Knowledge and Information Systems
https://doi.org/10.1007/s10115-019-01429-z

SURVEY PAPER

A survey of recent methods on deriving topics from Twitter: algorithm to evaluation

Robertus Nugroho[1,2] · Cecile Paris[2] · Surya Nepal[2] · Jian Yang[3] · Weiliang Zhao[4]

Received: 12 September 2018 / Revised: 26 November 2019 / Accepted: 30 November 2019

Abstract
In recent years, studies related to topic derivation in Twitter have gained a lot of interest from businesses and academics. The interconnection between users and information has made social media, especially Twitter, an ultimate platform for propagation of information about events in real time. Many applications require topic derivation from this social media platform. These include, for example, disaster management, outbreak detection, situation awareness, surveillance, and market analysis. Deriving topics from Twitter is challenging due to the short content of the individual posts. The environment itself is also highly dynamic. This paper presents a review of recent methods proposed to derive topics from social media platform from algorithms to evaluations. With regard to algorithms, we classify them based on the features they exploit, such as content, social interactions, and temporal aspects. In terms of evaluations, we discuss the datasets and metrics generally used to evaluate the methods. Finally, we highlight the gaps in the research this far and the problems that remain to be addressed.

Keywords Topic derivation · Twitter analysis · Algorithms · Evaluations

✉ Robertus Nugroho
nugroho@unika.ac.id

Cecile Paris
cecile.paris@data61.csiro.au

Surya Nepal
surya.nepal@data61.csiro.au

Jian Yang
jian.yang@mq.edu.au

Weiliang Zhao
weiliangzhao.email@gmail.com

1 Soegijaparanata Catholic University, Semarang, Indonesia

2 CSIRO Data61 Australia, Marsfield, Australia

3 Macquarie University, Macquarie Park, Australia

4 Donghua University, Changning, China

Published online: 09 January 2020

1 Introduction

Twitter is one of the social media platforms widely used worldwide enabling people to post short pieces of information on any matter. People might post a message for a wide range of reasons, such as to state someone's mood in a moment [56,108], to advertise one's business, to comment on current events, or to report an accident or disaster [48]. With the widespread and continuous use of social media by such a large community, there is a need to understand what the topics under discussion are. This is the goal of topic derivation. Deriving topics from Twitter is the task of identifying the main topic of each tweet (Twitter post) and listing the most important keywords to represent each topic.

Founded in 2006, Twitter now has millions of active users globally. It was based on the idea of Jack Dorsey (Twitter co-founder) to broadcast users' status update to friends utilizing an SMS-based messaging platform [14]. The limit of 140 characters used in earlier versions of Twitter was based on the limit of the number of characters for one SMS (short message service). Jack Dorsey (@jack) posted his first tweet on March 21, 2006. In March 2007, Twitter won the *Web Award* from the SXSW Interactive conference [116]. It is a prestigious award given to honor the best and most exciting technology development in the digital era. Just about three years after the first tweet, the number of tweets posted in Twitter had reached a billion [115].

Table 1 shows some statistics about Twitter in early 2019. There were around 330 million active Twitter users per month, posting roughly 500 million tweets per day ($\pm$6000 tweets per second). 82% of the active users posted their tweets from mobile devices. Twitter has 35 offices around the world with more than 4000 employees. With 79% of the accounts outside the USA, the Twitter platform now supports more than 40 languages. These facts make Twitter one of the most active social network platforms worldwide. In addition, Twitter also provides real-time access through API to stream the posts. This is one of the most important advantages of Twitter compared to Facebook, Instagram, or other social media platforms.

With its large number of users and its ability to deliver real-time updates, Twitter has been used as an important source by journalists and government organizations to obtain the latest information about unfolding events. The photograph of the US Airways plane crashed into the Hudson River was first posted and seen on Twitter before being reported by traditional news media [116]. In 2011, Twitter proved to be one of the mass communication media reporting on the unfolding events in the Arab spring movement, attracting the interests of

Table 1 Twitter facts

Items	Facts/number
Monthly active users	330M
Tweets per day	500M
Unique monthly visits to sites with embedded tweets	1B
Active users on mobile	82%
Accounts outside the USA	79%
Number of employees	4100
Number of offices around the world	35+
Number of supported languages	40+

The facts are compiled from https://about.twitter.com/company (accessed April 24, 2019) except the tweets per day, which are from http://www.internetlivestats.com/twitter-statistics/ (accessed April 24, 2019)

journalists to source news from this platform [39]. Information about other types of events was also frequently and widely spread through Twitter. They include, for example, the 2009 earthquake in Japan [89], the floods in Australia [19], and Obama's presidential election [116].

Nowadays, following the high level of user activities on this platform, most of the news channels have accounts on Twitter and post their current headlines to the platform [80]. Brands and public figures, including actors, athletes, and politicians, are taking advantages of the exponential rise of Twitter users to maximize their influence. More than 80% of world leaders are active on Twitter [24].[1] The above situation has attracted the interests of businesses and researchers to develop methods for topic derivation in Twitter: identifying what is being discussed. The ability to derive topics from this platform is very important for various critical applications [119], including disaster management, outbreak detection, situation awareness, surveillance [65], and market analysis [23]. The task of topic derivation in Twitter is challenging due to the short length of the posts, and the language used (which includes abbreviations, misspelling, and non-conventional terminology and syntax).

This paper presents our review of key techniques and current studies on deriving topics from Twitter. To get comprehensive review, we start the process by exploring the existing surveys related to the area of topic derivation for the Twitter platform. We then identify the state-of-the-art algorithms, the major extensions and their techniques used to derive topics in this environment. Different from existing surveys, our review ranges from algorithms to evaluations. In terms of algorithms, we classify the methods based on the features they exploit, such as content, social interactions, and temporal aspects. With regard to evaluations, we discuss the datasets and metrics used to evaluate the methods as an integral part of the study of topic derivation in Twitter environment.

This paper is organized as follows: In Sect. 2, we present the summary of existing survey papers. We describe the general task of topic derivation and its prominent methods in Sect. 3. Section 4 focuses on the review of recent topic derivation methods by classifying the literature based on their incorporated features. Section 5 discusses the datasets and common evaluation metrics used to measure the performance of the proposed solutions. In Sect. 6, we highlight the practical issues, challenges, and future directions. Section 7 concludes the paper.

2 Existing surveys

In this section, we present a comprehensive summary of recent surveys in the area of topic derivation, especially related to social media. They provide reviews of relevant works from many points of view, for example, the types of topics or applications (e.g., related to law enforcement, drug reaction detection, outbreaks, news, or disasters), the evolution of topics, and the detection methods (e.g., supervised vs unsupervised, offline vs online).

A paper by Atefeh and Khreich [6] presents a survey of event detection techniques in Twitter. It classifies the literature according to the event types (i.e., specified and unspecified events), and the detection methods (i.e., supervised and unsupervised). Alghamdi and Alfalqi [2] categorize the reviewed literature into two common approaches: normal topic models and topic evolution models with a time factor. Only popular methods in topic modeling are included. Song et al. [107] present a survey for the short text classification. It discusses the characteristics of a short text (like Twitter) and its challenges for the classification process.

[1] According to the Digital Policy Council (DPC) annual report on 2015 World Leader Ranking on Twitter [24], a total of 139 world leaders from 167 countries have an account in Twitter.

The works reviewed are divided into four major categories: text classification using semantic analysis, semi-supervised short text classification, ensemble short text classification, and real-time classification.

Rafeeque and Sendhilkumar [97] review several works on short text analysis in the following categories: semantic similarity using Web search for data enrichment, semantic similarity using data repositories for data enrichment, short text classification, and short text clustering (unsupervised). Jelisavčić et al. [46] provide an overview of popular probabilistic models used in topic modeling. Hong and Davison [42] conducted an empirical study of topic modeling methods in Twitter by comparing the performance of LDA [9] and the author–topic model [101].

Several surveys are focused on the evolution of topics. Aggarwal and Subbian [1] provide an overview of some literature on graph evolution analysis and application. The reviews are generally divided into two categories: the maintenance methods and the analysis of analytical evolution. The paper also discusses two different types of speed in the evolution of topics: slowly evolving networks and streaming networks. Campos et al. [12] survey studies related to temporal information retrieval and their applications. It also discusses the temporal-based clustering and classification of Web pages and social network results. The survey by Silva et al. [106] focuses on methods that address the temporal aspects involved in data stream clustering in the domain of network intrusion detection, sensor networks, and stock market analysis.

On the application level, Edwards et al. [30] conduct a survey of works related to data mining technology specifically intended for law enforcement. The works included in the survey were found through queries focused on crime, police, and law enforcement. These queries were augmented with keywords from data mining area, such as from artificial intelligence, data fusion, data mining, information fusion, NLP, machine learning, social network analysis, and text mining. Collected papers are then classified based on problem topics (e.g., financial crime, cybercrime, criminal threats or harassment, and police intelligence), and crossed with the data mining techniques. Topic detection in Twitter is specifically discussed in the social network and terrorism/extremism section. A paper by Karimi et al. [49] presents a survey of text mining techniques for the surveillance of adverse drug reaction on various platforms, including social media. Nurwidyantoro and Winarko [88] present limited survey of event detection methods in social media based on the types of event (disaster, traffic, outbreak, and news).

Different from existing surveys, our paper not only focuses on the review of the approaches but also discusses the features that are exploited to deal with the extreme sparsity and dynamics of the social media environment. The organization of the survey is shown in Fig. 1. We identify the main algorithms used for topic derivation and collect the major extensions and methods that works in the Twitter environment. We conduct a comprehensive review of the most prominent and recent algorithms for topic derivation classifying them based on the features they exploit, such as content, social interactions, and temporal aspects. We also

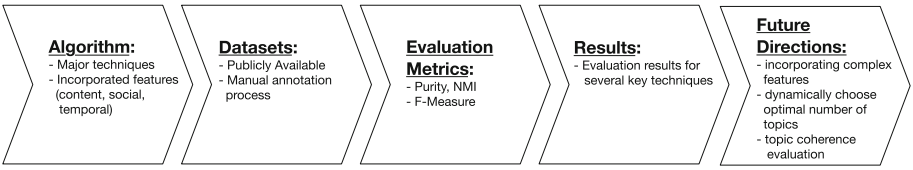


Fig. 1 Survey organization

investigate and discuss the datasets and evaluation metrics commonly used to test the performance of the methods. In addition, we present the experimental results for several key techniques for topic derivation that can be used as baseline methods. Finally, based on the review and experimental results, we discuss potential future directions for topic derivation-related research in the Twitter environment.

3 Deriving topics from a collection of documents

In general, a topic can be defined as a set of stories linked by some real-world event [3]. Topics of a specific music festival in the town could include, for example, reviews of the musicians that will perform on the stage, prices of the tickets, or even security issues at the event. For a document collection, a topic is formally defined as a distribution over a fixed set of terms (vocabulary); each document in the collection is a mixture of a set of topics [10]. Thus, topic derivation of a document collection can be defined as the (unsupervised) task of characterizing the main topic of each document in the collection (cluster documents based on their main topics) and listing the most important keywords to represent each discovered topic.

The task of topic derivation from a collection of documents has long been studied. One of the earliest approach is *latent semantic analysis (LSA)* [26]. LSA takes advantage of the relationship between documents and terms represented in the term-document matrix by decomposing the matrix into its lower representation using the singular value decomposition (SVD) method.

Hofmann presented the extension of LSA called *probabilistic latent semantic analysis (PLSA)* [41] to deal with the different meanings and types of words. The study of [58] investigated the properties of a method for matrix decomposition called *nonnegative matrix factorization (NMF)*. The method is now widely adopted for various matrix dimensional reduction problems, including document clustering and system recommendations. Later, the study of [9] introduced the *latent Dirichlet allocation (LDA)*, currently considered as the state-of-the-art method in topic modeling. LDA is a fully generative method in which, like PLSA, a document is a mixture of topics. These four major methods in topic derivation (i.e., LSA, PLSA, NMF, and LDA) share the property to be able to find k number of latent features (topics) through a dimensional reduction process. Each method is discussed in turn in the next subsections.

3.1 Latent semantic analysis (LSA)

LSA [26] is a text mining approach to derive the latent semantic structure of a document collection. It was designed to deal with the inability of existing techniques to retrieve information taking account of conceptual content rather than just matching words to queries. In this work, Deerwester et al. [26] highlight two issues pertaining to words matching that penalize the precision of the result: *synonymy* and *polysemy*. Synonymy is described as the use of various words to refer to the same object. Polysemy is the fact that a word can have more than one meaning, or can refer to more than one object.

LSA uses the relationship between documents and all unique terms (vocabulary) from the document collection to take the conceptual content into account. It constructs a term-document matrix V and performs matrix decomposition on this matrix to derive k number

of latent structures. LSA utilizes singular value decomposition (SVD) [32, Chapter 9] to decompose the term-document matrix into its lower-dimensional representation.

In LSA, SVD is viewed as a method for inferring a set of indexing variables to determine the latent structures. LSA simplifies the SVD method by taking only the first k largest singular values, so that the matrices produced by the decomposition process are of rank k. The term-document matrix decomposition in LSA is formulated as follows:

$$V = TSD^{\mathrm{T}} \tag{1}$$

where V is the term-document matrix with the size of $t \times d$ (t is the number of unique terms in the document collection, and d is the number of documents). Matrices T and D^{T} represent the rank k lower-dimensional matrix V. $k \leq \min(t, d)$ is the number of expected latent structures. T and D^{T} have the size of $t \times k$ and $k \times d$, respectively. S is the diagonal matrix of singular values with the size of $k \times k$.

In a document collection, the latent structures derived by LSA can be referred to as topics. Since matrix V is the representation of the relationship between documents and the unique terms available in the document collection, the matrix T can be viewed as the representation of term-topic relationships. The matrix D^{T} then can be viewed as the representation of relationships between the topics and documents. LSA has been successfully implemented for various applications, including document clustering (e.g., [8,71]) and language modeling (e.g., [7,35]).

3.2 Probabilistic latent semantic analysis (PLSA)

PLSA [41] was introduced to improve the performance of LSA. PLSA is claimed to have a more solid statistical foundation than LSA and is defined as a generative data model. The LSA model employs the Frobenius norm approximation in its objective function to get the most optimal decomposition result, which allows for negative values on the main matrix. In contrast, PLSA employs the likelihood principle for its objective function, and the model only allows positive entries to optimize the 'bag-of-words'-based data modeling approach.

PLSA derives the statistical latent class model as a mixture decomposition model. For a document collection, the latent variables of the model can be considered as the topics. In PLSA, the probability of the co-occurrences between documents and words ($P(d, w)$) is generated independently as a mixture of conditionally multinomial distributions:

$$P(d, w) = P(d)P(w|d) \tag{2}$$

$$\text{where} \quad P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \tag{3}$$

In the above equations, w is a word in the vocabulary $W = \{w_1, \ldots, w_N\}$, and d is a document in a document collection $D = \{d_1, \ldots, d_M\}$. z is the unobserved variable in the latent class $Z = \{z_1, \ldots, z_K\}$. Figure 2 shows the plate notation representation of the PLSA model. From this plate notation, we can see the process of generating z as the latent variable from the multinomial topic distribution in document $P(z|d)$, and w which is drawn from the word-topic distribution $P(w|z)$.

3.3 Nonnegative matrix factorization (NMF)

Nonnegative matrix factorization (NMF) is a method to decompose a matrix into its lower-dimensional matrix representations. The method only allows positive values for all involved

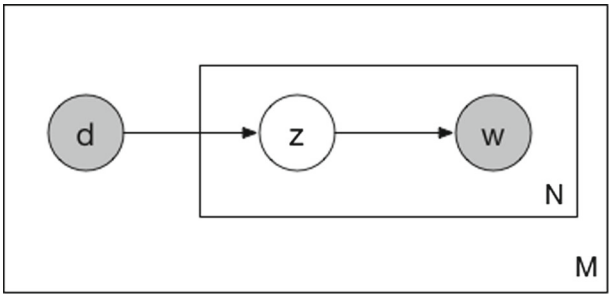


Fig. 2 PLSA model on a plate notation

matrices, including the decomposed matrix and the resulted matrices. NMF became popular after Lee and Seung [58] investigated two different multiplicative algorithms (least square error and Kullback–Leibler divergence) for the NMF implementation. NMF has been applied in numerous domains, including unsupervised clustering [38,51,53,105], recommendation systems [45,67,68,135,136], topic derivation [84,85,129], image processing [21,43,59,130], and bioinformatics [27,50,112].

For a document collection, NMF is able to uncover the hidden thematic structures of the collection by finding the factor matrices approximation for a document-term matrix. The document-term matrix represents the relationship of each document to every unique term in the document collection. The factorization process can be formulated as follows:

$$V \approx WH \tag{4}$$

Figure 3 illustrates the NMF method. Let the $V \in \mathbb{R}^{m \times n}$ be a document-term matrix with the size of $m \times n$ (m is the number of documents and n is the number of unique terms), the product of matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ is the approximation to the matrix V. In this process, rank $k < \min(m, n)$ can be considered as the number of expected latent topics. The main topic of each document can then be determined by choosing the maximum value of each vector in matrix W, and x number of keywords to represent each topic can be chosen by taking the $topx$ values from each vector in matrix H.

In NMF, the values of elements in the three matrices V, W, H are all positive. This nonnegativity feature is a useful constraint that allows only additive combination in the factorization process [59]. NMF is considered equivalent to PLSA method when the *Kullback–Leibler (KL) divergence* [54] is used as its objective function [33].

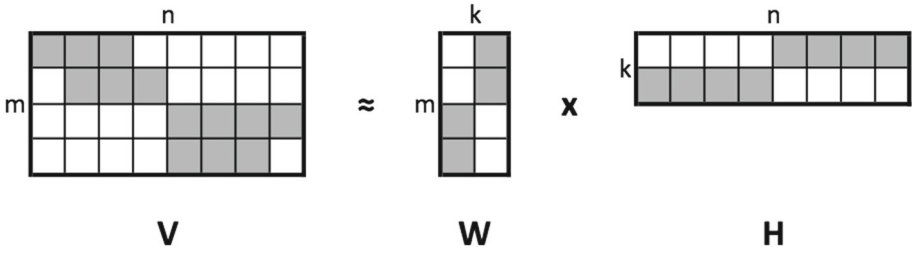


Fig. 3 Nonnegative matrix factorization process on document-term matrix $V \in \mathbb{R}^{m \times n}$ to derive the latent structures on factor matrices $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$

3.4 Latent Dirichlet allocation (LDA)

Blei et al. [9] introduced the LDA method: a generative probabilistic model for document collections. Like PLSA, in LDA, each document in the collection is modeled as a mixture of a set of latent topics. Blei et al. criticize the PLSA model to be not fully generative, as there is no generative probabilistic model for mixing the proportion of the latent variables, and thus, it becomes problematic for unseen documents (those which are outside the training set). In contrast, LDA uses Dirichlet prior for both the distribution of topics in the document collection and the distribution of words in every topic, making it fully generative and capable to infer topics from unseen documents. LDA has become very popular, and it is currently considered as the *state-of-the-art* method for topic derivation.

Figure 4 illustrates the generative process of LDA on a plate notation. In this figure, suppose we have M number of documents in a collection, α is the Dirichlet prior for the distribution of topics in document θ, and β is the Dirichlet prior for the distribution of words in topic ϕ with K being the number of the latent topics and N the number of words in the document. z is the topic assigned to word w in the current iteration. The generative process of LDA can be described as follows:

1. For each document in the collection, choose $\theta \sim \text{Dir}(\alpha)$.
2. For each topic, choose $\phi \sim \text{Dir}(\beta)$
3. For every word w_i in current document:
   (a) Choose a topic $z_i \sim \text{Multinomial}(\theta)$
   (b) Choose a word $w_i \sim \text{Multinomial}(\phi_{z_i})$

Mathematically, the probability of the LDA is formulated as follows:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{i=1}^{K} P(\phi_i; \beta) \prod_{j=1}^{D} P(\theta_j; \alpha) \prod_{t=1}^{N} P(z_{j,t}|\theta_j) P(w_{j,t}|\phi_{z_{j,t}}) \tag{5}$$

The original LDA model was based on the variational method and the expectation-maximization (EM) algorithm for Bayes parameter approximation. Later, Griffiths and

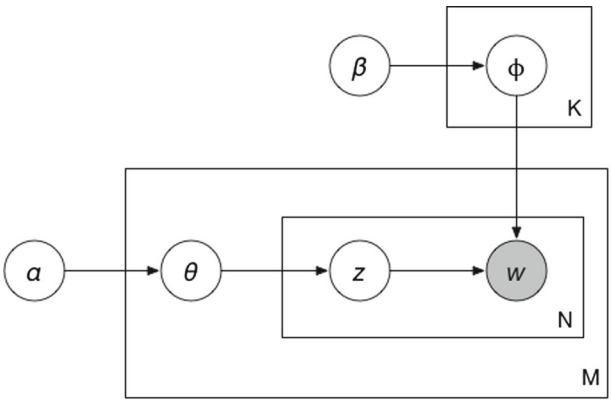


Fig. 4 LDA model on a plate notation

Steyvers introduced the use of the *Gibbs sampling* inference strategy as an alternative to the variational Bayes estimation [36]. The work shows that Gibbs sampling implementation on LDA model is simple and more efficient in memory in comparison with the previous approach.

Further studies show that in particular situations, LDA is equivalent to PLSA. The study in [34] presents the relationship between PLSA and LDA. It shows that PLSA is, in fact, a maximum likelihood (ML) estimated LDA model under a uniform Dirichlet prior. The work in [73] compares LDA and PLSA as a dimensionality reduction methods for the task of document clustering. It is found that both LDA and PLSA are far superior to a random projection. However, it did not find any meaningful difference between LDA and PLSA for a dimensionality reduction problem.

4 Deriving topics from the Twitter environment

The major techniques for topic derivation discussed in the previous section were mainly focused on uncovering various semantic relationships of words in documents. The methods have been applied and extended for many types of (lengthy) documents such as email [74,75,120], academic papers [11,28], and Web pages [64,99]. However, the Twitter environment poses new challenges. First is the severe sparsity of content. Posts are often very short and include many irrelevant characters (e.g., emoticons) and misspelled words. This can lead to an extremely low number of overlapping terms within a collection of Twitter posts (tweets). Defining text-based semantic relationships for topic identification thus becomes problematic. The next challenge is the dynamicity of the platform. With the speed of information propagation and a large number of incoming tweets, identifying topics on Twitter is a non-trivial task. A topic can quickly grow, decay, or even merge with another topic.

In this section, we review key studies that focus on deriving topics from Twitter. Most of these are still based on the major methods discussed in the previous section. Extensions have been proposed to take advantage of the unique features offered by social media to derive high-quality topics. As shown in Fig. 5, we classify the major features that are incorporated for topic derivation into three categories: *content, social, and temporal*. We discuss each feature and related existing methods in the subsections below.

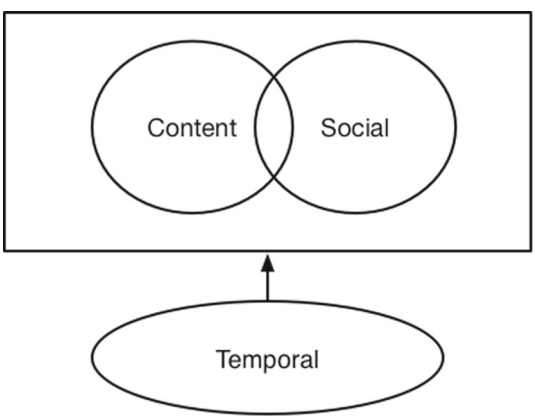


Fig. 5 Features incorporated for topic derivation in the Twitter environment

4.1 Focus on content exploitation

Despite the extreme sparsity of the posted messages, many studies still focus solely on the exploitation of content. Some have simply applied the methods described above. Others have built on them to include content merging (merge several or all tweets into a single entity), content expansion (expand every tweet with external resources), or various semantic relationships between terms in the collection. In this section, we review studies that primarily focus on content for topic derivation.

The direct application of the major topic derivation methods has shown a relatively good performance on specific Twitter datasets. The work in [94] successfully applies the LDA method to uncover topics in a public-health-related Twitter dataset. It uses a dataset based on tobacco-related terms such as '*smoking*,' '*tobacco*,' '*cigarette*,' '*cigar*,' '*hookah*,' and '*hooka*.' The study of [52] proposes a modified LDA method to identify topics from a disaster-related tweet collection. Instead of using an equal weight for the distribution of the topics in a document (θ), Kireyev et al. used a word's *specificity* weighting scheme, where more specific words have a higher weight in the topic assignment process to deal with the sparsity problem. The original LDA method is also used in the work of [121] to extract events from Twitter for an automatic crime prediction, focusing on hit-and-run cases. These methods share a common objective where they are aimed to derive targeted topics from Twitter. Predefined keywords are used to supervise the topic derivation process. However, in many cases, we might not know the incoming topics and thus are unable to infer particular keywords to help the process. In addition, the Twitter environment is dynamic and predefined keywords might change from time to time, making the approaches potentially failed to perform.

Other works find that merging the content from each tweet could have a positive impact to deal with the short content issue. In the work of Weng et al. [124], all tweets' contents are aggregated into a single big document to be processed with the original LDA method. The derived topics are then used as a factor for identifying influential Twitter users. Similarly, the study of [42] conducts an empirical study of topic modeling in Twitter using the LDA model and the LDA extension author–topic (AT) model [109]. The author–topic (AT) model considers the relationships between the authors of the posts and their topical distribution. Based on their experiments, the study finds that combining the content of tweets can improve the effectiveness of the trained topic models. It concludes that the performance of the standard LDA approach in the Twitter environment is better than its author–topic model extension. However, combining all tweets into a single document limits the ability to infer the topical cluster for each tweet.

Quite a few approaches employ various techniques for using external resources to tackle the sparsity problem. The works in [91,92] propose to convert an external knowledge base as an additional "universal dataset" to enhance the short content. Hu et al. [44] employ a hierarchical three-level structure that integrates multiple semantic knowledge bases such as *Wikipedia* and *WordNet*. Jin et al. [47] propose the *dual latent Dirichlet allocation (DLDA)* model to infer topics from Twitter data with the help of auxiliary lengthy datasets like *Wikipedia* content through a joint transfer learning process. DLDA derives the topics of the auxiliary data and the target data in two separate LDA processes and merges the results after filtering out the irrelevant topical knowledge. Since running a single LDA process itself has high computational complexity, integrating two LDA processes might increase the complexity exponentially.

The work in [66] expands the query using terms generated from *Freebase* as the knowledge base. Freebase is chosen as the main external resource as it consists of data harvested from various other sources like the Semantic Web and Wikipedia. Furthermore, the structure of

Freebase generally represents human knowledge. Yet other work [131] utilizes the title of articles in Wikipedia to represent the topic for every post in Twitter. In that work, topic identification relies mainly on the importance of words in the tweet collection calculated using the *TF-IDF* formula and the computation of words similarity from both the Twitter datasets and the Wikipedia title collection. The work in [96] presents better clustering results when bisecting the K-means algorithm in comparison with the simple TF-IDF. The study also finds that the word unigrams are the best feature to be incorporated in the clustering process.

Some studies adopt deep learning techniques for their topic modeling implementation. The work in [22] presents the topic-layer-adaptive stochastic gradient approach to jointly learn the simplex-constrained global parameters from all layers and topics. The parameters are then applied to the deep latent Dirichlet allocation (DLDA) method for topic modeling. The study in [133] presents the benefits of deep structures for learning word-topic distributions. It proposes a multilayer generative process on the word distributions of topic to discover interpretable topic hierarchies. A recent study [132] proposes a four-stage framework to extract hot topic from Twitter: data preprocessing, deep learning to enrich short text via image understanding, LDA to optimize the image effective word pairs, and last the integration of both text and images information. The improvement of deep learning performance over other techniques is still an open problem, especially in a streaming environment where the incoming data volume is big and the speed is high.

More recently, the increasingly popular word embedding techniques [77,78] have been applied for the task of topic derivation in Twitter. The study of [82] incorporates latent feature vector representations of words into two different Dirichlet multinomial topic models (LDA and Dirichlet multinomial mixture (DMM) [83]). The vector representation used is trained on very large corpora. The work in [60] also proposes a topic derivation method for short text like tweets using the help of auxiliary word embeddings. The Google News corpus with a vocabulary size of 3 million words is used for the word embedding learning purposes.

While incorporating external resources to deal with the sparsity issue in a short text environment looks promising, it might introduce extra burden to the system when dealing with the Twitter dynamic environment. The dynamicity of the environment with the high speed of incoming tweets, and the vast amount of noise make involving external resources not scalable for the task of online topic derivation.

A method proposed in [129] explores the correlation between terms for learning the topics from a sparse environment like Twitter. It reports that the correlation between terms in the document collection is much denser when compared to the generally used term-document relationship matrix. The term correlation matrix is considered to be more capable of capturing the latent structure for topic identification. Figure 6 shows the topic learning process proposed in that work. It employs a two-step matrix factorization process. The first step is the factorization of the term correlation symmetric matrix S to infer the term-topic matrix U. The second factorization is used to solve the topic-document matrix V by using the observed term-topic matrix U when factorizing the term-document matrix X. Experiments with the TREC2011[2] Twitter dataset and several other short-text type datasets show that the proposed method is able to outperform the state-of-the-art LDA model.

Similar to the work in [129], the study of Ma et al. [69] performs a factorization of the term correlation matrix to obtain the term-topic matrix as the first step of the topic derivation in a microblog environment. However, for the second step, instead of using matrix factorization approach, they employ the PLSA model on the term-topic matrix to infer the relationships

[2] http://trec.nist.gov/data/tweets/.

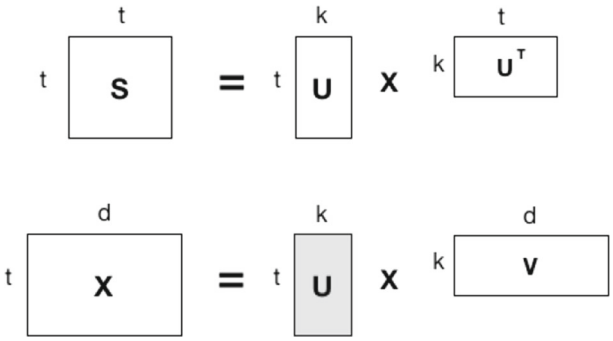


Fig. 6 Learning topics using a two-step matrix factorization process [129]. In the first step, the term-topic matrix U is inferred from the factorization of the term correlation matrix S. The observer term-topic matrix U is then used to learn the topic-document matrix V in the process of factorizing the term-document matrix X in the second process

between document and topic. Xu et al. [127] extend the *biterm topic model (BTM)* method, proposed in [18,128], to get the word co-occurrence pattern model as a parameter in the proposed semantically similar hashing method (SSHash). The hashing method provides fast and efficient matching techniques for mining semantically similar topics in short text environments. BTM directly models the co-occurrence of words patterns in the whole document collection to enhance the process of topic derivation. The work in [61] integrates the K-means algorithm into the BTM approach to derive topics from the dataset. First, BTM is applied to infer potential topics from the dataset, and next, K-means is used to get topic-based clusters.

The semantic relationship between words for topic modeling is also explored in [90,137]. The study of Ozdikis et al. implements semantic expansion techniques based on the statistics of co-occurrence words in a tweet collection [90]. The work in [137] proposes a word co-occurrence network-based model to deal with the sparsity problem. The method uses the sliding window technique to build the network of words where any two distinct words from a document occurring in the same window are considered as connected to each other. The resulted network of words is then turned into a pseudo-document set and processed with the Gibbs sampling for LDA [36] to observe the latent topics. Unfortunately, the performance of methods that focus only on word co-occurrence exploitation is still limited due to the extreme sparsity in the Twitter environment.

4.2 Incorporating social features

Unlike other types of short text (e.g., collections of titles, *RSS*, instant messages, image captions), social media platforms provide features to interact with other users or explicitly refer to events. *Mentions*, *replies*, *retweets*, and *hashtags* are some examples of social features that are popular among Twitter users. A *mention* is generally used to initialize a conversation with other users or to involve other users into the current discussion about a particular topic. Users can reply to or re-share someone's post. A hashtag is a specific term starting with '#' to tag the tweet. A hashtag in a tweet generally refers to a particular discussion, location, or event. Researchers find that exploiting such features along with the content can improve the quality of the derived topics in a social media environment.

Ramage et al. [100] address the problem of characterizing the information in microblogs with the help of a topic modeling approach. The work implements the *Labeled LDA* method

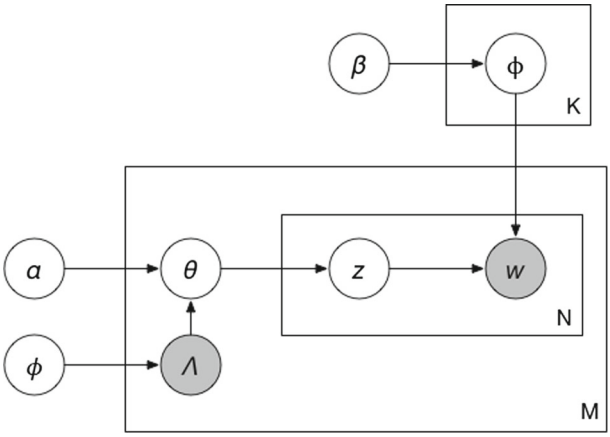


Fig. 7 Plate notation of Labeled LDA model [99,100]

[99] to analyze the content of Twitter posts. Labeled LDA is an extension of the LDA method that incorporates labels to partially supervise the learning process. *Hashtags*, *replies*, *@users*, and *emoticons* are used as predefined labels. Figure 7 shows the plate notation of the Labeled LDA model. The model assumes that each tweet will use only some labels from a set of labels Λ. It allows the modeling of a collection of tweets as a mixture of some labels and as a combination of latent topics as in the original LDA method. However, only small number of posts contains hashtags or other features to be used as predefined labels.

The work in [76] investigates methods to improve the original LDA when applied to Twitter. The paper proposes the combination of pooling scheme (to get more coherent input for the LDA learning process) and automatic topic labeling (to further improve the results of identified topics). Hashtags are used for both pooling the tweets to build the aggregated text and for labeling the derived topics automatically. The work in [93] incorporates another method for pooling to improve the input of LDA process, based on a community detection approach by aggregating content from groups of users who have common interests and interactions. Another pooling scheme is also introduced in [5]. It groups together the tweets that belong to specific conversation by analyzing the involved users and the replies of the tweet. The study in [104] proposes a community detection approach from Twitter environment by clustering users on their similar topical preferences based on the Louvain modularity.

Method to incorporate hashtags is also proposed in [37]. The incorporated hashtags are used as a specific feature of tweets along with external news entities to help extracting text-to-text correlations to enrich the short text data. The study of Wang et al. proposes a hashtag graph-based topic model to discover more distinct and coherent topics in Twitter [123]. In the work, a hashtag is used as a weakly supervised information point to model the topic. The work in [114] also uses hashtags to cluster short messages in general domains. The clustering process is broken down into two steps. The first step is to use a collection of hashtagged tweets to obtain stable clusters based on the hashtags. The clusters are then incorporated into the second step to do the clustering of tweets which mostly are not tagged.

Ma et al. [70] propose *tag-latent Dirichlet allocation* (TLDA), an extension of LDA that incorporates the observed hashtags as a mixture of topics into the process. The study of [117] proposes a unified framework that integrates social aspects and external resources as additional information to model the multifaceted topics in Twitter. The framework extracts

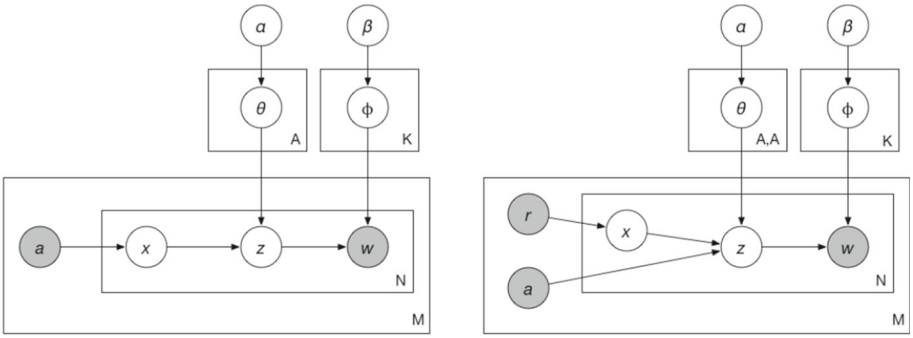


Fig. 8 Plate notation of **a** author–topic (AT) model [109] and **b** author–recipient–topic (ART) model [74]

all the hashtags from the tweet collection as social semantics, and it retrieves the top-n terms from the Web documents included as URL in tweets as the auxiliary semantics. Both social and auxiliary semantics are then used to enrich the content for the topic identification process.

Hashtags are often associated with users' interests in a particular topic. However, they might also be used to refer to something else. For example, a hashtag '#Sydney' might refer to a location instead of a topic. Furthermore, the number of tweets with hashtags is usually very low. Quite a few studies try to incorporate other social features to improve the quality of derived topics. Chierichetti et al. [20] investigate the behavior of tweets and retweets when a particular event is happening. The tweets and retweets form a "heartbeat" pattern that can be observed for event detection. Their work finds that looking only at the volume of tweets and retweets, an event being discussed on Twitter can be more accurately detected than with some baseline methods. Similar to hashtags, the volume of retweets is most likely will be very low. Relying only on this specific feature might not help much on the real-world topic.

Rajani et al. in [98] compare the application of the original LDA, the author–topic (AT) model [109], and the author–recipient–topic (ART) model [74] to extract topics from Twitter data. The ART model expands the AT model by incorporating the recipient of the posts. Both the AT and ART models are built on the original LDA model. Figure 8a, b shows the plate notations of the AT and ART models, respectively. In the AT model, each document is assumed to have a set of observed authors a. For every document, author x is sampled from the set of authors a, and the topic z is sampled from the distribution of the authors over topics θ. In the ART model, each document is assumed to have both sets of observed authors a and observed recipients r, and the process of topic sampling is influenced by both a and r. In Twitter, the author is the user who posts the tweet itself, and the recipient is the mentioned user in the tweet. The research finds that the ART model has the best performance, followed by the original LDA and AT, respectively. Unfortunately, ART model can only be applied to tweets that involve the mention feature as the recipient of the tweets.

The work in [95] proposes a behavior-topic model (B-LDA) to obtain topics from social media environments like Twitter. The proposed approach jointly models the users' social-based behavioral patterns and their interests in topics. Xia et al. propose another LDA extension (Plink-LDA) to incorporate links or similarity between documents to improve the quality of topic model [125]. In Twitter, the link between posts can be derived from hashtags or URLs. The link information is then used to control the topic sampling process along with the document collection itself.

Finally, the study in [86] proposes *intLDA*, an approach that directly incorporates the relationships between tweets into the topic derivation based on the LDA method. The tweets relationships are observed through both the social interactions (mentions, replies, and retweets) and content similarity. Later, the method was improved in [84] by introducing the weight of the relationships between tweets and uses the two steps of nonnegative matrix factorization process to incorporate the features when deriving topics from a static collection of tweets.

4.3 Incorporating the temporal aspect

In social media environments, users' posts continuously arrive as they are posted, and topics change rapidly. In Twitter, for example, a tweet posted by a user might not be about the same topic as the tweet posted by the same user several hours later. When a specific event happens, users tweets could be about the same topic during the time of the event, but the discussion can move quickly to various topics in periods when there are no major events. The fact that topics can change rapidly makes this environment very dynamic. Topic derivation methods applicable online (in real time) need to take the temporal aspect into account.

Lau et al. [57] propose a variant of the Online-LDA method [4,40]. In this model, new tweets are partitioned based on discretized time slices. The key difference between this approach and the Online-LDA method is that in Online-LDA, the vocabulary is assumed to be fixed. In the Lau et al.'s approach, the vocabulary is regenerated at each update by adding new incoming words and removing existing words with a frequency below a particular threshold. The method still depends solely on the content, which is sparse.

Saha and Sindhwani [103] introduce a variant of nonnegative matrix factorization method to work in an online environment like Twitter. Temporal regularization is used in the matrix factorization process to capture topics from the stream of incoming posts. The study in [122] presents temporal-LDA (TM-LDA), an extension of LDA to mine the text streams in social media. Specifically, TM-LDA learns the parameters for topic transitions dynamically when new messages arrive.

The work in [17] develops an incremental clustering framework to derive topics and characterize the emerging topics from the Twitter online environment. Starting with a crawling strategy to obtain more organized data, the proposed method employs temporal sequence features to detect the emerging topics in a semi-supervised way. The work in [29] proposes a nonparametric topics over time (npTOT) method to model the time-varying topics from a corpus that spans a long time period. The method employs Gibbs sampler based on the Chinese restaurant franchise approach [113]. The evaluation is conducted against a dataset of tweets obtained by the authors from January to March 2011, originating from Egypt. The study in [110] proposes the SAX* algorithm for discretizing a temporal series of terms to get the patterns of the collective attention to discover events in Twitter.

A number of studies focus on detecting bursty topics from the Twitter streaming environment. The study of Cataldi et al. proposes a real-time topic detection method, aimed especially to observe the most emergent topics in Twitter [15]. The approach includes the process of modeling the term life cycle according to the novel aging theory to automatically identify coherent topics across the different time intervals. The study in [126] proposes the TopicSketch, a framework to detect bursty topics in real time. It has a two-stage integrated solution. The first stage of the framework is used to continuously sketch statistical data related to word co-occurrence relationships. The second stage is used to infer the bursty topics based

on the statistical data sketch from the first stage. It uses the hash-based dimension reduction techniques to achieve the scalability of the process.

Methods that focus only on bursty topics often miss many important topics if there are no obvious burst captured. Recently, the work in [87] proposes a time-sensitive topic derivation approach to effectively capture the dynamic of topics from a streaming environment. The investigation shows that social features in Twitter, especially mentions, are sensitive to time. Tweets that mention the same users within a short period have a high chance to be about the same topic. The probability becomes exponentially smaller as the time posting difference increases. It suggests that time is an important factor in the task of topic derivation in the Twitter environment.

5 Datasets and evaluation metrics

Datasets and evaluation metrics are an integral component of performance measurement of topic derivation methods. In this section, we review several publicly available Twitter labeled datasets and metrics commonly used for topic derivation evaluation purposes.

5.1 Twitter datasets

To evaluate new methods, we require labeled datasets as ground truth. Each post in such datasets must be annotated with a label representing the topical membership in the collection. Several providers publish their labeled datasets, for example, *TREC* and *Sanders Analytics*. A number of researchers prefer to obtain and annotate a dataset by themselves to better represent their specific problems. In such cases, the annotations must meet a particular standard for an objective evaluation process.

5.1.1 TREC microblog datasets

This dataset is provided by *The Text REtrieval Conference* (TREC), a community co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. The TREC community regularly releases labeled datasets for research purposes, including a Twitter dataset with topics labels for the microblog track.[3] At the time of writing, the latest topic annotated dataset is the *TREC 2015 Microblog Track* [63]. It consists of more than 60,000 tweets posted between July 19, 2015, and July 30, 2015. Each of the tweets in this dataset has been annotated with one of 51 available topics. Some examples of the topics and their related tweets are shown in Table 2.

The previous version, TREC 2014, consists of more than 50,000 tweets posted between January 31, 2013, and March 31, 2013. This dataset is built based on the tweet timeline generation (TTG) task [62], which is to cluster relevant tweets ordered chronologically. It offers varying topics to represent the dynamics of the Twitter environment, where the number of tweets for each topic is changing over time. Each of the tweets in this dataset has been annotated for one of 55 available topics.

[3] http://trec.nist.gov/data/microblog.html, accessed April 24, 2019.

Table 2 Example of topics and their related tweets in the *TREC 2015* dataset

Code	Topic	Related tweets
MB226	Hershey, PA quilt show	We down here at Hershey????????????????????
		Tomorrow: We announce our 2015 #iHeartRadio Music Festival Lineup! Today: WIN THE FIRST TRIP TO THE SHOW!
		Sheepskin Saddle Comforter (Large Size for Western Saddle) http://goo.gl/T8NaSm
MB227	Pradaxa side effects	It seems that most of the side effects in the medicines we take are worse than the sicknesses we have!
		Holy crap my nose started bleeding and my face was full of blood
		Jwu... wo headache. The side effect of sleeping late xD
MB228	Coumadin dietary restrictions	Food for thought! All food is full of chemicals. http://m.huffpost.com/us/entry/55ad0ba1e4b0d2ded39f6b53 ...
		Food for thought... Abit of #WednesdayWisedom for you to ponder on before bed! self reflection... https://instagram.com/p/5vDSgSgCJQ/
		Please support this petition to support GMO food labelling...we have a right to know what we are eating!... http://fb.me/1R6okCqQC

Table 3 Examples of tweets for each topic in Sanders dataset

Topic	Tweet Code	Tweet
Apple	*a1*	Houston we have a problem!! My iPad has been restoring for 12+ hours after installing @apple IOS5. This can't be right....
	a2	hmmmm a lot of #siri feature don t work in canada location and direction seriously come on
	a3	#ios5 is nice and a it had to be thanks
Microsoft	*m1*	#Microsoft shows 'touch screen' for any surface | Nanotech - The Circuits Blog - CNET News http://cnet.co/oQKvoG via @cnet
	m2	Jus updated my computer to Windows 7 .....I'm on thanks to #microsoft
	m3	#Microsoft CEO Steve Ballmer on Not Buying #Yahoo: "Sometimes, You're Lucky" http://goo.gl/fb/KIrVu #uncategorized
Google	*g1*	#Android #Google Samsung and Google introduce GALAXY Nexus http://bit.ly/qfXlSU #DhilipSiva
	g2	The Samsung Galaxy Nexus and Ice Cream Sandwich are sick! #android #icecreamsandwich #google
	g3	Google is gonna need to do better than this to beat #iOS #Android #icecreamsandwich #Google http://youtu.be/android
Twiter	*t1*	62 Ways to Use #Twitter for Business: http://bit.ly/smbiz60 #tweets #socialmedia
	t2	My Facebook messed up and I had to make a new one so... add me! Haha at least twitter is reliable
	t3	My cute friend finally got a #twitter

Table 4 *Kappa* interpretation based on Landis and Koch [55]

Kappa value	Strength of agreement
< 0	Poor
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

5.1.2 Sanders Analytics

This dataset is available online and free to download.[4] It includes over 5500 tweets, each manually classified as belonging to one of four topics (Apple, Microsoft, Google, Twitter). The tweets are from more than 3000 different users. Table 3 shows the four topics and samples of tweets for each of them.

5.1.3 Self collected datasets

Twitter datasets available from providers like *TREC* and *Sanders Analytics*, although popular and used by many researchers, they might not be sufficient to represent specific problems addressed by researchers, such as disaster, traffic monitoring, or marketing. To get more varied tweets in different situations, researchers often need to obtain their own datasets using specific methods or queries. In such cases, two or more annotators are usually invited to label the collection. To get a good quality of dataset, a high inter-rater agreement should be achieved. The inter-rater agreement can be measured using the *Kappa coefficient* metric [31]. It measures the consistency of rating when several people assign a label to the same item. It is defined as:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{6}$$

$$\text{with,} \quad \overline{P} = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right)$$

$$\text{and,} \quad \overline{P}_e = \sum_{j=1}^{k} p_j^2 \tag{7}$$

where the quantity $1 - \overline{P}_e$ measures the degree of agreement attainable over and above what would be predicted by chance, and $\overline{P} - \overline{P}_e$ is the degree of agreement actually attained in excess of chance. In Eq. 7, $\overline{P}$ is the observed agreement, N is the total number of tweets, n is the number of annotators, k is the number of topics assigned to each tweet, and $\overline{P}_e$ is the mean proportion of agreement for agreement by chance. Table 4 shows the categorization of the kappa value based on Landis and Koch interpretation [55].

[4] https://github.com/zfz/twitter_corpus, accessed April 24, 2019.

5.2 Evaluation metrics

Topics are obtained from clusters of posts. To evaluate the derived topics, we use metrics appropriate to measure the quality of the clusters, with the labeled tweets as gold data. Major metrics include *purity*, *normalized mutual information (NMI)*, and *pairwise F-measure* [72].

5.2.1 Purity

Purity [134] evaluates the extent to which tweets are clustered in the same way as in the labeled datasets. The accuracy of the topic assignment is measured by the number of correctly assigned tweets divided by the total number of the labeled tweets in the dataset.

Let N be the number of labeled tweets in the gold standard, k the number of derived clusters, j the number of clusters in the gold standard. w_i is a cluster in the set of derived cluster W, and c_i is a cluster in the gold standard set C. The purity of W is defined to be:

$$\text{Purity}(W, C) = \frac{1}{N} \sum_{k} \max_{j} |w_i \cap c_j| \,. \tag{8}$$

The result of this metric ranges from 0 to 1. Low quality clustering has a purity value of 0, and a perfect clustering has a purity value of 1.

Using this metric, the perfect clustering value of 1 can be achieved regardless of the number of clusters. If the number of clusters (k) is the same as the number of the tweets (N), and each tweet gets its own cluster ($k = N$), the value of purity will be 1. A high purity value is easily achieved when the number of clusters is large [72]. Thus, the best way to evaluate the purity is using the same number of topics for both gold standard and the output of the algorithm.

5.2.2 Normalized mutual information (NMI)

Purity is a simple measure, but as explained above, a larger number of clusters tends to increase the purity value. To measure the trade-off between the quality of the clusters against the number of clusters, we employ *Normalized Mutual Information (NMI)* [111].

NMI measures the mutual information $I(W, C)$ shared between clusters W and the gold standard set C, normalized by the mean of the entropy of clusters $H(W)$ and classes $H(C)$. Similar to purity, the values of NMI range from 0 to 1, with larger the values of NMI meaning better clustering accuracy.

$$\text{NMI}(W, C) = \frac{I(W; C)}{[H(W) + H(C)]/2} \,. \tag{9}$$

In this metric, mutual information $I(W, C)$ quantifies the statistical information shared by the pair of clusters W and C [25], defined in Eq. 10.

$$I(W, C) = \sum_{k} \sum_{j} P(w_k \cap c_j) \log \frac{P(w_k \cap c_j)}{P(w_k) P(c_j)} \tag{10}$$

where k and j are the numbers of clusters in W and C, respectively. $P(w_k)$ is the probability of a tweet being in cluster w_k, $P(c_j)$ is the probability of a tweet being in cluster c_j, and $P(w_k \cap c_j)$ is the probability of a tweet being in both the cluster w_k and the gold standard

c_j. So, Eq. 10 is equivalent to Eq. 11 for maximum likelihood of the probabilities as the corresponding relative frequencies [72].

$$I(W, C) = \sum_k \sum_j \frac{|w_k \cap c_j|}{N} \log \frac{N|w_k \cap c_j|}{|w_k||c_j|} \tag{11}$$

where N is the total number of tweets in the gold standard, $|w_k|$ is the number of tweets in the cluster w_k, $|c_j|$ is the number of tweets in the cluster c_j, and $|w_k \cap c_j|$ is the number of tweets occurring in both the cluster w_k and the gold cluster c_j.

The minimum value of the mutual information $I(W, C)$ is 0, and the maximum is 1. This maximum value happens if clusters in W exactly recreate the gold standard C. However, it is reached also if the clusters in W, while recreating the gold, are further subdivided into smaller clusters. Thus, like purity, mutual information still faces a problem about the trade-off between the quality of the clusters and the number of clusters. To eliminate this bias, mutual information is normalized with the mean of the entropy of the clusters $H(W)$ and gold standards $H(C)$. Following [72], we use the arithmetic mean of $H(W)$ and $H(C)$ since $[H(W) + H(C)]/2$ is a tight upper bound on $I(W, C)$.

Entropy is a measure of uncertainty for a probability distribution [25]. Entropy $H(C)$ of a gold standard C is defined by:

$$H(C) = -\sum_j P(c_j) \log P(c_j) \tag{12}$$

Based on maximum likelihood estimates of the probabilities, Eq. 12 is equivalent to:

$$H(C) = -\sum_j \frac{|c_j|}{N} \log \frac{|c_j|}{N} \tag{13}$$

Similar to $H(C)$, the entropy $H(W)$ of a set of clusters W is defined by:

$$\begin{aligned} H(W) &= -\sum_k P(w_k) \log P(w_k) \\ &= -\sum_k \frac{|w_k|}{N} \log \frac{|w_k|}{N} \end{aligned} \tag{14}$$

5.2.3 *F*-Measure

As a final measure of the quality of the clustering result, we include the pairwise *F-Measure* metric [72] to compute the harmonic mean of precision P and recall R.

$$F = 2 \times \frac{P \times R}{P + R}\,. \tag{15}$$

where precision p is the fraction of pairs of tweets correctly put in the same cluster, and recall r is the fraction of actual pairs of tweets that were identified. Definition of both precision and recall is shown in Eqs. 16 and 17.

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

In this metric, TP (true positive) is the number of pairs of tweets from clusters in the gold standard which are correctly assigned to the same cluster in the output. TN (true negative) is the number of pairs of tweets from different clusters in the gold standard that are assigned to different clusters. The false positive (FP) is the number of pairs of tweets that should not be in the same cluster, but are assigned to the same cluster. False negative (FN) is the number of pairs of tweets that should be in the same cluster, but are assigned to different clusters.

Many researchers have long been using a likelihood-based metric like perplexity [118] to evaluate the quality of representative keywords for every topic learned from the derivation process. Perplexity is one of the most effective methods used to evaluate the log-likelihood of unseen documents. However, the investigation reported in [16] shows that this predictive likelihood approach does not address the goals of the topic model, and human judgment is often not being correlated to the perplexity.

To resemble the human judgment, topic coherence metrics are proposed for the automatic measurement of the interpretability of a topic based on its keywords representation. Two popular topic coherence measures are the extrinsic UCI measure [81] and the intrinsic UMass measure [79]. The extrinsic measure relies on external resources such as Wikipedia or Google 2-g to evaluate the coherence between words for each topic. The intrinsic approach by Mimno et al. [79] employs the pairwise function to evaluate the topics without collection reference from outside the dataset used in the topic derivation process. As it depends on the word co-occurrence relationship, the reliability for a sparse environment like Twitter can be very limited. Furthermore, the result is not symmetric and based on the order of the pair of words. (The first argument should be the rare word, followed by the common word as the second one.) Due to these limitations, the more subjective human judgment is still widely used. To automatically evaluate the quality of keywords representation for derived topics remains an open problem.

5.3 Experimental results

In this section, we present the results of our experiments as an example of using the datasets and evaluation metrics described above. We pick several key methods from the reviewed papers that represent different approaches. They include the original algorithms such as LDA and NMF, and their extensions that incorporate different features, such as TNMF [129] (exploiting the term correlation matrix), Plink-LDA [125] (incorporating social link between posts), and NMijF [87] (incorporating time-sensitive social features and content similarity).

Each experiment executes the topic derivation methods for k particular number of expected topics based on the labeled datasets. For every k and every method, we run the algorithms over both TREC 2015 and Sanders datasets 30 times and note the average value of each evaluation metric for comparison. All parameters for the methods are chosen to achieve the best performance for topic derivation in a Twitter environment based on the original papers. We applied typical data preprocessing to all datasets such as removing irrelevant terms or characters (stop-words, punctuations, emoticons, and terms with less than three characters) and stemming each term using the NLTK python package.[5]

The TREC 2015 dataset was designed to evaluate techniques to monitor the stream of posts from social media [63]. To evaluate the performance of the methods with respect to the dynamic nature of this social media environment, we divide the TREC 2015 dataset into a series of time periods (6-h interval) from July 20, 2015, to July 28, 2015. The number of tweets varies from 991 to 2240 in each interval. There are 36 intervals in total.

[5] http://www.nltk.org.

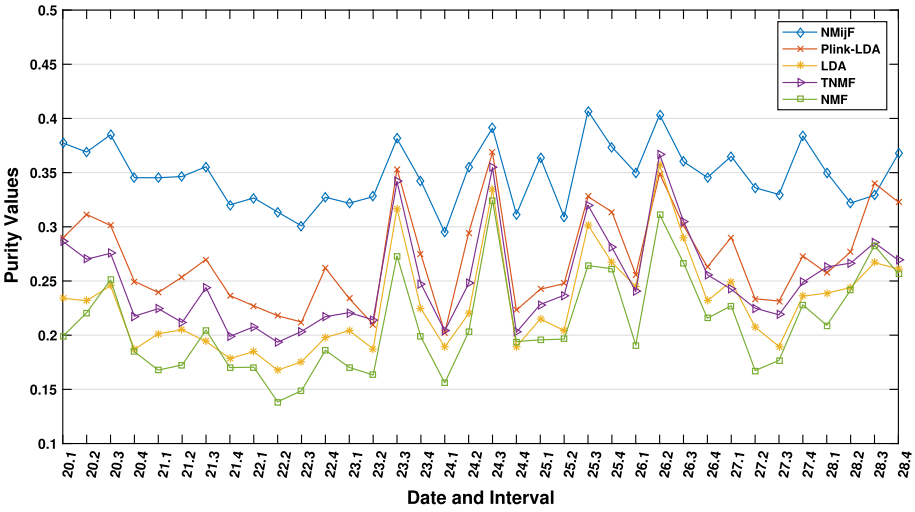


Fig. 9 Purity results for TREC 2015 dataset

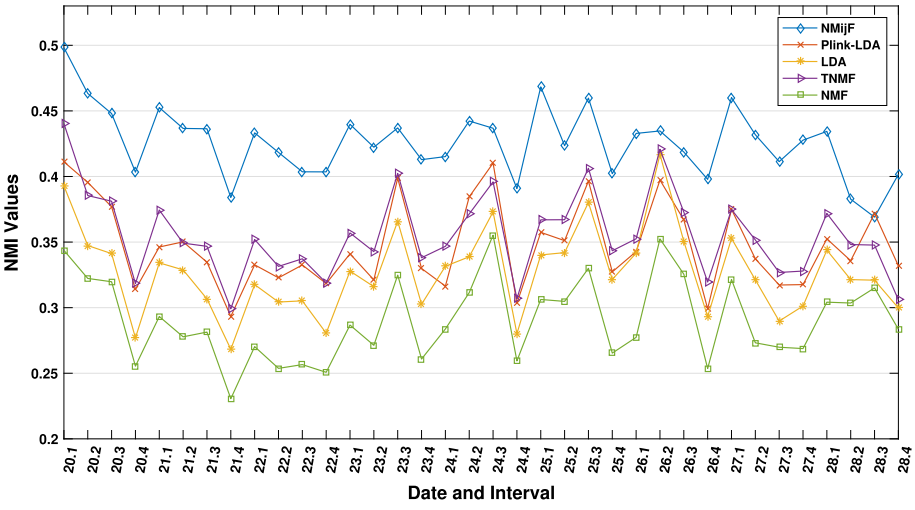


Fig. 10 NMI results for TREC 2015 dataset

Figure 9 shows the purity results for the experiment with TREC 2015 dataset. In this figure, the NMijF presents the best performance for most of the intervals with around 5–70% improvements over the other methods, followed by TNMF, Plink-LDA, LDA, and NMF. In specific cases, where there are prevalent topics discussed in the intervals, most methods can achieve high purity value. For example, in the interval 23.3, the majority of tweets in that time period discuss the topic MB383 (*Online dating for older woman*) and MB226 (*Quilt show being held in Hershey, PA*). In the interval 24.3, most of the tweets talk about the topic MB405 (*Experience about Rotterdam Unlimited events*) and MB409 (*Issues related to the airport TSA screening*).

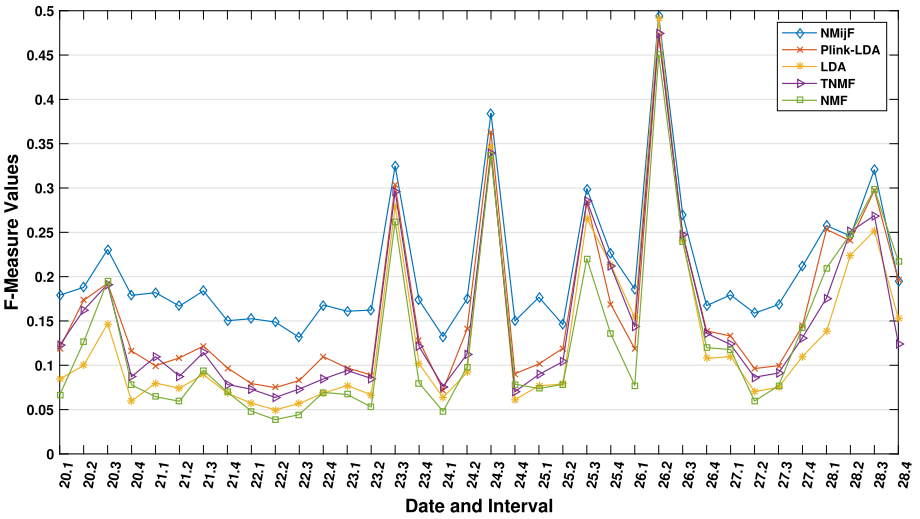
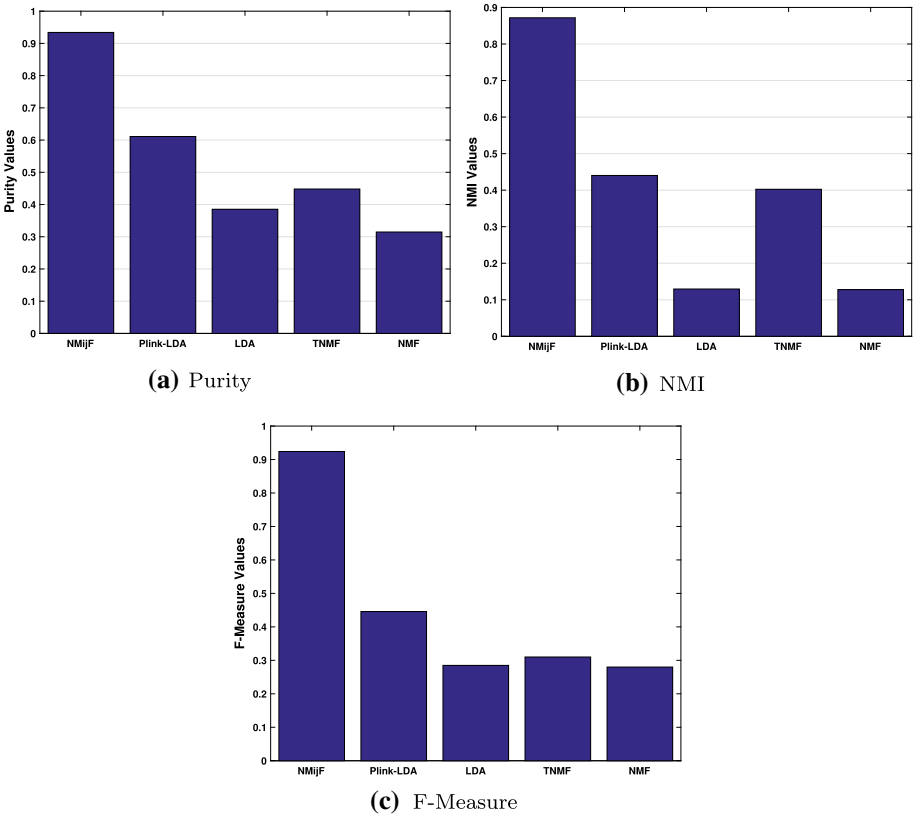


Fig. 11 F-Measure results for TREC 2015 dataset

The evaluation using NMI and F-measure for TREC 2015 are presented in Figs. 10 and 11, respectively. The NMI results in Fig. 10 show similar trends to the purity evaluation shown in Fig. 9, with NMijF having the best performance, followed by TNMF, Plink-LDA, and LDA. NMF remains in the last position in these experimental results. For the F-measure evaluation results, as shown in Fig. 11, all methods perform well for the intervals with prevalent topics, such as in 23.3, 24.3, 25.3, and 26.2. The highest F-measure values for all methods are achieved in interval 26.2, where the majority of tweets discuss the topic MB401 (*The television show "Knock Knock Live"*).

Our experiments with the Sanders dataset further show the importance of incorporating different features to deal with the sparsity problem. In these experiments, the Sanders dataset is not broken down into intervals as it contains tweets from only a very short period. Figure 12 shows the results of the purity evaluation (Fig. 12a), NMI (Fig. 12b), and F-measure (Fig. 12c). In these evaluations, we see that the methods that consider multiple features are always able to provide better results compared to those which only consider the content of the posts. Experiments with the TREC and Sanders datasets evaluate several aspects of deriving topics in the Twitter environment. First, the TREC dataset is used to test the performance of the methods with regard to the dynamicity of Twitter environment, including the sensitivity to time, sparsity issues, and diversity of topics. Second, the Sanders dataset is used to test the capability of the methods to deal with many overlapping words spread in all topics. In all cases, experimental results show that NMijF consistently performs the best, according to all the evaluation metrics.

We now look at the complexity of NMijF, to see if the superior performance comes at a computational cost. The computational complexity of NMijF in each of the multiplicative update rules is $\mathcal{O}(mnk)$ for every iteration, where m is the number of processed tweets, n is the number of unique terms from the tweets, and k is the number of derived topics. This complexity is similar to the TNMF- and the LDA-based method. With the same computational complexity, NMijF requires only 30 iterations to achieve the best result, while the other methods need at least 50–100 iterations to get the optimal results.



(a) Purity **(b)** NMI **(c)** F-Measure

Fig. 12 Purity, NMI, and F-measure evaluation results for sanders dataset

NMijF incorporates the relationships between tweets defined from the social interactions (user mentions, replies, and retweets) and content similarity. The time sensitivity of the interactions is also considered. Plink-LDA uses a similar approach to incorporate the links between tweets into the LDA process, but without considering the time aspect. In contrast, TNMF also observes the relationships between terms within dataset and builds the term correlation matrix using positive point mutual information (PPMI) [129]. Both LDA and NMF focus on the simple content exploitation of the social media posts. Incorporating all three main features (content, social interactions, and temporal), as in NMijF, can consistently improve the quality of derived topics.

6 Discussion

LSA, PLSA, NMF, and LDA are the major techniques to derive topics from a document collection in an unsupervised way. These methods work solely on the document content. A lot of extensions have been proposed to address the sparsity issue that arises in social media environments. Extensions include: incorporating more text, adding social features, and taking the temporal aspect into account. Some examples of the recent works in topic derivation and their incorporated features are summarized in Table 5. Based on our review of the current studies on topic derivation in the Twitter environment, we observe the following:

Table 5 Summary of the recent methods and their incorporated features for topic derivation from social media

Works	Techniques	Features		
		Content	Social	Temporal
Prier et al., Kireyev et al., Wang et al.	Straight LDA	✓		
Weng et al., Hong and Davison	Merging tweet content	✓		
Phan et al., Phan et al., Hu et al., Jin et al., Lv et al., Yıldırım et al., Nguyen et al., Li et al.	Incorporate external resources or word embedding	✓		
Yan et al., Ma et al., Cheng et al., Yan et al., Xu et al., Li et al., Ozdikis et al., Zuo et al.	Exploit correlation between terms	✓		
Ramage et al.	Social features as labels	✓	✓	
Mehrotra et al., Prateek and Vasudeva	Pooling scheme	✓	✓	
Guo et al., Wang et al., Tsur et al., Ma et al.	Incorporate hashtags	✓	✓	
Vosecky et al.	Incorporate hashtags and external resources (URLs)	✓	✓	
Chierichetti et al.	Incorporate retweets	✓	✓	
Rajani et al., Cheong and Cheong, Liu et al., Rosen-Zvi et al.	Incorporate authors and/or recipients of the posts	✓	✓	
Qiu et al.	Models the users' social-based behavioral pattern and their interests in topics	✓	✓	
Xia et al., Nugroho et al., Nugroho et al., Nugroho et al.	Incorporate link between posts	✓	✓	
Cataldi et al., Xie et al.	Bursty topics detection	✓		✓
Lau et al., Wang et al.	Variant of LDA for streamed data	✓		✓
Saha and Sindhwani	Incorporate temporal regularization	✓		✓
Chen et al.	Incremental clustering framework	✓		✓
Dubey et al.	Gibbs sampler	✓		✓
Stilo and Velardi	Discretizing the temporal series of terms	✓		✓
Nugroho et al.	Incorporate time-sensitive social interactions and content similarity	✓	✓	✓

- Methods that rely entirely on the tweet content still suffer from the sparsity issue. The density of the co-occurrence of terms matrix in a tweet collection can be as low as 0.274% on average [87]. With these very low rates of overlapping terms, exploiting various semantic relationships to derive topics solely from internal content is less likely effective for providing significant improvements over the state-of-the-art methods.
- Augmenting the short text data with auxiliary content from external resources seems to be a promising solution. However, the newly added terms inferred from the resources often include noise and are often unrelated to the context. It thus can be harmful to the learning process [13]. The informal language used in tweets, with a lot of misspelled words and abbreviations, can itself be very challenging for matching with the auxiliary content. Furthermore, relying on external resources faces scalability issues, as it could bring an extra burden when dealing with a highly dynamic environment like social media.
- Most methods that incorporate social features still focus on content-based interactions such as hashtags. Hashtags are often used by users to participate in discussions for a particular topic. However, they are still part of the tweet content, and most of the tweets do not include hashtags. The methods thus still suffer from the sparsity issue. Some methods try to include the tweets author and/or recipients. However, unlike in specific types of documents such as academic papers or news articles, where authors have a strong relationship with topics, in Twitter, a tweet is authored by only one user, and a user can post tweets in various topics. Furthermore, if a method requires recipients information to be available for the learning process, the method will not be suitable for the majority of the tweets, as most of them do not contain user mentions.
- In a highly dynamic environment like Twitter, time is an important feature to deal with varying topics, especially in real time. Most methods that incorporate temporal features still view the time aspect as a time slicing window to specify the interval of the serial or incremental learning process over time. Time aspect is not yet seen as a factor that can improve the quality of topic derivation for a static document collection.
- Due to the extreme sparsity of correlation between terms, a statistical analysis of the coherency between words in the topic representation might not give a reliable result for evaluation purposes. Researchers often do a qualitative analysis to evaluate the topical keywords. More advanced topic coherence measures that can deal with the extreme sparsity are required to evaluate the topical keywords in a more objective manner.

According to the experimental results presented in Sect. 5.3, the best performance is consistently achieved by the method that incorporates more features, including the social interactions and time sensitivity. Combined with semantic relationships of tweets content, more complex social interaction features need to be examined to deal with the sparsity issue. Moreover, the temporal aspect in Twitter should be considered as an important factor even in an offline situation. The relationships between time and the interaction features should be investigated to make sure that the proposed method can also handle the dynamic environment, both for static collections of tweets or for analysis in real time.

Topic derivation methods often require the number of topics as one of the input parameter. In the experiments, number of topics from the labeled dataset is used for this purpose. However, in the implementation level, especially in an online environment, the system should be able to dynamically pick the best number of topics for the current dataset. Much research uses social signals such as hashtags to determine the number of topics. However, this should be improved as hashtags might not necessarily represent the topics.

The purity, NMI, and F-measure metrics are used to evaluate the quality of derived topics based on one aspect: the accuracy of the unsupervised clustering results. The other aspect,

which is the quality of keywords learned to represent each topic, is often evaluated manually. Due to the extremely low co-occurrences between terms, an intrinsic statistical analysis of the coherency between words, like in [79], for the topic representation might not be reliable for different runs and methods. More robust metrics to evaluate the coherence of the most important words for every topic is on demand for the automatic evaluation.

7 Conclusions

In this paper, we looked at the task of topic derivation in Twitter and presented a review of key techniques and features used to improve the quality of the derived topics. We first provided insights into why Twitter is an important source of data for topic derivation work and why deriving topics in this platform is challenging. We then reviewed the popular and state-of-the-art methods to derive topics in a document collection, followed by a review of key studies. We classify the literature based on the features incorporated for topic derivation (i.e., *content*, *social interactions*, and *temporal aspect*)

Different from topic derivation in traditional documents with lengthy content, deriving topics from Twitter is challenging due to the short and unstructured content, and the dynamics of the environment itself. However, Twitter provides features that enable users to interact with other users or discussed events, for example: *mention, reply, retweet,* and *hashtag*. These social interaction features often show the sign of interests of discussion about a particular topic. They can help to deal with the sparsity issue where the frequency of term co-occurrences is extremely low.

The Twitter environment is also highly dynamic. Users' posts continuously arrive in real time, which makes the task of topic derivation more challenging than for other platforms. Time becomes an important aspect, as a message posted by a user might not be about the same topic as a message posted by the same user several hours later. Social interactions could also be sensitive to the temporal aspect. A topic can quickly disappear, tip, or evolve to another topic over the time. Methods that can effectively incorporate all of important features (i.e., *content*, *social*, and *temporal*) to improve the quality of topic derivation remain on demand.

Acknowledgements This work is partially supported by the CSIRO Data61, Macquarie University, Soegijapranata Catholic University, The Australian Research Council LP120200231, and The Australian Research Council DP140101369.

References

1. Aggarwal C, Subbian K (2014) Evolutionary network analysis: a survey. ACM Comput Surv 47(1):10:1–10:36. https://doi.org/10.1145/2601412
2. Alghamdi R, Alfalqi K (2015) A survey of topic modeling in text mining. Int J Adv Comput Sci Appl (IJACSA) 6(1):147–153
3. Allan J (2002) Topic detection and tracking: event-based information organization, vol 12. Springer, Berlin
4. AlSumait L, Barbarà D, Domeniconi C (2008) On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 2008 eighth IEEE international conference on data mining. pp 3–12. https://doi.org/10.1109/ICDM.2008.140
5. Alvarez-Melis D, Saveski M (2016) Topic modeling in twitter: aggregating tweets by conversations. In: Tenth international AAAI conference on web and social media
6. Atefeh F, Khreich W (2015) A survey of techniques for event detection in twitter. Comput Intell 31(1):132–164

7. Bellegarda JR (1998) Exploiting both local and global constraints for multi-span statistical language modeling. In: Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing, 1998. vol 2, pp 677–680. https://doi.org/10.1109/ICASSP.1998.675355
8. Bellegarda JR, Butzberger JW, Chow YL, Coccaro NB, Naik D (1996) A novel word clustering algorithm based on latent semantic analysis. In: Proceedings of the 1996 IEEE international conference on acoustics, speech, and signal processing, 1996. ICASSP-96. Conference proceedings. vol 1, pp 172–175. https://doi.org/10.1109/ICASSP.1996.540318
9. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
10. Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84
11. Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat 1:17–35
12. Campos R, Dias G, Jorge AM, Jatowt A (2014) Survey of temporal information retrieval and related applications. ACM Comput Surv 47(2):15:1–15:41. https://doi.org/10.1145/2619088
13. Cao G, Nie JY, Gao J, Robertson S (2008) Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp 243–250
14. Carlson N (2011) The real history of Twitter. http://www.businessinsider.com.au/how-twitter-was-founded-2011-4. Online, Accessed 6 Oct 2016
15. Cataldi M, Di Caro L, Schifanella C (2010) Emerging topic detection on Twitter based on temporal and social terms evaluation. In: Proceedings of the tenth international workshop on multimedia data mining. ACM, New York, NY, USA, MDMKDD '10, pp 4:1–4:10. https://doi.org/10.1145/1814245.1814249
16. Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM (2009) Reading tea leaves: how humans interpret topic models. In: Proceedings of the 22nd international conference on neural information processing systems (NIPS'09). Curran Associates Inc., Red Hook, NY, USA, pp 288–296
17. Chen Y, Amiri H, Li Z, Chua TS (2013) Emerging topic detection for organizations from microblogs. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, SIGIR '13, pp 43–52. https://doi.org/10.1145/2484028.2484057
18. Cheng X, Yan X, Lan Y, Guo J (2014) BTM: Topic modeling over short texts. Trans Knowl Data Eng 26(12):2928–2941. https://doi.org/10.1109/TKDE.2014.2313872
19. Cheong F, Cheong C (2011) Social media data mining: a social network analysis of tweets during the Australian 2010-2011 floods. In: Proceedings of the 15th Pacific Asia conference on information systems (PACIS). Queensland University of Technology, pp 1–16
20. Chierichetti F, Kleinberg JM, Kumar R, Mahdian M, Pandey S (2014) Event detection via communication pattern analysis. In: ICWSM
21. Cichocki A, Zdunek R, Amari Si (2006) New algorithms for non-negative matrix factorization in applications to blind source separation. In: Proceedings of the 2006 IEEE international conference on acoustics speech and signal processing proceedings. IEEE, vol 5, pp V–V
22. Cong Y, Chen B, Liu H, Zhou M (2017) Deep latent dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In: Proceedings of the 34th international conference on machine learning-volume 70, JMLR. org, pp 864–873
23. Cordero-Gutiérrez R, de la Prieta-Pintado F, Corchado-Rodríguez JM (2018) Decision support for digital marketing through virtual organizations-influencers on twitter. In: International conference on knowledge management in organizations. Springer, pp 574–585
24. Council DP (2015) Researh note: world leader ranking on Twitter. http://www.digitaldaya.com/admin/modulos/galeria/pdfs/73/161_o59ontgs.pdf. Online, Accessed 6 Oct 2016
25. Cover TM, Thomas JA (2012) Elements of information theory. Wiley, Hoboken
26. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391
27. Devarajan K (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol 4(7):e1000029
28. Dietz L, Bickel S, Scheffer T (2007) Unsupervised prediction of citation influences. In: Proceedings of the 24th international conference on machine learning. ACM, New York, NY, USA, ICML '07, pp 233–240. https://doi.org/10.1145/1273496.1273526
29. Dubey A, Hefny A, Williamson S, Xing EP (2013) A nonparametric mixture model for topic modeling over time. In: Proceedings of the 2013 SIAM international conference on data mining. pp 530–538, https://doi.org/10.1137/1.9781611972832.59
30. Edwards M, Rashid A, Rayson P (2015) A systematic survey of online data mining technology intended for law enforcement. ACM Comput Surv 48(1):15:1–15:54. https://doi.org/10.1145/2811403
31. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378

32. Forsythe GE, Moler CB, Malcolm MA (1977) Computer methods for mathematical computations. Prentice-Hall, Upper Saddle River
33. Gaussier E, Goutte C (2005) Relation between PLSA and NMF and implications. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 601–602
34. Girolami M, Kabán A (2003) On an equivalence between PLSI and LDA. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. ACM, pp 433–434
35. Gotoh Y, Renals S (1997) Document space models using latent semantic analysis
36. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235
37. Guo W, Li H, Ji H, Diab MT (2013) Linking tweets to news: a framework to enrich short text data in social media. In: Proceedings of the 2013 association for computational linguistics conference. Citeseer, Sofia, Bulgaria, pp 239–249
38. He Z, Xie S, Zdunek R, Zhou G, Cichocki A (2011) Symmetric nonnegative matrix factorization: algorithms and applications to probabilistic clustering. IEEE Trans Neural Netw 22(12):2117–2131
39. Hermida A, Lewis SC, Zamith R (2014) Sourcing the arab spring: a case study of andy carvin's sources on Twitter during the Tunisian and Egyptian revolutions. J Comput Med Commun 19(3):479–499
40. Hoffman MD, Blei DM, Bach F (2010) Online learning for latent Dirichlet allocation. In: Proceedings of the 23rd international conference on neural information processing systems (NIPS'10), vol 1. Curran Associates Inc., Red Hook, NY, USA, pp 856–864
41. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp 50–57
42. Hong L, Davison BD (2010) Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics. ACM, New York, NY, USA, SOMA '10, pp 80–88. https://doi.org/10.1145/1964858.1964870
43. Hoyer PO (2004) Non-negative matrix factorization with sparseness constraints. J Mach Learn Res 5(Nov):1457–1469
44. Hu X, Sun N, Zhang C, Chua TS (2009) Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: Proceedings of the 18th ACM conference on information and knowledge management. ACM, New York, NY, USA, CIKM '09, pp 919–928. https://doi.org/10.1145/1645953.1646071
45. Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on Recommender systems. ACM, pp 135–142
46. Jelisavčić V, Furlan B, Protić J, Milutinović V (2012) Topic models and advanced algorithms for profiling of knowledge in scientific papers. In: MIPRO, 2012 Proceedings of the 35th international convention. pp 1030–1035
47. Jin O, Liu NN, Zhao K, Yu Y, Yang Q (2011) Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings of the 20th ACM international conference on information and knowledge management. ACM, New York, NY, USA, CIKM '11, pp 775–784. https://doi.org/10.1145/2063576.2063689
48. Joshi A, Sparks R, McHugh J, Karimi S, Paris C, MacIntyre RC (2019) Harnessing tweets for early detection of an acute disease event. Epidemiology 31:90–97
49. Karimi S, Wang C, Metke-Jimenez A, Gaire R, Paris C (2015) Text and data mining techniques in adverse drug reaction detection. ACM Comput Surv 47(4):56:1–56:39. https://doi.org/10.1145/2719920
50. Kim H, Park H (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23(12):1495–1502
51. Kim J, Park H (2008) Sparse nonnegative matrix factorization for clustering
52. Kireyev K, Palen L, Anderson K (2009) Applications of topics models to analysis of disaster-related Twitter data. In: Proceedings of the NIPS workshop on applications for topic models: text and beyond. Whistler, Canada, vol 1
53. Kuang D, Park H, Ding C (2012) Symmetric nonnegative matrix factorization for graph clustering. In: Proceedings of the 2012 SIAM international conference on data mining. SIAM, California, USA, vol 12, pp 106–117
54. Kullback S (1997) Information theory and statistics. Courier Dover Publications, Mineola
55. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
56. Larsen ME, Boonstra TW, Batterham PJ, O'Dea B, Paris C, Christensen H (2015) We feel: mapping emotion on twitter. IEEE J Biomed Health Inform 19(4):1246–1252

57. Lau JH, Collier N, Baldwin T (2012) On-line trend analysis with topic models:\# Twitter trends detection topic model online. In: Proceedings of the 24th international conference on computational linguistics. Mumbai, India, pp 1519–1534
58. Lee D, Seung H (2000) Algorithms for non-negative matrix factorization. In: Proceedings of the advances in neural information processing systems 13 (NIPS 2000). Denver, CO, USA, pp 556–562
59. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791
60. Li C, Wang H, Zhang Z, Sun A, Ma Z (2016) Topic modeling for short texts with auxiliary word embeddings. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, SIGIR '16, pp 165–174. https://doi.org/10.1145/2911451.2911499
61. Li W, Feng Y, Li D, Yu Z (2016) Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm. Autom Control Comput Sci 50(4):271–277
62. Lin J, Efron M, Wang Y, Sherman G (2014) Overview of the trec-2014 microblog track. Tech. rep., NIST. http://trec.nist.gov/pubs/trec23/trec2014.html
63. Lin J, Efron M, Wang Y, Vorhees EM (2014) Overview of the trec-2015 microblog track. Tech. rep., NIST. http://trec.nist.gov/pubs/trec24/trec2015.html
64. Liu Y, Niculescu-Mizil A, Gryc W (2009) Topic-link LDA: Joint models of topic and author community. In: Proceedings of the 26th annual international conference on machine learning. ACM, New York, NY, USA, ICML '09, pp 665–672. https://doi.org/10.1145/1553374.1553460
65. López-Sáncez D, Revuelta J, de la Prieta F, Corchado JM (2018) Towards the automatic identification and monitoring of radicalization activities in twitter. In: International conference on knowledge management in organizations. Springer, pp 589–599
66. Lv C, Qiang R, Fan F, Yang J (2015) Proceedings of the information retrieval technology proceedings : 11th asia information retrieval societies conference, airs 2015, brisbane, qld, australia, december 2–4 (2015). Springer, Cham, pp 43–55
67. Ma H, Yang H, Lyu MR, King I (2008) Sorec: social recommendation using probabilistic matrix factorization. In: Proceedings of the 17th ACM conference on Information and knowledge management. ACM, pp 931–940
68. Ma H, Zhou D, Liu C, Lyu MR, King I (2011) Recommender systems with social regularization. In: Proceedings of the fourth ACM international conference on Web search and data mining. ACM, pp 287–296
69. Ma HF, Sun YX, Jia MHZ, Zhang ZC (2014) Microblog hot topic detection based on topic model using term correlation matrix. In: Proceedings of the 2014 international conference on machine learning and cybernetics. vol 1, pp 126–130. https://doi.org/10.1109/ICMLC.2014.7009104
70. Ma Z, Dou W, Wang X, Akella S (2013) Tag-Latent Dirichlet Allocation: Understanding hashtags and their relationships. In: 2013 IEEE/WIC/ACM international joint conferences on proceedings of the web intelligence (WI) and intelligent agent technologies (IAT). vol 1, pp 260–267
71. Maletic JI, Valluri N (1999) Automatic software clustering via latent semantic analysis. In: Proceedings of the 14th IEEE international conference on automated software engineering. pp 251–254. https://doi.org/10.1109/ASE.1999.802296
72. Manning C, Raghavan P, Schütze H (2008) Introduction to information retrieval, vol 1. Cambridge University Press, Cambridge
73. Masada T, Kiyasu S, Miyahara S (2008) Comparing LDA with pLSI as a dimensionality reduction method in document clustering. In: Large-scale knowledge resources. Construction and application. Springer, pp 13–26
74. McCallum A, Corrada-Emmanuel A, Wang X (2005) The author–recipient–topic model for topic and role discovery in social networks: experiments with enron and academic email
75. McCallum A, Wang X, Corrada-Emmanuel A (2007) Topic and role discovery in social networks with experiments on enron and academic email. J Artif Intell Res 30:249–272
76. Mehrotra R, Sanner S, Buntine W, Xie L (2013) Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, SIGIR '13, pp 889–892. https://doi.org/10.1145/2484028.2484166
77. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
78. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems (NIPS'13), vol 2. Curran Associates Inc., Red Hook, NY, USA, pp 3111–3119

79. Mimno D, Wallach H, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), association for computational linguistics. pp 262–272
80. Myers SA, Sharma A, Gupta P, Lin J (2014) Information network or social network?: The structure of the Twitter follow graph. In: Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, pp 493–498
81. Newman D, Noh Y, Talley E, Karimi S, Baldwin T (2010) Evaluating topic models for digital libraries. In: Proceedings of the 10th annual joint conference on digital libraries. ACM, New York, NY, USA, JCDL '10, pp 215–224. https://doi.org/10.1145/1816123.1816156
82. Nguyen DQ, Billingsley R, Du L, Johnson M (2015) Improving topic models with latent feature word representations. Trans Assoc Comput Linguist 3:299–313
83. Nigam K, McCallum AK, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using em. Mach Learn 39(2–3):103–134
84. Nugroho R, Yang J, Zhong Y, Paris C, Nepal S (2015) Deriving topics in Twitter by exploiting tweet interactions. In: Proceedings of the 2015 IEEE international congress on big data. pp 87–94. https://doi.org/10.1109/BigDataCongress.2015.22
85. Nugroho R, Zhong Y, Yang J, Paris C, Nepal S (2015) Matrix inter-joint factorization—a new approach for topic derivation in Twitter. In: Proceedings of the 2015 IEEE international congress on big data. pp 79–86. https://doi.org/10.1109/BigDataCongress.2015.21
86. Nugroho R, Molla-Aliod D, Yang J, Zhong Y, Paris C, Nepal S (2016) Incorporating tweet relationships into topic derivation. In: Hasida K, Purwarianti A (eds) Proceedings of the computational linguistics: 14th international conference of the pacific association for computational linguistics. PACLING 2015, Bali, Indonesia, May 19-21, 2015, Revised Selected Papers, Springer Singapore, Singapore, pp 177–190
87. Nugroho R, Zhao W, Yang J, Paris C, Nepal S (2016) Using time-sensitive interactions to improve topic derivation in Twitter. World Wide Web 20:1–27
88. Nurwidyantoro A, Winarko E (2013) Event detection in social media: a survey. In: Proceedings of the 2013 international conference on ICT for smart society (ICISS). IEEE, pp 1–5
89. Ostrow A (2009) Japan earthquake shakes Twitter users... and beyonce. http://mashable.com/2009/08/12/japan-earthquake/#4IvI9oMp8kqd, [Online, Accessed 6 October 2016]
90. Ozdikis O, Senkul P, Oguztuzun H (2012) Semantic expansion of tweet contents for enhanced event detection in Twitter. In: Proceedings of the 2012 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). pp 20–24. https://doi.org/10.1109/ASONAM.2012.14
91. Phan XH, Nguyen LM, Horiguchi S (2008) Learning to classify short and sparse text and web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on World Wide Web. ACM, New York, NY, USA, WWW '08, pp 91–100. https://doi.org/10.1145/1367497.1367510
92. Phan XH, Nguyen CT, Le DT, Nguyen LM, Horiguchi S, Ha QT (2011) A hidden topic-based framework toward building applications with short web documents. IEEE Trans Knowl Data Eng 23(7):961–976. https://doi.org/10.1109/TKDE.2010.27
93. Prateek M, Vasudeva V (2016) Improved topic models for social media via community detection using user interaction and content similarity. In: Proceedings of the 2016 international fruct conference on intelligence, social media and web (ISMW FRUCT). pp 1–7. https://doi.org/10.1109/FRUCT.2016.7584770
94. Prier KW, Smith MS, Giraud-Carrier C, Hanson CL (2011) Identifying health-related topics on Twitter. In: Salerno J, Yang SJ, Nau D, Chai SK (eds) Proceedings of the social computing, behavioral-cultural modeling and prediction: 4th international conference, SBP 2011. College Park, MD, USA, March 29–31, 2011., Springer, Berlin, Heidelberg, pp 18–25
95. Qiu M, Zhu F, Jiang J (2013) It is not just what we say, but how we say them: LDA-based behavior-topic model. In: Proceedings of the 2013 SIAM international conference on data mining. pp 794–802. https://doi.org/10.1137/1.9781611972832.88
96. Rafea A, Mostafa NA (2013) Topic extraction in social media. In: 2013 International conference on collaboration technologies and systems (CTS). IEEE, pp 94–98
97. Rafeeque PC, Sendhilkumar S (2011) A survey on short text analysis in web. In: Proceedings of the 2011 third international conference on advanced computing. pp 365–371. https://doi.org/10.1109/ICoAC.2011.6165203
98. Rajani NFN, McArdle K, Baldridge J (2014) Extracting topics based on authors, recipients and content in microblogs. In: Proceedings of the 37th international acm sigir conference on research and development

in information retrieval. ACM, New York, NY, USA, SIGIR '14, pp 1171–1174. https://doi.org/10.1145/2600428.2609537
99. Ramage D, Hall D, Nallapati R, Manning CD (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 conference on empirical methods in natural language processing: volume 1–volume 1, association for computational linguistics. Stroudsburg, PA, USA, EMNLP '09, pp 248–256. http://dl.acm.org/citation.cfm?id=1699510.1699543
100. Ramage D, Dumais ST, Liebling DJ (2010) Characterizing microblogs with topic models, vol 10. AAAI, Washington, pp 130–137
101. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: Proceedings of the 20th conference on uncertainty in artificial intelligence. AUAI Press, Arlington, Virginia, United States, UAI '04, pp 487–494. http://dl.acm.org/citation.cfm?id=1036843.1036902
102. Rosen-Zvi M, Chemudugunta C, Griffiths T, Smyth P, Steyvers M (2010) Learning author–topic models from text corpora. ACM Trans Inf Syst 28(1):4:1–4:38. https://doi.org/10.1145/1658377.1658381
103. Saha A, Sindhwani V (2012) Learning evolving and emerging topics in social media: a dynamic NMF approach with temporal regularization. In: Proceedings of the fifth ACM international conference on web search and data mining. ACM, New York, NY, USA, WSDM '12, pp 693–702. https://doi.org/10.1145/2124295.2124376
104. Sánchez DL, Revuelta J, De la Prieta F, Gil-González AB, Dang C (2016) Twitter user clustering based on their preferences and the Louvain algorithm. In: International conference on practical applications of agents and multi-agent systems. Springer, pp 349–356
105. Shahnaz F, Berry MW, Pauca VP, Plemmons RJ (2006) Document clustering using nonnegative matrix factorization. Inf Process Manag 42(2):373–386
106. Silva JA, Faria ER, Barros RC, Hruschka ER, Carvalho ACPLF, Gama JA (2013) Data stream clustering: a survey. ACM Comput Surv 46(1):13:1–13:31. https://doi.org/10.1145/2522968.2522981
107. Song G, Ye Y, Du X, Huang X, Bie S (2014) Short text classification: a survey. J Multimedia 9(5):635–643
108. Sparks RS (2018) Sentiment monitoring of social media from Oceania. Glob J Med Res 18(5-K). Retrieved from https://medicalresearchjournal.org/index.php/GJMR/article/view/1568
109. Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T (2004) Probabilistic author-topic models for information discovery. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, KDD '04, pp 306–315. https://doi.org/10.1145/1014052.1014087
110. Stilo G, Velardi P (2014) Time makes sense: event discovery in Twitter using temporal similarity. In: Proceedings of the 2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT)—Volume 02. IEEE Computer Society, Washington, DC, USA, WI-IAT '14, pp 186–193. https://doi.org/10.1109/WI-IAT.2014.97
111. Strehl A, Ghosh J (2003) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
112. Taslaman L, Nilsson B (2012) A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. PLoS ONE 7(11):e46331
113. Teh YW, Jordan MI, Beal MJ, Blei DM (2004) Sharing clusters among related groups: hierarchical Dirichlet processes. In: Proceedings of the 17th international conference on neural information processing systems (NIPS'04). MIT Press, Cambridge, MA, USA, pp 1385–1392
114. Tsur O, Littman A, Rappoport A (2013) Efficient clustering of short messages into general domains. In: Proceedings of the international AAAI conference on web and social media. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6103
115. Twitter (2011) #numbers. https://blog.twitter.com/2011/numbers. Online, Accessed 6 Oct 2016
116. Twitter (2011) Twitter milestones. https://about.twitter.com/company/press/milestones. Online, Accessed 6 Oct 2016
117. Vosecky J, Jiang D, Leung KWT, Xing K, Ng W (2014) Integrating social and auxiliary semantics for multifaceted topic modeling in Twitter. ACM Trans Internet Technol (TOIT) 14(4):27
118. Wallach HM, Murray I, Salakhutdinov R, Mimno D (2009) Evaluation methods for topic models. In: Proceedings of the 26th annual international conference on machine learning. ACM, New York, NY, USA, ICML '09, pp 1105–1112. https://doi.org/10.1145/1553374.1553515
119. Wan S, Paris C (2014) Improving government services with social media feedback. In: Proceedings of the 19th international conference on intelligent user interfaces. ACM, New York, NY, USA, IUI '14, pp 27–36. https://doi.org/10.1145/2557500.2557513
120. Wang X, McCallum A (2006) Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, KDD '06, pp 424–433. https://doi.org/10.1145/1150402.1150450

121. Wang X, Gerber MS, Brown DE (2012) Automatic crime prediction using events extracted from Twitter posts. In: Yang SJ, Greenberg AM, Endsley M (eds) Proceedings of the social computing, behavioral—cultural modeling and prediction: 5th international conference, SBP 2012. College Park, MD, USA, April 3–5, 2012. Proceedings, Springer, Berlin, Heidelberg, pp 231–238
122. Wang Y, Agichtein E, Benzi M (2012) TM-LDA: Efficient online modeling of latent topic transitions in social media. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, KDD '12, pp 123–131. https://doi.org/10.1145/2339530.2339552
123. Wang Y, Liu J, Qu J, Huang Y, Chen J, Feng X (2014) Hashtag graph based topic model for tweet mining. In: Proceedings of the 2014 IEEE international conference on data mining. pp 1025–1030. https://doi.org/10.1109/ICDM.2014.60
124. Weng J, Lim EP, Jiang J, He Q (2010) Twitterrank: finding topic-sensitive influential Twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. ACM, pp 261–270
125. Xia H, Li J, Tang J, Moens MF (2012) Plink-lda: Using link as prior information in topic modeling. In: Lee Sg, Peng Z, Zhou X, Moon YS, Unland R, Yoo J (eds) Proceedings of the database systems for advanced applications: 17th international conference, DASFAA 2012. Busan, South Korea, April 15–19, 2012, Part I, Springer, Berlin, Heidelberg, pp 213–227
126. Xie W, Zhu F, Jiang J, Lim EP, Wang K (2016) Topicsketch: real-time bursty topic detection from twitter. IEEE Trans Knowl Data Eng 28(8):2216–2229
127. Xu J, Liu P, Wu G, Sun Z, Xu B, Hao H (2013) A fast matching method based on semantic similarity for short texts. In: Zhou G, Li J, Zhao D, Feng Y (eds) Proceedings of the natural language processing and chinese computing: second CCF conference, NLPCC 2013. Chongqing, China, November 15–19, 2013,, Springer, Berlin, Heidelberg, pp 299–309
128. Yan X, Guo J, Lan Y, Cheng X (2013) A biterm topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web. ACM, New York, NY, USA, WWW '13, pp 1445–1456. https://doi.org/10.1145/2488388.2488514
129. Yan X, Guo J, Liu S, Cheng X, Wang Y (2013) Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In: Proceedings of the SIAM international conference on data mining (SIAM 2013). SDM, San Diego, California, USA
130. Yang Z, Yuan Z, Laaksonen J (2007) Projective non-negative matrix factorization with applications to facial image processing. Int J Pattern Recognit Artif Intell 21(08):1353–1362
131. Yıldırım A, Üsküdarlı S, Özgür A (2016) Identifying topics in microblogs using wikipedia. PLoS ONE 11(3):e0151885
132. Zhang C, Lu S, Zhang C, Xiao X, Wang Q, Chen G (2019) A novel hot topic detection framework with integration of image and short text information from twitter. IEEE Access 7:9225–9231
133. Zhao H, Du L, Buntine W, Zhou M (2018) Dirichlet belief networks for topic structure learning. In: Advances in neural information processing systems. pp 7955–7966
134. Zhao Y, Karypis G (2001) Criterion functions for document clustering: experiments and analysis. Tech. rep, Citeseer
135. Zhong Y, Yang J, Nugroho R (2015) Incorporating tie strength in robust social recommendation. In: Proceedings of the 4th IEEE international congress on big data. IEEE Services Computing Community, New York, USA, pp 63–70
136. Zhou T, Shan H, Banerjee A, Sapiro G (2012) Kernelized probabilistic matrix factorization: exploiting graphs and side information. In: Proceedings of the 2012 SIAM international conference on data mining. SIAM, California, USA, vol 12, pp 403–414
137. Zuo Y, Zhao J, Xu K (2016) Word network topic model: a simple but general solution for short and imbalanced texts. Knowl Inf Syst 48(2):379–398

Robertus Nugroho is a lecturer at Soegijapranata Catholic University in Semarang Indonesia and Visiting Scientist at CSIRO Data61, Australia. He received his PhD from the Department of Computing, Macquarie University, Australia, in 2018 and masters degree in computing and information technology from the University of New South Wales, Australia, in 2009. Before returning to Soegijapranata Catholic University, Robertus worked as a research fellow at Macquarie University and CSIRO Data61. He was awarded a postgraduate studentship position at CSIRO Australia (2014–2017). In 2015, he received the best student paper award at IEEE BigData Congress 2015 and the best paper award at Web Information System Engineering (WISE) 2015 Conference. His current research interests include big data, social network analysis, and machine learning.

Cecile Paris is the Science Leader for the Knowledge Discover and Management Research Group at Data61, CSIRO. Her expertise is in natural language processing, user modeling, social media analytics, social computing and, more generally, in artificial intelligence and communication. She is interested in understanding how people communicate, in facilitating communication with information and information environments and in making sense of big data. With a bachelors degree from the University of Berkeley (California) and a PhD from Columbia University (New York), she has over 25 years of experience in research and research management, in CSIRO and other research laboratories overseas. Her group develops systems that are being used in government and in industry, in a wide variety of domains, including service delivery, digital libraries, e-research, mental health, business intelligence, media monitoring, customer relationship, and service delivery. Dr Paris is a Fellow of the Australian Academy of Technology and Engineering (ATSE). She has authored over 280 refereed technical papers. Dr Paris is very active in the research community, in Australia and internationally, serving on numerous conference and workshop committees, on review boards of grant-giving bodies and journals.

Surya Nepal received BE degree from the National Institute of Technology, Surat, India, ME degree from the Asian Institute of Technology, Bangkok, Thailand, and PhD degree from RMIT University, Australia. He is a principal research scientist at CSIRO Data61. His main research interest includes the development and implementation of technologies in the area of distributed systems and social networks, with a specific focus on security, privacy, and trust. At CSIRO, he undertook research in the area of multimedia databases, web services and service-oriented architectures, social networks, security, privacy and trust in collaborative environment, and cloud systems and big data. He has more than 150 publications to his credit. Many of his works are published in top international journals and conferences such as VLDB, ICDE, ICWS, SCC, CoopIS, ICSOC, International Journals of Web Services Research, IEEE Transactions on Service Computing, ACM Computing Survey, and ACM Transaction on Internet Technology.

Jian Yang is a professor at Department of Computing, Macquarie University. She received her PhD in Multidatabase Systems area from The Australian National University in 1995. Prior to joining Macquarie University, she was an associate professor at Tilburg University, Netherlands (2000–2003), a senior research scientist at the Division of Mathematical and Information Science, CSIRO, Australia (1998–2000), and a lecturer (assistant professor) at Department of Computer Science, The Australian Defence Force Academy, University of New South Wales (1993–1998). Her main research interests are web service technology; business process management; interoperability, trust and security issues in digital libraries and e-commerce; and social networks.

Weiliang Zhao is working at the College and Computer Science and Technology at Donghua University, China. He received his PhD at the School of Mathematics and Computing, University of Western Sydney, in 2009. Before joining Donghua University, he worked as a research fellow at Macquarie University, research fellow at University of Wollongong, data analyst at the Copyright Agency, software developer at ROAMZ, programmer at ANZ bank, and researcher at Chinese Academy of Science. His main research interests are social networks, service computing, trust management in distributed systems, and security in electronic commerce applications.