International Journal on Digital Libraries
https://doi.org/10.1007/s00799-018-0257-7

Tracking the history and evolution of entities: entity-centric temporal analysis of large social media archives

Pavlos Fafalios[1] · **Vasileios Iosifidis**[1] · **Kostas Stefanidis**[2] · **Eirini Ntoutsi**[1]

Received: 7 February 2018 / Revised: 6 September 2018 / Accepted: 13 October 2018

Abstract
How did the popularity of the Greek Prime Minister evolve in 2015? How did the predominant sentiment about him vary during that period? Were there any controversial sub-periods? What other entities were related to him during these periods? To answer these questions, one needs to analyze archived documents and data about the query entities, such as old news articles or social media archives. In particular, user-generated content posted in social networks, like Twitter and Facebook, can be seen as a comprehensive documentation of our society, and thus, meaningful analysis methods over such archived data are of immense value for sociologists, historians, and other interested parties who want to study the history and evolution of entities and events. To this end, in this paper we propose an *entity-centric* approach to analyze social media archives and we define measures that allow studying how entities were reflected in social media in different time periods and under different aspects, like popularity, attitude, controversiality, and connectedness with other entities. A case study using a large Twitter archive of 4 years illustrates the insights that can be gained by such an entity-centric and multi-aspect analysis.

Keywords Social media archives · Entity analytics · Entity linking · Sentiment analysis

1 Introduction

Social networking services have now emerged as central media to discuss and comment on breaking news and noteworthy events that are happening around the world. In Twitter, for example, every second around 6000 tweets are posted, which corresponds to over 350,000 tweets per minute, 500 million tweets per day, and around 200 billion tweets per year.[1]

Such large amount of user-generated content produced continuously in social media is considered of immense historical value for future generations [6]. However, although there are initiatives that aim to collect and preserve social media archives, like the Twitter Archive at the Library of Congress [46], the absence of meaningful access and analysis methods still remains a major hurdle in the way of turning such archives into useful sources of information for historians, journalists, and other interested parties [6].

When exploring archived data, analysts are not interested in the documents per se, but instead they want to see, compare, and understand the behavior of (and trends about) entities, like companies, products, politicians, athletes, celebrities, or music bands, thus calling for *entity-level* analytics over the archived data [41].

In this paper, we propose an *entity-centric* approach to analyze social media archives. Our approach allows tracking of how entities are reflected in a collection of user-generated content (e.g., tweets) in different time periods and how such information evolves over time and also with respect to other entities. Specifically, we propose a multi-aspect description of an entity in terms of its *popularity* (how much discussion it generates), *attitude* (predominant sentiment toward the entity), *sentimentality* (magnitude of sentiment toward the

[1] http://www.internetlivestats.com/twitter-statistics/ (August 30, 2018).

✉ Pavlos Fafalios
fafalios@l3s.de

Vasileios Iosifidis
iosifidis@l3s.de

Kostas Stefanidis
kostas.stefanidis@uta.fi

Eirini Ntoutsi
ntoutsi@l3s.de

[1] L3S Research Center, University of Hannover, Hannover, Germany

[2] Faculty of Natural Sciences, University of Tampere, Tampere, Finland

entity), *controversiality* (whether there is a consensus about the sentiment toward the entity), *connectedness* to another entity (how strong is its connection to another entity), and *network* (strongly connected entities). We propose measures that capture all these aspects in a given time period (e.g., day, week, or month). A distinctive characteristic of our approach is that it does not rely on service-specific labels (like #hashtags and @mentions), rather it exploits *entity linking* [37] and thus can be applied over any type of time-annotated texts.

We examine the insights gained by the proposed measures on a large collection of billions of tweets spanning a period of 4 years (Jan 2013–Jan 2017). Such analytics enable to answer questions like the following:

- How did the popularity of Greek Prime Minister, Alexis Tsipras, evolve in 2015? Were there any "outlier" periods, i.e., periods of extremely high or low popularity? What were the entities discussed in social media together with Alexis Tsipras during these periods? How did the "connectedness" of Alexis Tsipras with Vladimir Putin evolve in 2015?
- How did the predominant sentiment about Donald Trump and Hillary Clinton vary during 2016? Were there any "controversial" time periods related to these two politicians, i.e., time periods in which there were many positive and negative tweets? What other entities were discussed together with Donald Trump and Hillary Clinton in tweets with predominant positive or negative sentiment?

In a nutshell, this paper makes the following contributions:

- We propose a set of measures for capturing important entity features in a given time period. A sequence of such captures comprises a multi-variate time series in which each point is a multi-aspect description of the entity at a certain time period. We demonstrate the usefulness of our approach through illustrative examples.
- We provide an open-source Apache Spark library for computing the proposed measures efficiently.
- We analyze a large Twitter archive spanning 4 years and containing billions of tweets and make publicly available the entity and sentiment annotations of this archive. This dataset can foster further research in related topics like topic evolution, entity recommendation, and concept drift.

This paper is an extension of [13]. The major changes include: (i) an extensive survey of the related literature, (ii) a new family of time-related measures (*Entity-Time Measures*), (iii) an extension of the entity-relation measures with new measures for identifying the networks with positive or negative sentiment of a given entity (*Positive and Negative k-Networks*), and (iv) an extension of the case study with results related to the new measures.

The rest of this paper is organized as follows: Sect. 2 provides the required background and related works. Section 3 motivates and introduces the proposed measures. Section 4 describes a library for the distributed computation of the measures. Section 5 presents the results of a case study. Finally, Sect. 6 concludes the paper and identifies interesting directions for future reasearch.

2 Background and related literature

2.1 Entity and sentiment annotations

Our analysis is based on two different types of annotations applied in short texts from social media archives: *entity linking* and *sentiment analysis*.

2.1.1 Entities and entity linking

Following Chen's definition [11], an entity is *"a thing which can be distinctly identified"*. In our problem, an entity has a Web identity expressed through a unique URI [19]. This does not only include persons, locations, organizations, etc., but also events (e.g., *US 2016 presidential election*) and general concepts (e.g., *democracy* or *abortion*). A knowledge base contains information about a set of entities, like properties or relations with other entities. This information is described using one or more ontologies/vocabularies [8]. DBpedia, for instance, is a cross-domain knowledge base derived from Wikipedia that makes use of the DBpedia Ontology for describing information about its entities [22].

Entity linking is the task of automatically identifying entity mentions in a piece of text and resolving them to their corresponding entries in a reference knowledge base [37]. For example, given the text *"Obama visited Cuba"* and the reference knowledge base DBpedia, an effective entity linking system should link the text "Obama" to the former USA president Barack Obama (http://dbpedia.org/resource/Barack_Obama), and the text "Cuba" to the country Cuba (http://dbpedia.org/resource/Cuba). For each annotation, an entity linking system also provides a *confidence score* representing the confidence that the corresponding mention has been correctly disambiguated. The survey by Shen et al. [37] presents a thorough overview and analysis of the main approaches to entity linking and discusses various applications as well as the evaluation of entity linking systems. In our case studies, we used the system Yahoo FEL [4] which has been specially designed for linking entities from short texts to DBpedia/Wikipedia.

2.1.2 Sentiment analysis

Sentiment analysis refers to the problem of assigning a sentiment label (e.g., positive, negative) or sentiment score to a document [27]. We opt for the latest and we use SentiStrength, a robust tool for sentiment strength detection on social web data [39]. SentiStrength assigns both a positive and a negative score (since both types of sentiment can occur simultaneously in a text). The strength score of a positive sentiment ranges from +1 (not positive) to +5 (extremely positive). Similarly, negative sentiment strength scores range from −1 (not negative) to −5 (extremely negative). For example, given the text *"I love you but hate the current political climate"*, SentiStrength provides the positive sentiment score +3 and the negative sentiment score −4.

In Sect. 5.1, we report evaluation results regarding the accuracy of Yahoo FEL and SentiStrength.

2.2 Related works

The availability of web-based application programming interfaces (APIs) provided by popular social media services like Twitter and Facebook has led to an "explosion" of techniques, tools, and platforms for social media analytics. Batrinca and Treleaven [3] survey analytics tools for social media as well as tools for scraping, data cleaning, and sentiment analysis on social media data. There is also a plethora of works on exploiting social media for a variety of tasks, like opinion summarization [24], event and rumor detection [16,28], topic popularity and summarization [2,42], information diffusion [18], popularity prediction [34], and reputation monitoring [1]. Furthermore, social media is exploited by research communities for research and experimentation in a variety of research problems. Examples include the *Making Sense of Microposts* series of workshops [30,31], or the *Sentiment Analysis in Twitter* tasks of the International Workshop on Semantic Evaluation [26,32]. Below, we describe works related to the *temporal* analysis of topics and entities in social media.

Stefanidis and Koloniari [38] propose a query-answering framework to allow entity search in social networks by exploiting the underlying social graph and temporal information. To satisfy the varying search needs, the framework includes a time-aware query model and a corresponding logical algebra. To deal with the temporal aspect, the authors adopt an annotated graph model that incorporates time by associating each element in the graph with its temporal information. The nodes (representing users and objects) and the edges between them (representing social relationships) have a label that indicates their valid time. The proposed query model allows for time-dependent queries that exploit time explicitly by using it as a hard constraint to filter out irrelevant results.

Ardon et al. [2] perform a spatiotemporal analysis of tweets, investigating the time-evolving properties of the subgraphs formed by the users discussing each topic. The focus is on the network topology formed by follower-following links on Twitter and the geospatial location of the users. The authors investigate the effect of initiators on the popularity of topics and find that users with a high number of followers have a strong impact on popularity. They also showed that topics become popular when disjoint clusters of users discussing them begin to merge and form one giant component that grows to cover a significant fraction of the network.

Bruns and Stieglitz [5] introduce a catalogue of metrics for analyzing hashtag-based communication on Twitter, with particular focus on hashtagged Twitter conversations. The proposed metrics can be categorized into: metrics that examine the total activity and visibility of individual participants, metrics that establish the temporal flow of conversation and of specific forms of conversation, and metrics that combine the activity of the users and the flow of conversations to examine the relative contributions of specific user groups in different time points.

Saleiro and Soares [34] tackle the problem of predicting entity popularity on Twitter based on the news cycle. The authors apply a supervised learning approach and extract four types of features (signal, textual, sentiment, and semantic) which are used to predict whether the popularity of a given entity will be high or low in the following hours. The results of an experimental evaluation showed that news performs better on predicting entity popularity on Twitter when it is the primary information source of the event, in opposition to events such as live TV broadcasts, political debates, or football matches.

Celik et al. [7] investigate whether semantic relationships between entities can be learned by analyzing microblog posts published on Twitter. The authors developed a relation discovery framework that allows for the detection of typed relations that may have temporal dynamics. The evaluation results showed that co-occurrence-based strategies allow for high precision and perform particularly well for relations between persons and events. Our entity-to-entity connectedness scores are also based on entity co-occurrences (more in Sect. 3). The authors also analyzed the performance in learning relationships that are valid only for a certain time period and revealed that Twitter is a suitable source for this type of relationships because it allows the discovery of trending topics with high accuracy and low delay.

Ren et al. [29] consider the task of time-aware tweets summarization exploiting user's history and collaborative social influences from social circles. The authors propose a time-aware user behavior model, called Tweet Propagation Model, in which dynamic probabilistic distributions over interests and topics are inferred. In the same context, Zhao et al. [45] study how to incorporate social attention in the generation

of timeline summaries. Given a topic, the authors propose learning users' collective interests in the form of word distributions from Twitter which are subsequently incorporated into a unified framework for timeline summary generation. In a similar problem, Chang et al. [9] introduce the task of Twitter context summarization, which generates a succinct summary from a large but noisy Twitter context tree. The authors study how user influence models, which project user interaction information onto a Twitter context tree, can help Twitter context summarization within a supervised learning framework.

Regarding more recent works on timeline summarization, Yao et al. [42] focus on how to select a small set of representative tweets to generate a meaningful timeline, which provides enough coverage for a given topical query. The proposed approach jointly models individual topical relevance and overall diversity within a probabilistic model. Chang et al. [10] propose a framework called *Timeline-Sumy*, which consists of two main components: *episode detecting*, and *summary ranking*. Episode detecting aims to identify key episodes in a timeline, while summary ranking ranks the social media posts in each episode via a learning-to-rank approach.

Finally, Li and Cardie [23] propose an unsupervised framework for creating a chronological list of a user's personal important events. The authors introduce a non-parametric multi-level Dirichlet Process model to recognize four types of tweets: personal time-specific, personal time-general, public time-specific, and public time-general. These tweets, in turn, are used for further personal event extraction and timeline generation.

To our knowledge, our work is the first that models multi-aspect *entity-centric* analytics for social media archives, by combining automatically extracted entities with sentiment information expressed in the tweets. The proposed measures capture the multi-aspect behavior of an entity in different time periods and can be exploited in a variety of tasks, like entity evolution, event detection, and entity recommendation. In addition, our approach does not rely on service-specific labels (likes hashtags) and thus can be applied over any type of time-annotated short texts.

3 Multi-aspect entity measures

3.1 Motivation

According to Weikum et al. [41], when exploring archived data, like old web archives, analysts prefer to deal with semantically rich entities like people, places, organizations, and ideally relationships among them, instead of documents containing such references. The authors envision a system that should support a wide spectrum of analytical tasks that span the text, entity, and time dimensions, such as identification of salient entities for different subsets of an archive, entity-to-entity co-occurrences, or detection of interesting time points or periods for a given entity. In addition, to preserve Twitter as a historical source, Bruns and Weller [6] suggest that important events should be monitored while systems should offer the possibility to collect tweets for single events in order to document important background information or other contextual information (like related entities).

Considering the above, we propose a set of *entity-centric* measures that allow studying how entities (including events) are reflected in social media in different time periods and under different aspects. We propose a multi-aspect description of an entity in terms of the following aspects (computed for a given time period, like a specific day, week, or month):

- *entity popularity* (how much discussion it generates)
- *entity attitude* (predominant sentiment toward the entity)
- *entity sentimentality* (magnitude of sentiment toward the entity)
- *entity controversiality* (whether there is a consensus about the sentiment toward the entity)
- *entity-to-entity connectedness* (how strong is its connection to another entity)
- *entity network* (strongly connected entities)

These time-dependent entity features can facilitate research in a plethora of related problems, including prediction tasks [34,43] (by exploiting *popularity*, *attitude* and *sentimentality*), controversy detection [17] (by exploiting *controversiality*), time-aware entity relatedness [25] (by exploiting *entity-to-entity connectedness*), and time-aware entity recommendation [44] (by exploiting *entity network*).

Below, we formally introduce the proposed measures by classifying them into three categories: (i) *single-entity measures*, (ii) *entity-time measures*, and (iii) *entity-relation measures*.

3.2 Single-entity measures

First, let C be a collection of short texts (e.g., tweets) covering the time period $T = [t_s, t_e]$ (where t_s, t_e are two different time points with $t_s < t_e$), and let U be the total set of users who posted these texts. Let also E denote a finite set of entities, e.g., all Wikipedia entities.

Popularity

Let $e \in E$ be a given entity and $T_i \subseteq T$ a given time period. Let also $C_i \subseteq C$ be the collection of short texts posted during T_i. The popularity of e during T_i equals to the percentage of

texts mentioning e during that period. Formally:

$$popularity_c(e, T_i) = \frac{|C_{e,i}|}{|C_i|}, \quad (1)$$

where $C_{e,i} \subseteq C_i$ denotes the set of texts mentioning e during T_i.

Using the above measure, an entity can be very popular even if it is discussed by a few users but in a large number of texts. A more fine-grained indication of popularity is given by the number of different users discussing the entity. In that case, if $u_c \in U$ denotes the user who posted the text c, the popularity of an entity $e \in E$ during T_i can be defined as the percentage of different *users* discussing e during that period, i.e.,

$$popularity_u(e, T_i) = \frac{|\cup_{c \in C_{e,i}} u_c|}{|\cup_{c \in C_i} u_c|}. \quad (2)$$

We can now combine both aspects (percentage of *texts* and *users*) in one popularity score:

$$popularity_{c,u}(e, T_i) = popularity_c(e, T_i) \cdot popularity_u(e, T_i). \quad (3)$$

An entity has now a high popularity score if it is discussed in many tweets and by many different users.

Attitude and sentimentality

We use two measures (proposed by Kucuktunc et al. [21] for the case of questions and answers) for capturing a text's *attitude* (predominant sentiment) and *sentimentality* (magnitude of sentiment). First, for a text $c \in C$, let $s_c^+ \in [1, 5]$ be the text's positive sentiment score and $s_c^- \in [-5, -1]$ be the text's negative sentiment score (according to SentiStrength, c.f. Sect. 2.1). The attitude of a text c is given by $\phi_c = s_c^+ + s_c^-$ (i.e., $\phi_c \in [-4, 4]$) and its sentimentality by $\psi_c = s_c^+ - s_c^- - 2$ (i.e., $\psi_c \in [0, 8]$).

We now define the *attitude* of an entity e in a time period T_i as the average attitude of texts mentioning e during T_i. Formally:

$$attitude(e, T_i) = \frac{\sum_{c \in C_{e,i}} \phi_c}{|C_{e,i}|}. \quad (4)$$

Likewise, the *sentimentality* of an entity e in a time period T_i is defined as the average sentimentality of texts mentioning e during T_i:

$$sentimentality(e, T_i) = \frac{\sum_{c \in C_{e,i}} \psi_c}{|C_{e,i}|}. \quad (5)$$

Controversiality

An entity e can be considered controversial in a time period T_i if it is mentioned in plenty of both positive and negative texts. First, let $C_{e,i}^+$ be the set of texts mentioning e during T_i with strong positive attitude, i.e., $C_{e,i}^+ = \{c \in C_{e,i} \mid \phi_c \geq \delta\}$, where $\delta \in [0, 4]$ is a strong attitude threshold (e.g., $\delta = 2.0$). Likewise, let $C_{e,i}^-$ be those with strong negative attitude, i.e., $C_{e,i}^- = \{c \in C_{e,i} \mid \phi_c \leq -\delta\}$. We now consider the following formula for entity *controversiality*:

$$controversiality(e, T_i) = \frac{|C_{e,i}^+| + |C_{e,i}^-|}{|C_{e,i}|} \cdot \frac{\min(|C_{e,i}^+|,\ |C_{e,i}^-|)}{\max(|C_{e,i}^+|,\ |C_{e,i}^-|)}. \quad (6)$$

Intuitively, a value close to 1 means that the probability of the entity being "controversial" is high since there is a big percentage of texts with strong attitude (first part of the formula) and also there are many texts with both strong positive attitude and strong negative attitude (second part of the formula).

3.3 Entity-time measures

By exploiting the single-entity measures, we can now compute important time sub-periods of granularity Δ (e.g., day, week, or month) for a given entity in a given time period. For instance, given the entity *Barack Obama*, the time period *2015*, and the granularity *month*, we can find the top-3 months of 2015 of high or low Obama's popularity. Then, for a specific *month*, we can find the top-5 *days* of high or low Obama's popularity. We define the following measures:

- Top-K time sub-periods of high/low popularity
- Top-K time sub-periods of high/low attitude
- Top-K time sub-periods of high/low controversiality

Top-K time periods of high/low popularity

Given an entity e, a time period T_i and a granularity Δ, the *top-k time periods of high popularity* of e during T_i is the set of k time (sub-)periods of granularity Δ with the highest entity popularity score (cf. Formula 3). Let first $T_{i,\Delta}$ be the set of all time (sub-)periods of granularity Δ covering the time period T_i (for example, all days in a month). Now, the top-K time (sub-)periods of high popularity of e during T_i can be defined as:

$$k\text{-}HighPopularPeriods(e, T_i, \Delta) = \underset{T' \subseteq T_{i,\Delta},\ |T'|=k}{\operatorname{argmax}} \sum_{t \in T'} popularity_{c,u}(e, t). \quad (7)$$

Likewise, the set of top-K time (sub-)periods of low popularity is defined as:

$$k\text{-}LowPopularPeriods(e, T_i, \Delta) = \underset{T' \subseteq T_{i,\Delta},\ |T'|=k}{\text{argmin}} \sum_{t \in T'} popularity_{c,u}(e, t). \quad (8)$$

Top-K time periods of high/low attitude

By exploiting the attitude measure (cf. Formula 4), we can find time periods of high or low entity attitude. Given an entity e, a time period T_i, and a granularity Δ, the *top-k time periods of high attitude* of e during T_i is the set of k time (sub-)periods of granularity Δ with the highest entity attitude score. Formally:

$$k\text{-}HighAttitudePeriods(e, T_i, \Delta) = \underset{T' \subseteq T_{i,\Delta},\ |T'|=k}{\text{argmax}} \sum_{t \in T'} attitude(e, t). \quad (9)$$

Likewise, the set of *top-k time periods of low attitude* is defined as:

$$k\text{-}LowAttitudePeriods(e, T_i, \Delta) = \underset{T' \subseteq T_{i,\Delta},\ |T'|=k}{\text{argmin}} \sum_{t \in T'} attitude(e, t). \quad (10)$$

Top-K time periods of high/low controversiality

Given an entity e, a time period T_i and a granularity Δ, the *top-k controversial time periods* of e during T_i is the set of k time (sub-)periods of granularity Δ with the highest entity controversiality score (cf. Formula 6). Formally:

$$k\text{-}HighControversialPeriods(e, T_i, \Delta) = \underset{T' \subseteq T_{i,\Delta},\ |T'|=k}{\text{argmax}} \sum_{t \in T'} controversiality(e, t). \quad (11)$$

Likewise, the set of *top-k time periods of low controversiality* is defined as:

$$k\text{-}LowControversialPeriods(e, T_i, \Delta) = \underset{T' \subseteq T_{i,\Delta},\ |T'|=k}{\text{argmin}} \sum_{t \in T'} controversiality(e, t). \quad (12)$$

3.4 Entity-relation measures

Here, we define measures that quantify the degree of association (or connectedness) of a query entity with other entities mentioned in the same collection.

Entity-to-entity connectedness

We define a *direct-connectedness* score between an entity $e \in E$ and another entity $e' \in E$ in a time period T_i, as the number of texts in which e and e' co-occur within T_i. Formally:

$$direct\text{-}connectedness(e, e', T_i) = \frac{|C_{e,i} \cap C_{e',i}|}{|C_{e,i}|}. \quad (13)$$

Notice that the relation is not symmetric. We consider that if an entity e_1 is strongly connected with an entity e_2, this does not mean that e_2 is also strongly connected with e_1. For example, consider that *Alexis Tsipras* is mentioned in only 100 texts during T_i, *Barack Obama* in 1M texts, while 90 texts mention both entities. We notice that *Barack Obama* seems to be a very important entity for *Alexis Tsipras* during T_i, since it exists in 90/100 of texts mentioning *Alexis Tsipras*. On the contrary, *Alexis Tsipras* seems not to be important for *Barack Obama*, since it exists in only 90/1M of texts mentioning *Barack Obama*.

Two entities may not co-occur in texts, but they may share many common co-occurred entities. For example, both *Barack Obama* and *Donald Trump* may co-occur with entities like *White House*, *US Election*, and *Hillary Clinton*. For an input entity $e \in E$ and another entity $e' \in E$, we define an *indirect-connectedness* score which considers the number of *common entities* with which e and e' co-occur in a time period T_i:

$$indirect\text{-}connectedness(e, e', T_i) = \frac{|(\cup_{c \in C_{e,i}} E_c) \cap (\cup_{c \in C_{e',i}} E_c)|}{|(\cup_{c \in C_{e,i}} E_c)|}, \quad (14)$$

where $E_c \subseteq E$ is the entities mentioned in text c. Also in this case, the relation between the two entities is not symmetric.

Entity *k*-network

This measure targets at finding a list of entities strongly connected to the query entity in a given time period T_i. First, we define a connectedness score between an entity $e \in E$ and a set of entities $E' \subseteq E$ within T_i, as the average direct-connectedness score of the entities in E'. Formally:

$$connectedness(e, E', T_i) = \frac{\sum_{e' \in E'} direct\text{-}connectedness(e, e', T_i)}{|E'|}. \quad (15)$$

The k-Network of an entity e during T_i is the set of k entities $E' \subseteq E$ with the highest average connectedness score. Namely:

$$k\text{-}Network(e, T_i) = \underset{E' \subseteq E,\ |E'|=k}{\text{argmax}} \ connectedness(e, E', T_i). \quad (16)$$

In simple terms, the k-Network of an entity e consists of the k entities with the highest *direct-connectedness* scores.

Positive and negative *k*-networks

Based on the attitude of the texts mentioning two entities, we can compute the corresponding positive and negative k-Networks. First, for an entity $e \in E$ and a time period T_i, the set of positive entities $E^+_{e,i} \subseteq E$ can be defined as the set of entities co-occurring with e during T_i in texts with strong average positive attitude, i.e.,

$$E^+_{e,i} = \left\{ e' \in E \mid \frac{\sum_{c \in C_{e,i} \cap C_{e',i}} \phi_c}{|C_{e,i} \cap C_{e',i}|} \geq \delta \right\}, \quad (17)$$

where $\delta \in [0, 4]$ is a strong attitude threshold (e.g., $\delta = 2.0$). Likewise, the set of entities co-occurring with e during T_i in texts with strong average negative attitude can be defined as:

$$E^-_{e,i} = \left\{ e' \in E \mid \frac{\sum_{c \in C_{e,i} \cap C_{e',i}} \phi_c}{|C_{e,i} \cap C_{e',i}|} \leq -\delta \right\}. \quad (18)$$

Now, the positive and negative k-Networks of an entity e in a time period T_i can be defined as:

$$k\text{-}Network^+(e, T_i) = \underset{E' \subseteq E^+_{e,i},\ |E'|=k}{\text{argmax}} \ connectedness(e, E', T_i), \quad (19)$$

$$k\text{-}Network^-(e, T_i) = \underset{E' \subseteq E^-_{e,i},\ |E'|=k}{\text{argmax}} \ connectedness(e, E', T_i). \quad (20)$$

3.5 Discussion

The above-presented measures capture the multi-aspect behavior of a given entity at a certain time period. In the long run, a multi-variate time series is formed where each point represents the multi-aspect description of the entity at a certain period in time.

An important characteristic of our approach is that we can support both entity-specific queries referring to a single-entity and cross-entity queries involving more than one entities (e.g., a category of entities). This is achieved through the *entity linking* process in which entities are extracted from the texts and are linked to knowledge bases like Wikipedia/DBpedia. In that way, we can collect a variety of properties for the entities extracted from our archive. This enables us to aggregate information and capture the behavior of sets of entities. For example, by accessing DBpedia, we can collect a list of German politicians, derive their popularity, and then compare it with that of another set of entities.

In addition, the proposed measures can be easily computed by submitting queries on related knowledge bases that contain metadata and annotation information about a collection of archived documents or social media posts [12,14,15]. This enables the production of time series at query execution time, thereby allowing the answer of complex information needs (through structured SPARQL queries) as well as "on the fly" data integration (by exploiting other knowledge bases like DBpedia).

4 Library for computing the measures

For computing the measures, we provide an Apache Spark library. Apache Spark[2] is a cluster-computing framework for large-scale data processing. The library contains functions for computing the proposed measures for a given entity and over a specific time period. It operates over an annotated (with entities and sentiments) dataset split per year-month (the dataset should be in a simple CSV format). The library is available as open source.[3]

The time for computing the measures highly depends on the dataset volume, the used computing infrastructure as well as the available resources, and the load of the cluster at the analysis time. The Hadoop cluster used in our experiments for analyzing a large Twitter archive of more than 1 billion tweets consisted of 25 computer nodes with a total of 268 CPU cores and 2688 GB RAM (more about the dataset in the next section). Indicatively, the time for computing each of the measures was on average less than a minute (without using any index, apart from the monthly-wise split of the dataset).

5 Case study: entity analytics on a Twitter archive

In this section, we first describe the results of the analysis and annotation of a large Twitter archive. Then, we present examples of case studies illustrating the insights gained from the proposed measures.

5.1 Annotating a large Twitter archive

We analyzed a large Twitter archive spanning 4 years (January 2014–January 2017) and containing more than 6

[2] http://spark.apache.org/.

[3] https://github.com/iosifidisvasileios/Large-Scale-Entity-Analysis.

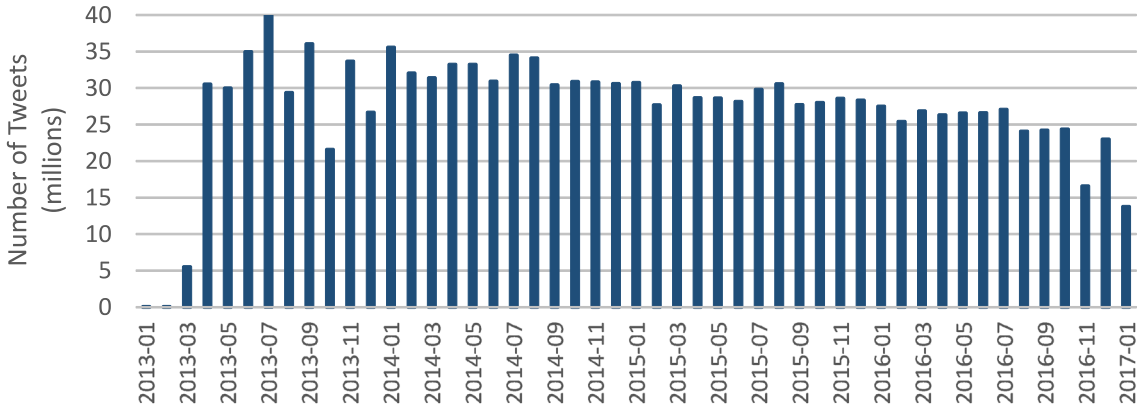

Fig. 1 Number of tweets per month

billion tweets. The tweets were collected through the Twitter streaming API. Our analysis comprised the following steps: (i) filtering (filtering out re-tweets, keeping only English tweets), (ii) spam removal, (iii) entity linking, and (iv) sentiment analysis. The filtering step reduced the number of tweets to about 1.5 billion tweets (specifically to 1,486,473,038 tweets). For removing the spam tweets, we trained a multinomial naive Bayes (MNB) classifier over the HSpam dataset [36], which consists of tweets annotated as either spam or not. We applied the learned model to our dataset and removed all tweets classified as spam. This removed about 150 million tweets (around 10% of the tweets). The final dataset consists of 1,335,324,321 tweets posted by 110,548,539 users. Figure 1 shows the number of tweets per month on the final dataset.

For the *entity linking* task, we used Yahoo FEL [4] with a confidence threshold score of −3. In total, 1,390,286 distinct entities were extracted from the collection. For each extracted entity, its confidence score provided by FEL is also stored. Thereby, data consumers can select suitable confidence ranges to consider, depending on the specific requirements with respect to precision and recall. For *sentiment analysis*, we used SentiStrength [39]. The average sentimentality of all tweets is 0.92, the average attitude is 0.2, while 622,230,607 tweets have no sentiment (−1 negative sentiment and 1 positive sentiment). Table 1 shows the number of tweets per attitude value.

Table 1 Number of tweets per attitude value

Attitude value	Number of tweets
−4	2,234,887 (0.17%)
−3	34,666,708 (2.60%)
−2	68,812,370 (5.15%)
−1	104,628,022 (7.84%)
0	670,484,267 (50.2%)
1	301,635,430 (22.6%)
2	138,197,637 (10.3%)
3	13,610,492 (1.02%)
4	1,054,508 (0.08%)

Quality of entity annotations

We used the ground truth dataset provided by the 2016 NEEL challenge of the 6th workshop on "Making Sense of Microposts" (#Microposts2016)[4] [31] for evaluating the quality of the entity annotations produced by FEL. The dataset consists of 9289 English tweets of 2011–2015. We considered all tweets from the provided training, dev and test files. The results are the following: *Precision* = 86%, *Recall* = 39%, *F1* = 54%. We notice that FEL achieves high precision; however, recall is low. The reason is that FEL did not manage to recognize several difficult cases, like entities within hashtags and nicknames, which are common in Twitter. Nevertheless, FEL's performance is comparable to existing approaches [30,31].

Quality of sentiment annotations

We evaluated the accuracy of SentiStrength using the ground truth datasets *SemEval2017* (Task 4, Subtask A)[5] [32] and *TSentiment15*[6] [20]. The *SemEval2017* dataset consists of 61,853 English tweets collected during the period 2013–2017 and labeled by human annotators as positive, negative, or neutral. We run the evaluation on all the provided training files (of 2013–2016) and on the 2017 test file. SentiStrength achieved the following scores: *AvgRec = 0.54* (recall averaged across the positive, negative, and neutral classes [35]), $F1^{PN}$=

[4] http://microposts2016.seas.upenn.edu/.

[5] http://alt.qcri.org/semeval2017/task4/.

[6] https://l3s.de/~iosifidis/TSentiment15/.

0.52 (F1 averaged across the positive and negative classes), *Accuracy* = 0.57. The performance of SentiStrength is good considering that this is a multi-class classification problem. The user can also achieve higher precision by selecting only tweets with high positive or negative SentiStrength score (e.g., $>+2$ for positive or <-2 for negative sentiment). Regarding the *TSentiment15* dataset, it consists of 2,527,753 English tweets collected during 2015, labeled as either positive or negative through semi-supervised learning [20]. SentiStrength achieved the following scores: $F1^{PN} = 0.80$, *Accuracy* = 0.91. Here, we notice that SentiStrength achieves very good performance. To conclude, our evaluation on Twitter ground truth datasets showed that SentiStrength achieves good performance in sentiment annotation of tweets.

Dataset availability

The annotated dataset is publicly available in CSV format.[7] For each tweet, the dataset includes the following information: ID, user (encrypted), post date, extracted entities, and positive and negative sentiment values. (The text of the tweets is not provided for copyright purposes.[8])

We make the dataset available so anyone interested can use it together with the library (described in Sect. 4) to extract the measures for any entity at the desired level of temporal granularity. We believe that such efforts can foster further research in a variety of areas like *entity recommendation*, *entity summarization*, and *concept drift*.

5.2 Case studies

Entity popularity

Figure 2 shows the popularity of *Alexis Tsipras* (Greek prime minister) within 2015. We notice that his popularity highly increased in July. Indeed, in July 2015 the Greek bailout referendum was held following the bank holiday and capital controls of June 2015. This event highly increased the popularity of the Greek prime minister. Moreover, by comparing the trend of the two different popularity scores (Formulas 1 and 2), we notice that, during June and July 2015, the percentage of different users discussing about *Alexis Tsipras* increased in bigger degree compared with the percentage of tweets, implying that more people were engaged in the discussion. As regards his *top-K time periods of high/low popularity* (Formulas 7 and 8), we notice that the top-3 months of high popularity in 2015 are [July, June, February], while the corresponding top-3 months of low popularity are [December, November, May].

[7] http://l3s.de/~iosifidis/tpdl2017/.

[8] https://help.twitter.com/en/rules-and-policies/copyright-policy.

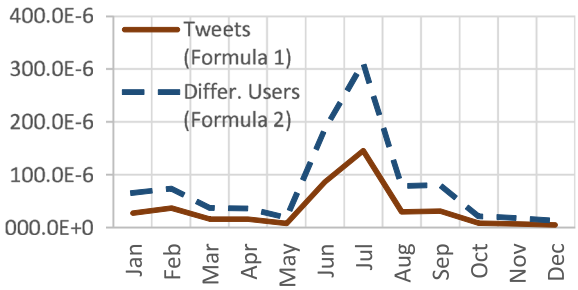

Fig. 2 Popularity evolution of *Alexis Tsipras* in 2015

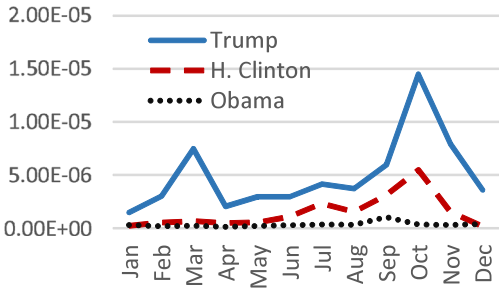

Fig. 3 Popularity evolution of *Donald Trump*, *Hillary Clinton*, and *Barack Obama* in 2016

Likewise, we can compare the popularity of multiple entities within the same time period. For example, Fig. 3 shows the popularity of *Donald Trump*, *Hillary Clinton*, and *Barack Obama* within 2016 (according to Formula 3). We notice that *Donald Trump* is much more popular in all months. We also notice that in October 2016 the popularity of *Donald Trump* and *Hillary Clinton* highly increased compared with the other months. This is an indicator of possible important events related to these two entities in October 2016. (Indeed, two presidential general election debates took place in that period.) The top-3 months of high popularity in 2016 are [October, November, March] for *Donald Trump*, [October, September, July] for *Hillary Clinton*, and [September, December, July] for *Barack Obama*.

Entity attitude

Figure 4 (left) depicts the attitude of *Donald Trump* and *Hillary Clinton* within 2016. We notice that both entities had constantly a negative attitude; however, that of *Hillary Clinton* was worse in almost all months. Moreover, we notice that *Hillary Clinton*'s attitude highly decreased in May 2016 (possibly, for example, due to a report issued by the State Department related to Clinton's use of private email), while October 2016 was the month with the lowest attitude value for Donald Trump (possibly due to the several sexual assault allegations leveled against Donald Trump during that period). The top-3 months of *high attitude* in 2016 are [February, December, April] for Donald Trump and [February, March,

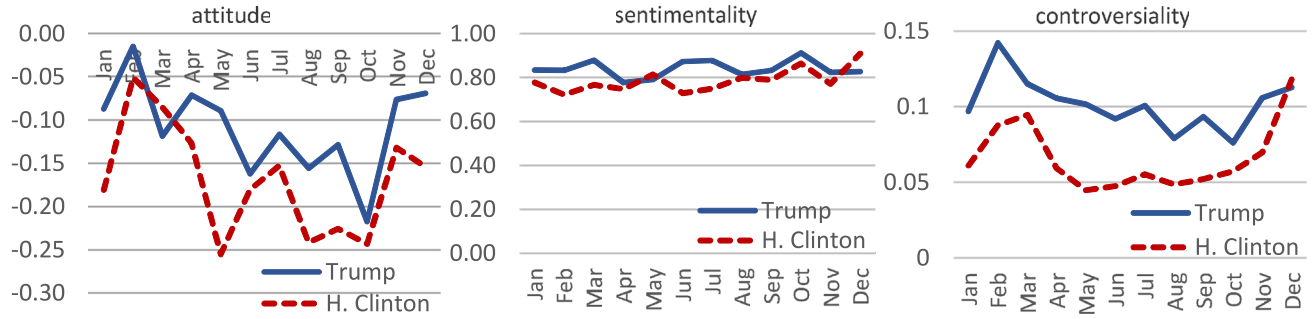

Fig. 4 Evolution of attitude (left), sentimentality (middle), and controversiality (right) of *Donald Trump* and *Hillary Clinton* in 2016

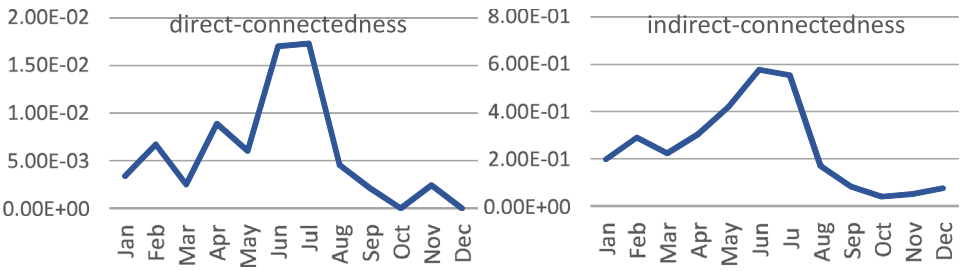

Fig. 5 Direct (Formula 13) and indirect (Formula 14) connectedness of *"Alexis Tsipras"* with *"Greek withdrawal from the eurozone"* in 2015

April] for Hillary Clinton, while the corresponding top-3 months of *low attitude* are [October, June, August] for Donald Trump and [May, October, August] for Hillary Clinton.

In general, we notice that the attitude values are relatively small and close to zero. This is due to the very big number of tweets with no sentiment (almost half of the tweets).

Entity sentimentality

Figure 4 (middle) depicts the sentimentality of *Donald Trump* and *Hillary Clinton* within 2016. We notice that for the majority of months the tweets mentioning *Donald Trump* are a bit more sentimental than those mentioning *Hillary Clinton*.

October seems to be one of the most "sentimental" months for both *Donald Trump* and *Hillary Clinton*, possibly due to the several revelations that were uncovered for both candidates the period before the US presidential election (held on November 8).

Entity controversiality

Figure 4 (right) shows the controversiality of *Donald Trump* and *Hillary Clinton* within 2016 (using $\delta = 2.0$). We notice that *Donald Trump* induces more controversial discussions in Twitter than *Hillary Clinton*, while February was his most "controversial" month, probably because of his references to some debatable topics during his campaign trail. It is interesting also that *Hillary Clinton*'s controversiality has an exponential increment from September to December 2016 (the period before, during, and after the US presidential election).

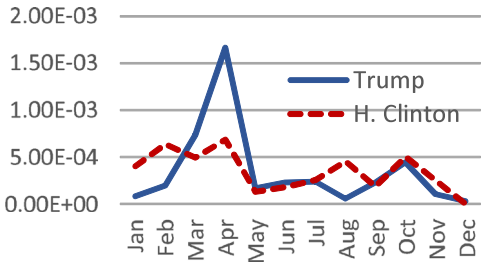

Fig. 6 Direct connectedness (Formula 13) of *"Donald Trump"* and *"Hillary Clinton"* with *"Abortion"* in 2016

As regards their *top-K time periods of high/low controversiality* (Formulas 11 and 12), the top-3 months of *high controversiality* in 2016 are [February, March, December] for *Donald Trump* and [December, March, February] for *Hillary Clinton*, while the corresponding top-3 months of *low controversiality* are [October, August, June] for *Donald Trump* and [May, June, August] for *Hillary Clinton*.

Entity-to-entity connectedness

Figure 5 depicts the connectedness of *Alexis Tsipras* with the concept *Greek withdrawal from the eurozone* within 2015. We notice that these two entities are highly connected in June and July, while after August, their connectedness is very close to zero. Indeed, important events related to Greece's debt crisis took place in June and July 2015, including the bank holiday, the capital controls, and the Greek bailout referendum.

Likewise, Fig. 6 shows the connectedness of both *Donald Trump* and *Hillary Clinton* with the concept *Abortion* in 2016.

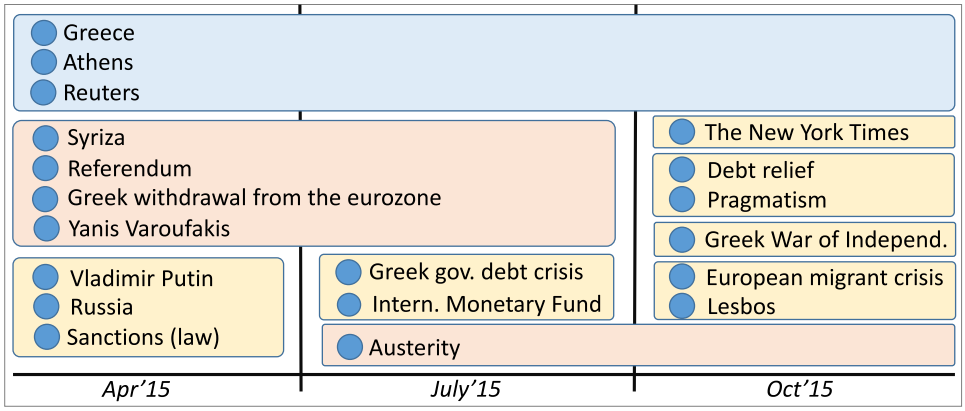

Fig. 7 10-Network of *Alexis Tsipras* in April, July, and October 2015

Table 2 Negative 5-Network of Donald Trump in October 2016 (using $\delta = 2.0$)

Rank	Entity
1	Iraq War
2	Bill Clinton
3	Toddler
4	Embarrassment
5	Central Park

Here, we notice that the connectedness is almost constant for *Hillary Clinton*, while for *Donald Trump*, there is a very large increment in March and April. During these two months, Donald Trump made several antiabortion comments, like that *"there has to be some form of punishment"* for women who have abortions.[9]

Entity *k*-network

Figure 7 shows the 10-Network of *Alexis Tsipras* in three different time periods (April, July, and October 2015). We notice that there are three general entities that exist in all time periods (*Greece*, *Athens*, *Reuters*). For April and July, we notice that the 10-Network contains four common entities (*Syriza*, *Referendum*, *Greek withdrawal from the eurozone*, and *Yanis Varoufakis*), while for July and October, *Austerity* is the only common entity (probably related to the approval of strict measures required by the creditors). For April, the 10-Network contains three entities related to Russia (due to Tsipra's visit in Moscow to meet Russian president Vladimir Putin), while for October, it contains two entities related to European migrant crisis (probably due to Tsipra's visit in Lesvos island).

Entity positive and negative *k*-networks

In Fig. 4 (left), we notice that the attitude toward *Donald Trump* highly decreased in October 2016. To understand this decrement, we can inspect his *Negative k-Network* during the same period, i.e., the entities that co-occur with *Donald Trump* in tweets with strong average negative attitude during October 2016. Table 2 shows the results for $k = 5$ and $\delta = 2.0$. We notice that the top-5 list contains entities related to important events that happened during this period and which are related to *Donald Trump*, including *Iraq War* (Donald Trump said that he opposed Iraq War from the start; however, there appeared audio evidence of him saying he supported it), *Bill Clinton* (Donald Trump appeared with Bill Clinton accusers before a debate), *Toddler* (Donald Trump brought a toddler to the stage during a campaign rally), and *Central Park* (related to the Central Park jogger case: Donald Trump declared that the Central Park Five were guilty).

In Fig. 4 (left), we also notice that the attitude toward *Donald Trump* highly increased in November–December 2016. Table 3 shows the corresponding *Positive 5-Network* of *Donald Trump* for this time period (using again $\delta = 2.0$). The top-5 list contains entities related to Donald Trump's election (*Cold open*, *God Bless America*, *Cheers*, and *Excite*) as well as the entity *Henry Kissinger* with whom Donald Trump met on November 17.

5.3 Limitations and problems

Although the proposed analysis approach is generic and can be applied over different types of social media archives, it is clear that the quality of the generated data depends on the quality of the input data. Twitter, for example, provides 1% random sample, which though is subject to bias, fake news, and possibly other adversarial attacks. In our case study, although we remove spam, we do not take similar actions to deal with bias and other data peculiarities. This also means that high-profile entities might occupy a big volume in the archive, whereas long-tail entities might be underrepresented or not represented at all.

[9] https://www.nytimes.com/2016/03/31/us/politics/donald-trump-abortion.html (August 30, 2018).

Table 3 Positive 5-Network of Donald Trump in November–December 2016 (using $\delta = 2.0$)

Rank	Entity
1	Cold open
2	God Bless America
3	Henry Kissinger
4	Cheers
5	Excite

Except for the quality of the original data, the different preprocessing steps (spam removal, entity linking, sentiment analysis) are also prone to errors. This means that the data produced by the proposed measures are also prone to errors. For instance, regarding the entity linking task, selecting a low threshold for the confidence score of the extracted entities can result in many false annotations (disambiguation errors), which in turn can affect the quality and reliability of the produced time series. For the case of *Entity k-Networks* in particular, one may get some unexpected and surprising results due to disambiguation errors. For instance, the *Negative 10-Network* of Donald Trump for October 2016 returns the entity *Harrow (tool)*, while that of Hillary Clinton for May 2016 returns the entity *Clueless (film)*. Both these two entities have been incorrectly disambiguated by FEL.

6 Conclusion

We have proposed an entity-centric and multi-aspect approach to analyze social media archives. For this, we defined a set of measures that allow studying how entities are reflected in social media in different time periods as well as how entity-related information evolves over time and also with respect to other entities. The proposed measures enable the temporal analysis of an entity in terms of its: *popularity* (how much discussion it generates), *attitude* (predominant sentiment toward the entity), *sentimentality* (magnitude of sentiment toward the entity), *controversiality* (whether there is a consensus about the sentiment toward the entity), *connectedness* to another entity (how strong is its connection to another entity), and *network* (strongly connected entities).

We believe that such a multi-aspect analysis approach is the first step toward more advanced and meaningful exploration of social media archives, while it can facilitate research in a variety of fields, such as data science, sociology, and digital humanities.

As part of our future work, we plan to exploit the rich amount of generated data for *prediction* of entity-related features. In particular, given an entity, our focus will be on how we can predict future values of the proposed measures (e.g., popularity or attitude in a given horizon) [34]. We also intend to study approaches on *understanding* and *representing* the dynamics of such evolving entity-related information, using for instance an RDF-based approach [33]. Another interesting direction is the exploitation of the entity-relation measures on the related problems of *time-aware entity relatedness* [25] and *event timeline summarization* [40].

Acknowledgements The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA (No. 339233) and by the German Research Foundation (DFG) project OSCAR (Opinion Stream Classification with Ensembles and Active leaRners).

References

1. Amigó, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2014: Author profiling and reputation dimensions for online reputation management. In: CLEF (2014)
2. Ardon, S., Bagchi, A., Mahanti, A., Ruhela, A., Seth, A., Tripathy, RM., Triukose, S.: Spatio-temporal analysis of topic popularity in Twitter. arXiv preprint arXiv:1111.2904 (2011)
3. Batrinca, B., Treleaven, P.C.: Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY **30**(1), 89–116 (2015)
4. Blanco, R., Ottaviano, G., Meij, E.: Fast and space-efficient entity linking for queries. In: WSDM (2015)
5. Bruns, A., Stieglitz, S.: Towards more systematic Twitter analysis: metrics for tweeting activities. Int. J. Soc. Res. Methodol. **16**(2), 91–108 (2013)
6. Bruns, A., Weller, K.: Twitter as a first draft of the present: and the challenges of preserving it for the future. In: WebSci (2016)
7. Celik, I., Abel, F., Houben, G.J.: Learning semantic relationships between entities in Twitter. In: ICWE (2011)
8. Chandrasekaran, B., Josephson, J.R., Benjamins, V.R.: What are ontologies, and why do we need them? IEEE Intell. Syst. Appl. **14**(1), 20–26 (1999)
9. Chang, Y., Wang, X., Mei, Q., Liu, Y.: Towards twitter context summarization with user influence models. In: WSDM (2013)
10. Chang, Y., Tang, J., Yin, D., Yamada, M., Liu, Y.: Timeline summarization from social media with life cycle models. In: IJCAI (2016)
11. Chen, P.P.S.: The entity-relationship model toward a unified view of data. ACM Trans. Database Syst. (TODS) **1**(1), 9–36 (1976)
12. Fafalios, P., Holzmann, H., Kasturia, V., Nejdl, W.: Building and querying semantic layers for web archives. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp 1–10. IEEE (2017a)
13. Fafalios, P., Iosifidis, V., Stefanidis, K., Ntoutsi, E.: Multi-aspect entity-centric analysis of big social media archives. In: International Conference on Theory and Practice of Digital Libraries, pp 261–273. Springer (2017b)
14. Fafalios, P., Holzmann, H., Kasturia, V., Nejdl, W.: Building and querying semantic layers for web archives (extended version). Int. J. Digit. Libr. (2018a) https://doi.org/10.1007/s00799-018-0251-0
15. Fafalios, P., Iosifidis, V., Ntoutsi, E., Dietze, S.: Tweetskb: A public and large-scale rdf corpus of annotated tweets. In: European Semantic Web Conference, pp. 177–190. Springer (2018b)
16. Farzindar, A., Khreich, W.: A survey of techniques for event detection in twitter. Comput. Intell. **31**(1), 132–164 (2015)

17. Garimella, K., Morales, G.D.F., Gionis, A., Mathioudakis, M.: Quantifying controversy on social media. ACM Trans. Soc. Comput. **1**(1), 3 (2018)
18. Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. SIGMOD Rec. **42**(2), 17–28 (2013)
19. Heath, T., Bizer, C.: Linked data: evolving the web into a global data space. Synth. Lect. Semant. Web Theory Technol. **1**(1), 1–136 (2011)
20. Iosifidis, V., Ntoutsi, E.: Large scale sentiment learning with limited labels. In: KDD (2017)
21. Kucuktunc, O., Cambazoglu, B.B., Weber, I., Ferhatosmanoglu, H.: A large-scale sentiment analysis for Yahoo! answers. In: WSDM (2012)
22. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S.: Dbpedia-a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web **6**(2), 167–195 (2015)
23. Li, J., Cardie, C.: Timeline generation: Tracking Individuals on Twitter. In: WWW (2014)
24. Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., Wang, H.: Entity-centric topic-oriented opinion summarization in Twitter. In: KDD (2012)
25. Mohapatra, N., Iosifidis, V., Ekbal, A., Dietze, S., Fafalios, P.: Time-aware and corpus-specific entity relatedness. In: Workshop on Deep Learning for Knowledge Graphs and Semantic Technologies (DL4KGS)—In conjunction with ESWC 2018, Heraklion, Greece (2018)
26. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: Semeval-2016 task 4: Sentiment analysis in twitter. In: SemEval@ NAACL-HLT (2016)
27. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2007)
28. Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: Identifying misinformation in microblogs. In: EMNLP (2011)
29. Ren, Z., Liang, S., Meij, E., de Rijke, M.: Personalized time-aware tweets summarization. In: SIGIR (2013)
30. Rizzo, G., Basave, A.E.C., Pereira, B., Varga, A.: Making sense of microposts (#microposts2015) named entity recognition and linking (NEEL) challenge. CEUR-WS.org (2015)
31. Rizzo, G., van Erp, M., Plu, J., Troncy, R.: Making sense of microposts (#microposts2016) named entity recognition and linking (NEEL) challenge. CEUR-WS.org (2016)
32. Rosenthal, S., Farra, N., Nakov, P.: Semeval-2017 task 4: Sentiment analysis in twitter. In: SemEval (2017)
33. Roussakis, Y., Chrysakis, I., Stefanidis, K., Flouris, G., Stavrakas, Y.: A Flexible Framework for Understanding the Dynamics of Evolving RDF Datasets. In: ISWC (2015)
34. Saleiro, P., Soares, C.: Learning from the news: Predicting entity popularity on twitter. In: International Symposium on Intelligent Data Analysis, pp. 171–182. Springer (2016)
35. Sebastiani, F.: An axiomatically derived measure for the evaluation of classification algorithms. In: ICTIR (2015)
36. Sedhai, S., Sun, A.: Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In: SIGIR (2015)
37. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2015)
38. Stefanidis, K., Koloniari, G.: Enabling Social Search in Time through Graphs. In: Web-KR@CIKM (2014)
39. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social Web. J. Am. Soc. Inf. Sci. Technol. **63**(1), 163–173 (2012)
40. Tran, T.A., Niederée, C., Kanhabua, N., Gadiraju, U., Anand, A.: Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1201–1210. ACM (2015)
41. Weikum, G., Spaniol, M., Ntarmos, N., Triantafillou, P., Benczúr, A., Kirkpatrick, S., Rigaux, P., Williamson, M.: Longitudinal Analytics on Web Archive Data: It's About Time! In: CIDR (2011)
42. Yao, J.g., Fan, F., Zhao, W.X., Wan, X., Chang, E., Xiao, J.: Tweet timeline generation with determinantal point processes. In: AAAI (2016)
43. Yu, S., Kak, S.: A survey of prediction using social media (2012). arXiv preprint arXiv:1203.1647
44. Zhang, L., Rettinger, A., Zhang, J.: A probabilistic model for time-aware entity recommendation. In: International Semantic Web Conference, pp. 598–614. Springer (2016)
45. Zhao, X.W., Guo, Y., Yan, R., He, Y., Li, X.: Timeline generation with social attention. In: SIGIR (2013)
46. Zimmer, M.: The Twitter Archive at the Library of Congress: challenges for information practice and information policy. First Monday (2015). https://doi.org/10.5210/fm.v20i7.5619