $\phi_j^{t'}$ is created within the topic chain if:

$$div(\phi_i^t, \phi_j^{t'}) \leq th_L \tag{3}$$

The introduction of the exponential decay allows to overcome the issues of the original methodology, which is not able to detect connections at different probability levels between topics at an arbitrary temporal distance. Indeed, by eliminating the time window, we are able to connect topics even if they belong to time slices distant from each other, by controlling at the same time the length of the chain through the decrease of the connection probabilities. This allows the formation of links between topics located at any point within the global time axis, and avoids the introduction of noise in chains of excessive length. Specifically, as the length of the chain increases, topics in subsequent time slices encounter greater resistance in forming a connection, which leads to the continuation of the chain only if the link is significant enough to overcome this resistance. Otherwise, the process iterates backward trying to link that topic to another in an earlier time slice. In that case, the topic will be connected to a shorter chain which is thus forked into a new, separate one.

5 Case Studies

In this section, we discuss the extensive experimental evaluation carried out by applying the proposed methodology to two different case studies, concerning social media conversation on the Twitter platform. Specifically, the first case study relates to the 2016 US presidential election, characterized by the rivalry between Hillary Clinton and Donald Trump, while the second focuses on content posted by users during the Covid19 pandemic. In the following, we provide a detailed description of the hyper-parameter tuning phase, together with an in-depth analysis of the identified chains. In our experiment, each time slice coincides with an entire day, and we run the LDA algorithm with a number of latent topics to be discovered equal to 10. In addition, the discovered chains will be analyzed at two different levels of granularity, to better grasp the connections between discussion topics and the daily evolution of social media conversation:

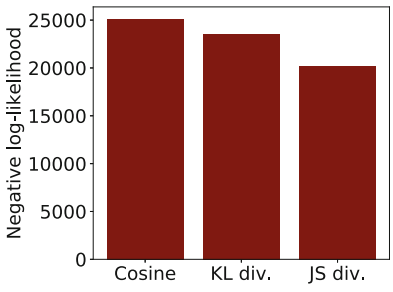
- *Topic-level*: $\{\phi_i^t, \phi_j^{t'}, \ldots, \phi_z^{t''}\}$, i.e., fine-grained chains identified by the connection between topics, in which there exists a link between each topic and its predecessor in the chain.
- *Day-level*: $\{t, t', \ldots, t''\}$, i.e., coarse-grained chains in which two days t and t' are connected if there exists a link between two topics ϕ_i^t and $\phi_j^{t'}$ detected in those days.

5.1 The 2016 US Presidential Election

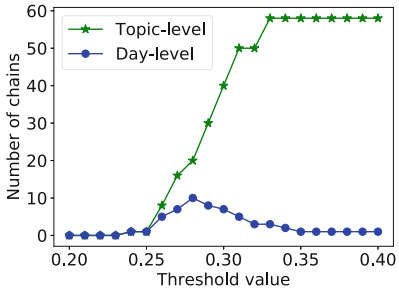
This corpus comprises about 2.5 million tweets, posted by 521, 291 users regarding the 2016 US elections, published from October 10 2016 to November 7 2016. The analysis has been performed on tweets published in the main US swing states (Colorado, Iowa, Florida, Ohio, Michigan, Pennsylvania, Wisconsin, New Hampshire, North Carolina, and Virginia), characterized by high political uncertainty. In this way we obtained a representative corpus balanced with respect to the political polarization of the contained posts (*pro-Clinton* or *pro-Trump*). Tweets were collected through the public Twitter API, by using a set of keywords related to the presidential election, such as *#votetrump*, *#maga*, *#voteblue*, and *#USAelection2016*. Then, collected tweets were preprocessed in order to make them suitable for the subsequent analysis steps, as described below.

- We lowercased the text of each tweet, filtering out URLs, emojis, punctuation, and stopwords.
- We normalized each word by replacing accented characters with regular ones and by performing lemmatization.
- Tweets in a language other than English were filtered out.
- The most frequent bigrams in the corpus were found and collapsed in a single word, such as *hillary_clinton*, *donald_trump*. *bernie_facts*.

Hyper-parameter Tuning. Here we describe how the different hyper-parameters needed by the methodology were tuned. Firstly, we determined the most suitable topic similarity/divergence measure. In particular, we followed the approach described in Sect. 3.2. As shown in Fig. 1(a), we found that the best measure, which minimizes negative log-likelihood, is the Jansen-Shannon divergence. Therefore, we used this measure throughout the whole experimental evaluation.



(a) Choice of the best similarity/divergence measure.



(b) Choice of the best threshold (cut value for JS divergence).

Fig. 1. Tuning of the main hyper-parameters of the methodology.

Once the divergence measure was chosen, we focused on the tuning of its cut value, i.e. the threshold used to test if two topics can be connected to create a

new link within the chain. For this purpose, we plotted the number of topic-level and day-level chains, varying the threshold, as shown in Fig. 1(b). Two limit cases can be identified:

- If the cut value of the JS divergence is too low (below 0.24) no chain is detected, as this threshold results to be too strict.
- if the threshold is too high (above 0.35), all connections are merged in a single global day-level chain, as almost all subsequent days are connected by at least a pair of topics.

In order to ensure a trade-off between the number of chains at day and topic level, we selected a threshold value $th_0 = 0.28$. Indeed, this value shows a good ability in discriminating different trends evolving in subsequent days, causing at the same time the formation of a reasonable number of topic-level chains.

Discovered Chains. In the following we describe the most relevant chains identified by the proposed methodology, analyzing also the effects of the introduction of the exponential decay with a factor $\lambda = 0.05$. For a better understanding, chains are reported at day-level together with the general topic of discussion. Then the different connections between fine-grained topics are investigated, by providing various example tweets. Specifically, we found out what follows.

- ***Sexism:*** this chain connects 11, 12, 13, 15, and 16 October, and is characterized by a series of criticisms leveled at Donald Trump and its supporters. In particular, Trump was criticized for his sexually aggressive comments, which he justified by defining them *locker room talk*. This topic is characterized by the words *locker*, *room*, *talk*, and by the hashtag *#nevertrump*, which confirms the anti-trump polarity of the topic, which is constant throughout the entire chain. The chain continues with another anti-trump topic linked to sexism. This second topic is related to the tweets published by Trump's supporters, favorable to the repeal of amendment 19, which grants women the right to vote. As an example: *"Women are not fit for politics. #RepealThe19th"*. The publication of such content generated a lot of criticisms on Twitter, aliment-ing the anti-Trump discussion on the social platform: *"For anyone who thinks sexism doesn't exist and fighting for women's equality doesn't matter anymore: #repealthe19th is an actual hashtag"*. The main words and hashtags characterizing this topic are *women*, *#repealThe19*, *#nevertrump*, and *#imwithher*, a pro-Clinton hashtag. This chain continues with the same topic about sexism, identified in the following days by the words *women*, *inappropriate*, *predator*, and *#frankentrump*, a hashtag through which Trump was compared to Mary Shelley's Frankenstein in a derogatory way.
- ***Disputes over the Clintons and elitism:*** this chain covers the 17, 18, 22, 23, and 24 October, and is characterized by the discussion on Twitter about a series of disputes related to Hillary Clinton and her husband, the ex-president Bill Clinton. The first controversy is about Hillary Clinton's six-years tenure as a director of Walmart, and is characterized by words like *walmart*, *board*,

#corrupt, and *#podestamails*. Indeed, several emails were stolen from Hillary Clinton campaign chairman John Podesta's mail account, that document a close relationship between Clinton and Walmart. The chain continues with another topic about *elitism*, in which Hillary Clinton is accused of being supported by the American elite, pursuing the interests of a small circle of influential people. This topic is identified by the negative hashtag *#neverhillary* and the word *elite*, also used as a hashtag in tweets like the following: *"No doubt she has already been crowned queen by the US #elite"*. The last topic in this anti-Clinton chain relates to the controversies concerning the connection between Bill Clinton and Jeffrey Epstein, a millionaire accused of sexual abuse and child trafficking. In particular, conversation on social media focuses on Epstein's private island commonly referred to as Pedophile Island, and on the accusations made against the Clintons of having visited that place. Therefore, this topic is characterized by the words *pedophile*, *island*, *#lockherup*, and *#draintheswamp*, which are pro-Trump hashtags.

- ***Trump's rhetoric:*** this is a short chain linking 26 and 27 October, characterized by an anti-Trump topic. In particular, the republican candidate was criticized for his rhetoric, often considered violent, homophobic, and racist. Therefore, this topic is identified by the words *rhetoric*, *violent*, *trump*, and *#voteblue*, a hashtag in favor of Hillary Clinton.
- ***Support from prominent public figures for Hillary Clinton:*** this chain covers 28, 29 October, and 1, 2 November, days in which social media conversation focused on the support for the democratic candidate from public figures. As an example, Michelle Obama supported Hillary Clinton's candidacy by speaking at the rally held by Clinton on October 27 in Winston-Salem, North Carolina. Words and hashtags referring to this event are *women*, *rally*, *#imwithher*, and *#strongertogether*. Clinton also had the support of senator Jeanne Shaheen (*#senatorshaheen*, *#imwithher*) and the billionaire Richard Branson. In particular, the words *richard*, *branson*, *quote*, *trump*, refers to an interview released by Brandson in which the entrepreneur criticized Trump's violent temper, defining him as irrational and aggressive.
- ***US elections and propaganda:*** this is a short chain, covering 3 and 4 November, in which both pro-Clinton and pro-Trump supporters published content in favor of the two main candidates. Tweets are characterized by the main faction hashtags, such as *#maga* and *#votehillary*, and by hashtags encouraging people to vote, like *#vote*, *#vote2016* *#election2016*.

It is worth noting that the identified chains are quite meaningful and coherent from the viewpoint of political polarization. They also represent well the main topics underlying social media conversation, as well as the relationships among them and their evolution through time. Furthermore, we achieved these results thanks to the introduction of the exponential decay mechanism. Indeed, by using the traditional methodology, with a sliding time window of fixed size and a constant value for the threshold, we observed a degradation in the quality of the detected chains. As an example, the first chain about sexism is merged with the first part of the second chain, about Clinton and Walmart. Due to this, the first

chain, characterized by anti-Trump themes, is polluted by a topic against Hillary Clinton. This introduces noise into the results due to an inversion of political polarization within the same chain. One way to avoid these noisy links could be to lower the cut value for JS divergence, but this would result in the loss of other links and small chains, such as the one referring to Trump's violent rhetoric.

5.2 Coronavirus Pandemic (Covid19)

This case study analyzes the tweets published in December 2020 related to the Covid19 pandemic. By applying the proposed methodology, we discovered five different chains, whose macro-topic is Covid19. These chains, which represent some of the aspects that the conversation on social media has focused on most, span the entire month under consideration, covering almost every day. This means that the identified chains do not present clear boundaries, contrary to those described in the previous case study. This is because these chains are not connected to specific events, but deal with the main topics on which the general discussion related to the pandemic is focused. Specifically, we found what follows:

- ***General conversation about Covid19:*** this chain connects topics that refer to Covid19 in a generic way, characterized by words like *global* and *covid*, and hashtags like *#covid19* and *#coronaviruspandemic*.
- ***Anti-contagion protocols:*** this chain connects a series of topics about the different protocols for contagion prevention, identified by trending hashtags on Twitter, such as *#washyourhands*, *#socialdistancing*, and *#wearamask*.
- ***Remote job:*** in this chain, the advantages of remote job are discussed, causing the presence of topics identified by words like *work* and *job* and hashtags like *#workfromhome* and *#remotejob*.
- ***Vaccination and medical personnel:*** this chain is characterized by published content regarding Covid19 vaccines, a topic identified by words and hashtags like *vaccine*, *#vaccine*, *#covidvaccine*. In addition, other hashtags like *#healthcare* and *#frontlineheroes* refer to healthcare workers and their vital contributions during the pandemic.
- ***Christmas:*** in this chain, the discussion on social media was about the effect of the pandemic on how people spent the Christmas holidays. The main words and hashtags are *christmas*, *#covidchristmas*, and *#christmas2020*.

6 Conclusions and Final Remarks

This paper describes a Topic Detection and Tracking methodology, namely *length-weighted topic chain*, aimed at finding the main topic of discussion in a given corpus, tracking their evolution over time, and detecting the relationships among them. The proposed methodology is based on the *topic chain* model and introduces an exponential decay mechanism applied to the probability of topic connection, which is computed with respect to the length of the topic chain. In this way, the main limitations of the original topic chain model can be overcome,

allowing the identification of links at different probability levels, between topics located at any point within the global time axis, as well as an overall reduction of noise in the discovered chains.

The effectiveness of the proposed methodology was assessed over two real-world case studies, related to the 2016 United States presidential election and the Covid19 pandemic. Achieved results confirm the quality and meaningfulness of the identified chains, which represent well the main topics underlying social media conversation as well as their temporal evolution. Detected chains are also quite coherent, with no conflicting topics within the same chain, which is desirable in the case of politically-oriented news.

References

1. Allan, J.: Topic Detection and Tracking: Event-Based Information Organization, vol. 12. Springer, Heidelberg (2002). https://doi.org/10.1007/978-1-4615-0933-2
2. Arnaboldi, V., Contia, M.: Passarella, A., Dunbar, R.: Online social networks and information diffusion: the role of ego networks (2017). Preprint submitted to Elsevier, 8 November 2017
3. Belcastro, L., Branda, F., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: Analyzing voter behavior on social media during the 2020 us presidential election campaign. Soc. Netw. Anal. Min. **12**(1), 1–16 (2022)
4. Belcastro, L., Cantini, R., Marozzo, F., Orsino, A., Talia, D., Trunfio, P.: Programming big data analysis: principles and solutions. J. Big Data **9**(1), 1–50 (2022). https://doi.org/10.1186/s40537-021-00555-2
5. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: Learning political polarization on social media using neural networks. IEEE Access **8**, 47177–47187 (2020)
6. Blei, D.M., Lafferty, J.: Topic Models. Text Mining: Theory and Applications (2009)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
8. Cantini, R., Marozzo, F., Bruno, G., Trunfio, P.: Learning sentence-to-hashtags semantic mapping for hashtag recommendation on microblogs. ACM Trans. Knowl. Discov. Data (TKDD) **16**(2), 1–26 (2021)
9. Cantini, R., Marozzo, F., Mazza, S., Talia, D., Trunfio, P.: A weighted artificial bee colony algorithm for influence maximization. Online Soc. Netw. Media **26**, 100167 (2021)
10. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(suppl_1), 5228–5235 (2004)
11. Kim, D., Oh, A.: Topic chains for understanding a news corpus. In: Gelbukh, A. (ed.) CICLing 2011. LNCS, vol. 6609, pp. 163–176. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19437-5_13
12. Makkonen, J., et al.: Semantic classes in topic detection and tracking (2009)
13. Mohd, M., Crestani, F., Ruthven, I.: Construction of topics and clusters in topic detection and tracking tasks. In: 2011 International Conference on Semantic Technology and Information Retrieval, pp. 171–174. IEEE (2011)
14. Ren, X., Zhang, Y., Xue, X.: Adaptive topic tracking technique based on k-modes clustering. Comput. Eng. **35**(9), 222–224 (2009)

15. Trieschnigg, D., Kraaij, W.: TNO hierarchical topic detection report at TDT 2004. In: Topic Detection and Tracking Workshop Report (2004)
16. Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., Yao, H.: Research on topic detection and tracking for online news texts. IEEE Access **7**, 58407–58418 (2019)
17. Zhe, G., Zhe, J., Shoushan, L., Bin, T., Xinxin, N., Yang, X.: An adaptive topic tracking approach based on single-pass clustering with sliding time window. In: Proceedings of 2011 International Conference on Computer Science and Network Technology, vol. 2, pp. 1311–1314. IEEE (2011)
18. Zubiaga, A.: Mining social media for newsgathering: a review. Online Soc. Netw. Media **13**, 100049 (2019)