An Overview of Topic Discovery in Twitter Communication through Social Media Analytics

Full Paper

Andrey Chinnov
University of Münster, Germany
Information Systems
a_chin01@uni-muenster.de

Pascal Kerschke
University of Münster
Information Systems
kerschke@uni-muenster.de

Christian Meske
University of Duisburg-Essen
Professional Communication in Electronic Media
christian.meske@uni-due.de

Stefan Stieglitz
University of Duisburg-Essen
Professional Communication in Electronic Media
stefan.stieglitz@uni-due.de

Heike Trautmann
University of Münster
Information Systems
trautmann@uni-muenster.de

Abstract

The need for automatic methods of topic discovery in the Internet grows exponentially with the amount of available textual information. Nowadays it becomes impossible to manually read even a small part of the information in order to reveal the underlying topics. Social media provide us with a great pool of user generated content, where topic discovery may be extremely useful for businesses, politicians, researchers, and other stakeholders. However, conventional topic discovery methods, which are widely used in large text corpora, face several challenges when they are applied in social media and particularly in Twitter – the most popular microblogging platform. To the best of our knowledge no comprehensive overview of these challenges and of the methods dedicated to address these challenges does exist in IS literature until now. Therefore, this paper provides an overview of these challenges, matching methods and their expected usefulness for social media analytics.

Introduction

The Internet and mobile technologies, which have been continuously developed over the last decades, led to the rise of social media (Zeng et al., 2010), which can be defined as a "group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the

creation and exchange of User Generated Content" (Kaplan and Haenlein, 2010, p. 61).Besides private individuals, organizations can also make use of social media for different purposes (see e.g. Meske and Stieglitz, 2013; Stieglitz et al., 2013). The world-wide adoption of social media such as social networks increased tremendously (Larosiliere et al., 2015). For instance, according to the official statistics web pages, an average of 1.28 billion people use Facebook at least once a month (Facebook, 2014). For Twitter there are 255 million monthly users (Twitter, 2014b). In Twitter, currently the most popular microblogging platform, messages (tweets) are limited to 140 characters. Twitter users can easily create tweets and spread information by retweeting the initial tweet (Zhao et al., 2011). Zu et al., 2014 (p. 199) state, that Twitter "has excellent features for knowledge exploration: open to the public, abbreviated, easy to analyze, and constantly generated". The usage of twitter leads to (big) data that provides rich and unfiltered insights into primary communication data. Twitter provides an application programming interfaces (API) (Twitter, 2014a) that allows for automatical collection of such data.

Social Media Analytics (SMA) can hence bring benefit to many different contexts, e.g. economics, politics, health, science, and others (Zeng et al., 2010). For example, businesses may gain benefit by deriving innovative marketing ideas through the collection of customer opinions and sentiments regarding different products. Furthermore, social media monitoring can be performed for exploration purposes: a company might be interested in knowing about what kind of topics related to the company's brand are discussed and how such discussions take place in social media (Stieglitz and Dang-Xuan, 2013). Similarly, politicians become able to analyze opinions and ideas about certain candidates or their campaigns, offering additional information to adjust e.g. political communication strategies. In addition, social media can be crucial in crisis situation, providing authorities with actual information for instance regarding the current progress of a crisis, posted by e.g. twitter users in the respective crisis area. Therefore, in recent years social media became an object of scientific research. Such research is multidisciplinary in nature, including e.g. information systems, communication studies, sociology and others. It also involves different methods like statistics, graph theory, social network analysis, text mining methods, and others.,

Different approaches have been proposed and applied for SMA purposes such as topic or issue discovery, sentiment analysis or structural analysis. Topic discovery can be considered a discipline in academia that deals with the automatic uncovering of latent structures and patterns of text documents, mostly applying text mining and data mining techniques. We will show that social media, and Twitter in particular, dispose of several challenges to conventional topic discovery methods, which make it difficult to discover meaningful and coherent topics. These main challenges are dynamism and continuity, network structure, noisiness, metadata enrichment and text shortness. In this paper, these challenges are described in detail, each followed by a discussion of existing SMA-methods to meet these challenges. To the best of our knowledge, until now in SMA literature no comprehensive overview of these challenges and matching methods exists.

Topic Discovery in Social Media – Challenges and Approaches

Topic discovery has its roots in closely related research areas such as automatic document classification (ADC) and information retrieval (IR). The main purpose of ADC is to classify the documents of a corpus, given a number of "headings" (classes), where a heading is simply a collection of words. The main purpose of IR is to find all the relevant documents from the corpus given a heading (query).

The modern topic discovery techniques can be separated in two main categories: distance-based document clustering and probabilistic approaches. A *distance-based document clustering process* consists of the following four steps: definition of a document representation as well as a similarity or a distance measure, followed by the selection and application of a clustering algorithm. Different representations, measures and algorithms subsequently form different approaches (see e.g. Dhillon and Modha, 2001; Kim et al., 2012; Larsen and Aone, 1999; McCallum et al., 2000; Popescul et al., 2000; Seo and Sycara, 2004). As a result, similar documents are clustered into one group and the corresponding topic is assigned to all the documents within the group. In contrast to document clustering, in the context of probabilistic approaches each document is considered a mixture of topics, rather than a sequence of words drawn from the same topic. It is assumed that documents in a corpus are generated by generative probabilistic processes (GPP). In such a process, documents and words (perceived as observed random variables), as well as topics (perceived as latent random variables) are generated according to the

predefined distributions, presumably depending on a number of parameters. The distribution parameters and the number of topics are the parameters of the model. The topic discovery task in this case is equivalent to the maximum likelihood parameter estimation. The *Latent Dirichlet Allocation (LDA)* approach (Blei et al., 2003), which is the simplest probabilistic topic model, can be considered one of the most fundamental works in the topic discovery research area. Its main advantage is that it can be embedded in more complicated models that capture and reflect richer assumptions about the analyzed data. Moreover, the distribution families in LDA can be easily changed, allowing us to apply the model to many kinds of data. Therefore, LDA has been adapted and extended in many ways. Each of its extensions relaxes some of the LDA assumptions. For example, the *correlated topic model (CTM)* (Blei and Lafferty, 2006a), the *author-topic (AT)* model (Steyvers et al., 2004) and the *hierarchical Dirichlet language model (HDLM)* (MacKay and Peto, 1995) are such LDA extensions.

Social media and Twitter in particular dispose of several challenges for the topic discovery task. For example, Kireyev et al. (2009) emphasize the informality of the language, the lack of proper written grammar and the shortness of messages. The standard topic discovery methods, which require large document collections, may not work properly in such a domain. Also, social media inherits the challenges of social networks and text data streams since the communication usually has a network structure (for example, the network of followers in Twitter) or stream characteristics (messages are distributed over time via timestamps). Moreover, communication in social media is usually appended with a lot of meta-data that should also be used for improving the quality of the standard topic discovery techniques. In this regard, especially the following challenges are subject to research:

- Dynamism and Continuity
- Network Structure
- Noisiness
- Metadata Enrichment
- Text Shortness

The following subsections describe each challenge in details and present possible solutions that are discussed in related literature.

Dynamism and Continuity

The dynamism property reflects the fact that communication in social media is distributed over time, whereas the continuity property reflects the infinity of the message flow. In contrast to large size document collections, communication in social media has no determined ending. Therefore, the standard static offline methods for topic discovery may be inappropriate. Topic discovery in text data streams (TDS) can be considered a part of the more general study temporal text mining (TTM) that "...is concerned with discovering temporal patterns in the text information collected over time" (Mei and Zhai, 2005, p. 198). The documents in a collection are tagged with timestamps forming the text data stream. In such a stream, besides the ordinary topic modeling task, the following two questions appear: (1) How do topics evolve over time and (2) what are the emerging topics?

Conventional topic models (LDA, CTM, and others) assume that documents are exchangeable within a corpus (the order does not matter) and, therefore, such methods were inappropriate for data streams, e.g. scholarly journals, news articles or emails. The articles about topic discovery published half a century ago substantially differ from the articles about this topic published today. One of the problems related to continuity is that "...the whole text data cannot be fit into memory at once and multiple scans of the data kept in secondary storage are not possible due to real-time response requirements (...)" (Liu et al., 2008, p. 113). Addressing these problems, Mei and Zhai (2005) divided the sequentially organized collection of documents by the time slice (e.g. year) and modelled topics within a separate time partition using the simple mixture models. The authors leveraged the KL divergence to discover coherent topics over time. Using this metric, the evolution graph of a topic is constructed, what can support to reveal the topic's development over time as well as its influence on other topics. Blei and Lafferty (2006b) showed how LDA can be extended for topic modeling in text data streams by proposing the dynamic topic model (DTM).

The authors used the similar idea of time-based document partitioning and showed that DTM can provide more accurate predictive models, than LDA does (Blei and Laerty, 2006b).

Event detection, as a sub-field of topic discovery in streams, also attracts high academic interest (Allan et al., 1998a,b; Luo et al., 2007; Mehmood et al., 2013; Weng and Lee, 2011, and others). An overview can be found in (Nurwidyantoro and Winarko, 2013). Yao et al. (2009) emphasized the computational complexity of existing topic discovery approaches in the large-scale and real-time context. Addressing this problem, the authors proposed the more flexible topic modeling framework, SparseLDA, which can be up to 20 times faster than the traditional LDA approach, while using substantially less memory. The idea behind this method is to pre-compute some of the calculations in the inference procedure improving the overall performance (see Yao et al., 2009 for more details). Wang et al. (2012) announced that their novel method, temporal-LDA (TM-LDA) works better in the microblogging domain than previous state-of-the-art methods. The idea behind this method is the modeling of the topic transitions in temporarily-sequenced documents (streams). The authors noticed that users tend to post messages about different topics instead of simply developing the previous topic. It means that, in order to improve models regarding the dynamic semantics of microblog post streams, the changing pattern among topics should be captured. For example, the topic "Food" usually succeeds the topic "Drink". The understanding of such topic transitions may help: (1) to dynamically predict future trends, (2) to get a more in-depth view of temporal relationships among social behaviour, and (3) to detect unusual events when the topics fail to follow expected transitions (Wang et al., 2012).

In continuous flows, it is inefficient to retain all the communication messages. Therefore, the existing approaches manage such flows by retaining some globally applicable statistics such as topic-word counters. The document-level information is usually lost since it is not clear how to retain such information in a scalable and meaningful way (Slutsky et al., 2014). Addressing this limitation, (Slutsky et al., 2014) introduced the hash-based stream LDA (HS-LDA), which is a "...framework that makes it possible to retain the knowledge of historical stream messages in a scalable way and uses this knowledge to improve the quality of topic discovery in social streams" (Slutsky et al., 2014, pp. 151-152). The main idea of HS-LDA was borrowed from particle physics where nearly massless uncharged particles called neutrinos exist and which can only be detected through their interactions with other matters. According to HS-LDA, the corresponding generative process, in addition to words, also emits "pseudo-neutrinos", certain auxiliary objects associated with the words that are not directly observable in the corpus. It is assumed that these pseudo-neutrinos belong to the fixed set of possible types. If this set is restricted to a single type, then HS-LDA is equivalent to LDA. Therefore, HS-LDA is a generalization of LDA. The authors presented the inference algorithm appending the standard Gibbs sampling with the locality sensitive hashing (LSH) technique, which allows to detect pseudo-neutrinos in the documents. It was shown that HS-LDA significantly outperforms LDA in terms of perplexity (Slutsky et al., 2014).

Network Structure

As long as social media inherits the network structure of the underlying communication, the full spectrum of the social network analysis (SNA) techniques could be applied to gain additional information that can be worthful in different contexts. The main focus of SNA lies on the relationships among social entities (users), as well as the patterns and implications of these relationships (Wasserman, 1994). The first works in SNA treated social networks as graphs and studied basically their topological properties like connectivity, degree distribution, correlation and centrality (see e.g. Albert and Barabási, 2002; Newman, 2003; Wasserman, 1994). It was shown that social networks are not the ordinary graphs and have some particular properties. For example, the small world property (Milgram, 1967) or "six degrees of separation" means that each two nodes in the graph are in general connected with a path of length six. Therefore, different mathematical social network models have been proposed.

One of the most famous social network models is the random graph proposed by Erdos and Rényi (1959). It was proven that such a model satisfies the small world property (Albert and Barabási, 2002). However, topological network analysis is not enough to discover all the roles and patterns in a social network (McCallum et al., 2005). The content of communications and the information flow between the social network nodes subsume a rich source of additional information about the communication. Therefore, at the beginning of the 21st century, with the growth of available text corpora of social network communications - for example, the Enron email dataset (Cohen, 2009) -- the content and text mining techniques became one of the main SNA tools. The Enron email dataset became also an object of study in the topic discovery research area. McCallum et al. (2005) argued that at that point of time none of the

existing topic models were appropriate for SNA, in which the goal is to capture the directed interactions and relationships between people. Indeed, the discovered topics from the set of documents having the same authors and recipient should be highly correlated. To capture this issue, McCallum et al. (2005) proposed the author-recipient-topic (ART) model, which extends the author-topic (AT) model, appending an additional observed variable of the document's recipient. The authors showed that the predictions of the user roles, made by the ART model, are better than the ones made by the AT model. Subsequently, Wang et al. (2005) extended the ART model, which does not explicitly capture the groups formed by entities in the communication network. Therefore, the group-topic (GT) model was presented capturing different relationships between people and other entities. In this way, the modeled topics became more group dependent, better corresponding to the reality.

Mei et al. (2008) formalized topic discovery in social networks by introducing the notion of topic modeling with network structure (TMN) where a collection of documents (e.g. messages or articles) is associated with a network (e.g. sender-recipient network or the network of co-authority). The three major TMN tasks formulated by the authors are: (1) topic extraction, (2) topic map extraction, and (3) topical community discovery. Roughly speaking, the first task is equivalent to the topic discovery task in a collection of documents where the number of topics is predefined. The topic map extraction (2) shows how the topic is distributed in the network. The third task helps to identify topical communities in the network, e.g. groups of entities associated with similar topics. It is assumed that the nodes of the network associated with similar topics should be tightly coupled within the networks, whereas the node associated with different topics should be loosely coupled. Applying this assumption, the authors proposed a framework to model topics with a network structure, by regularizing the underlying statistical topic model with the special regularizer on the network, which makes the weights of topics for the connected nodes similar. Subsequently, the word distributions for a connected pair of nodes become similar. Although the framework was initially applied to probabilistic Latent Semantic Indexing (pLSI), it is an approach, which generally can be useful for regularization of any other statistical topic model.

Noisiness

The noisiness challenge reflects the fact that communication in social media is usually informal, spam affected and saturated with a high amount of misspelling and grammar errors. Therefore, the traditional methods based on the bag-of-words model have limitations. Sriram et al. (2010) proposed one of the possible solutions: information filtering. A set of domain-specific features are extracted from the text and the author's profile in order to classify all messages into fixed predefined sets of generic classes such as news, events, opinions and others. Spam detection, a method dedicated to detect spam-related messages and users within communications, is another possible solution of this problem, which is also actively discussed in academia (McCord and Chuah, 2011; Miller et al., 2014; Wang, 2010). Most of the approaches are trying to discover some content-based and user-based features that distinguish spammers from non-spammers (legitimate users). The LDA model was also extended in a way that overcomes the noisiness problem. The Topic-coreterms Latent Dirichlet Allocation (TC-LDA) proposed by Ge et al. (2013) increases the layer of background model to reduce the noise in microblogging data. In contrast to the traditional layers “document-topic-word“, the modification “document-topic-coreterms-word“ is applied. The authors demostrated the out-performance of TC-LDA over LDA in terms of perplexity.

Metadata Enrichment

One of the social media characteristics nowadays is the high amount of meta-information that is simultaneously transmitted together with the main information. Hyperlinks, tags, geo-location and different labels are all examples of metadata that could be used in order to improve the performance of traditional topic discovery methods. Most of the standard methods are unsupervised trying to find hidden structures and patterns in unlabeled data. However, a lot of data are paired with the response variables (labels). For example, user reviews are paired with the number of stars, web pages are paired with the number of “likes“, and so on. This information should be also used in predictive models. Consider the prediction of a movie rating based on the words in the user reviews. Appropriate predictive topic models need to differentiate words like “good”, “bad”, and “average” regardless the genre. Yet, the general unsupervised topic modeling techniques relate topics to genres as soon as there is such a dominant structure in the corpus (Mcauliffe and Blei, 2008). Addressing the labeled data, the supervised LDA (sLDA) model (Mcauliffe and Blei, 2008) was proposed. In this model the LDA module is extended by a

response variable, which is associated with each document (e.g. the number of stars given to a movie). The documents and the responses are jointly modeled in order to find the latent topics that predict the response variables the best (Mcaulffe and Blei, 2008). Relational topic models (RTM) use sLDA assumptions with pair-wise responses to model networks of documents. The model helps "... to summarize a network of documents, predict links between them, and predict words within them" (Chang and Blei, 2009, p. 81). With the increase of mobile technologies an increasing amount of messages in social media are extended by geo-location information. This offers an opportunity to monitor how information is created and shared across different regions as well as how users and the users' terminology differ across the regions. In the context of topic discovery, this may impose some challenges due to the diversity of language variations.

A number of studies contribute to this problem (see e.g. Hong et al., 2012; Hu and Ester, 2013; Zhang et al., 2013). In further studies, authors try to make use of hashtags widely used in social media (Feng and Wang, 2014; Ma et al., 2013; Mehmood et al., 2013). Hashtags are usually user driven and serve as significant metadata to categorize messages, to easily code and spread ideas or trends. However, sometimes it is difficult to interpret hashtags and to discover their relationships due to their free-form nature (Ma et al., 2013). For example, different users may use different hashtags to annotate the same event(s). Ma et al. (2013) proposed tag-latent Dirichlet allocation (TLDA), a topic modeling approach dedicated to bridge topics and hashtags. In this method, the hashtags are integrated into the LDA model by means of the additional observed variables. As a result, the TLDA model learns the latent topics for each hashtag represented in the common topic space. This offers an opportunity to measure similarities between the hashtags. The experiments showed that TLDA outperforms the AT model in terms of perplexity. Feng and Wang (2014) tried to standardize the process of hashtag creation by proposing the statistical model for personalized hashtag recommendation, which "... suggests both content-relevant and user-relevant hashtags when users are composing tweets" (Feng and Wang, 2014, p. 866).

Text Shortness

In the context of short messages, insignificant words and stopwords are a substantial part and significantly influence their representation in the discovered topics (Kireyev et al., 2009). Moreover, the word co-occurrences, which play the main role in the traditional clustering and probabilistic approaches, are weakly pronounced in such short messages, what results in the poor accuracy of the discovered topics (Ni et al., 2011). Due to the lack of specific approaches for topic discovery in short texts, some researchers applied conventional methods (or slight modifications) for the short text analysis (Ramage et al., 2010; Wang et al., 2012). Others tried to implement a more intuitive solution to this problem, combining short texts into the larger "macro-documents", where the conventional methods may properly work (Aiello et al., 2013; Mehrotra et al., 2013; Weng et al., 2010; Zhao et al., 2011).

For example, the recent work of Mehrotra et al. (2013) focuses on obtaining better LDA topics from Twitter content without modifying basic machinery of the standard LDA model. Different pooling (aggregating) schemes are proposed for this purpose. The authors showed that the aggregation of tweets, which are similar in some sense -- semantically, temporally, same author or similar hashtags, etc -- enriches the content presented in the individual tweets so that LDA can learn topics easier (Mehrotra et al., 2013). Hong and Davison (2010) conducted an empirical study of topic modeling in Twitter applying different aggregation strategies and illustrated the need for new methods of topic discovery applicable in the context of short messages.

There have been several attempts to cluster short texts (Banerjee et al., 2007; Kim et al., 2012; Ni et al., 2011; Petkos et al., 2014, and others). Banerjee et al. (2007) proposed a method for improving the accuracy of clustering short texts using Wikipedia as an additional data source. In this approach, the titles of Wikipedia articles matching the tweets serve as additional features, which can be used in the clustering process. However, Twitter evolves much faster than Wikipedia, resulting in a significant influence wrt. the accuracy of such an approach (Kim et al., 2012). Ni et al. (2011) proposed the graph-based strategy TermCut for clustering short texts, which is different from the traditional distance-based clustering approaches. Consider the following three tweets:

T1 : "Topic discovery in Twitter"
T2 : "Topic discovery in Facebook"
T3 : "Clustering Twitter messages."

The traditional approaches would assign T1 and T2 to the same cluster since they have more words in common. However, they do not share the core terms ("Twitter" and "Facebook" in this case) and should be considered dissimilar by another strategy. In contrast to this, Ni et al. (2011) consider a collection of short text snippets as a graph where each vertex represents some piece of a short text snippet and each weighted edge between two vertices measures the relationship between the two vertices. Based on that, the algorithm TermCut recursively selects the core term and bisects the graph so that the short text snippets in the one part of the graph contain the term, whereas the snippets in the other part do not. In order to find the core terms, the authors introduce the criterion function RMcut. A term is a core term if the value of RMcut is minimal, after the cluster bisection associated to this term (see Ni et al., 2011, for more details). As a result, the tweets T1 and T3 might be considered similar by this approach since they share the core term "Twitter". The efficiency of the previous algorithm was improved by Kim et al. (2012), who proposed a similar method, the core-topic-based clustering (CTC). The authors show that the clustering performance is significantly better than the performance of the k-means algorithm.

4 Conclusion and Outlook

The tremendous growth of user-generated content in social media during the last decades elicitated the appearance of social media analytics as a self-contained research area. The generated information contains people's opinions or sentiments about certain entities and subjects (like companies, brands, politics, …) as well as events or other important discussion topics. The main intention of social media analytics is to utilize this generated information for beneficial purposes. Twitter is the most popular microblogging platform, allowing to easily create and spread information over the entire network. As long as Twitter provides convenient APIs, it remains possible to track Twitter and hence primary (big) communication data. By doing so, companies, for example, could monitor and react to the discussed topics related to the companies or the companies' brands. Therefore, topic discovery methods became an important part of social media analytics in the commercial context. Topic discovery also can be considered a self-contained research area, being the methodological basis of many interdisciplinary research areas such as information systems, communication studies and others. The main advantage of the probabilistic approaches is that the words of a document can be drawn from different topics (mixture of topics). This fact provides much more flexibility. The systematic literature review revealed that social media, and particularly Twitter, impose several challenges on the topic discovery task. Twitter communication is a continuous flow of time specific documents (tweets) consisting of a network structure (e.g. retweeting network). Moreover, tweets are usually "noisy" and short text documents enriched with additional metadata. All that differ Twitter corpora from other corpora of text documents, where the conventional topic discovery methods have been successfully applied.

To meet these challenges, several approaches have been proposed. For example, to capture topic evolution over time one could apply DTM. SparseLDA and HS-LDA address the infinity of social media communication trying to overcome the memory and computational limitations. Communication network structure can be captured by the ART and GT approaches. The information filtering techniques together with TC-LDA help to reduce noise in text documents. In addition, different kinds of metadata in text documents are leveraged to improve quality of topic discovery in approaches like sLDA, RTM and TLDA. Finally, the text shortness problem could be addressed by some aggregating schemes or by some more sophisticated approaches like CTC.

Apparently, it is very difficult, and probably impossible, to address all the challenges at once. Each of the mentioned methods is concentrating on a particular problem, which it is dedicated to. Therefore, it is up to the practitioner to choose an appropriate method basing on the underlying circumstances. For instance, the shortness problem is much more relevant for communication in Twitter than in Facebook. On the other hand, the flexibility of probabilistic approaches is quite doubtable due to the shortness of tweets. The words of a short tweet are more likely to be drawn from the same topic. Therefore, the conventional clustering approaches usually work well for Twitter.

Future work should deal with a systematic empirical study comparing the discussed approaches in order to be able to give sound recommendations for their practical usage. In this context, appropriate evaluation measures should be proposed, which help to assess the quality of the resulting topics (topic collections). Furthermore, the influence and importance of data pre-processing needs to be investigated in detail.

REFERENCES

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., and Jaimes, A. 2013. "Sensing Trending Topics in Twitter," *Multimedia, IEEE Transactions on*, (15:6), pp. 1268-1282.

Albert, R., Barabási, A.-L. 2002. "Statistical mechanics of complex networks," *Reviews of modern physics*, (74:1), p. 47.

Allan, J., Carbonell, J., and Doddington, G. 1998a. "Topic detection and tracking pilot study final report," *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, pp. 194-218.

Allan, J., Papka, R., and Lavrenko, V. 1998b. "On-line New Event Detection and Tracking," In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98*, New York, NY, USA, pp. 37-45.

Banerjee, S., Ramanathan, K., and Gupta, A. 2007. "Clustering Short Texts Using Wikipedia," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval , SIGIR '07*, New York, NY, USA, pp. 787-788.

Blei, D. M., and Lafferty, J. D. 2006a. "Correlated Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113-120.

Blei, D. M., and Lafferty, J. D. 2006b. "Dynamic Topic Models," in *Proceedings of the 23rd international conference on Machine learning*, pp. 113-120.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3:4-5), pp. 993-1022.

Chang, J., Blei, D. M. 2009. "Relational Topic Models for Document Networks," In *International Conference on Artificial Intelligence and Statistics*, pp. 81-88.

Dhillon, I. S., and Modha, D. 2001. "Concept Decompositions for Large Sparse Text Data Using Clustering," *Machine Learning* (42:1-2), pp. 143-175.

Erdos, P. and Rényi, A. 1959. "On random graphs," *Publ. Math. Debrecen* (6), pp. 290-297.

Facebook 2014. Statistics. http://newsroom.fb.com/company-info/. Accessed 21-July-2014.

Feng, W., Wang, J. 2014. "We can learn your #hashtags: Connecting tweets to explicit topics," in *Proceedings of 2014 IEEE 30th International Conference on Data Engineering (ICDE)*, pp. 856-867.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K. 2012. "Discovering geographical topics in the twitter stream," in *Proceedings of the 21st international conference on World Wide Web - WWW '12*, p. 769.

Hong, L., Davison, B. D. 2010. "Empirical study of topic modeling in Twitter," in *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pp. 80-88.

Hu, B., Ester, M. 2013. "Spatio-Temporal Topic Modeling in Mobile Social Media for Location Recommendation," in *Proceedings of 2013 IEEE 13th International Conference on Data Mining (ICDM)*, pp. 1073-1078.

Kaplan, A. M., Haenlein, M. 2010. "Users of the World, Unite! The Challenges and Opportunities of Social Media," *Business Horizons*, (53:1), pp. 59-68.

Kim, S., Jeon, S., Kim, J., Park, Y.-H., and Yu, H. 2012. "Finding Core Topics: Topic Extraction with Clustering on Tweet," in *Proceedings of 2nd International Conference on Cloud and Green Computing*, pp. 777-782.

Kireyev, K., Palen, L., and Anderson, K. 2009. "Applications of topics models to analysis of disaster-related twitter data," In *Proceedings of NIPS Workshop on Applications for Topic Models: Text and Beyond* 1.

Larosiliere, G., Meske, C. and Carter, L. 2015. "Determinants of Social Network Adoption: A Country-Level Analysis," in *Proceedings of the 48th Hawaii International Conference on System Sciences (HICSS)*, pp. 3424-3433.

Larsen, B., Aone, C. 1999. "Fast and Effective Text Mining Using Linear-time Document Clustering," in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 16-22.

Liu, Y.-B., Cai, J.-R., Yin, J., and Fu, A.-C. 2008. "Clustering Text Data Streams," *Journal of Computer Science and Technology*, (23:1), pp.112-128.

Luo, G., Tang, C., and Yu, P. S. 2007. "Resource-adaptive Real-time New Event Detection," in *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, SIGMOD '07*, New York, NY, USA, pp. 497-508.

Ma, Z., Dou, W., Wang, X., and Akella, S. 2013. "Tag-Latent Dirichlet Allocation : Understanding Hashtags and Their Relationships," pp. 260-267.

MacKay, D. J. C., Peto, L. C. B. 1995. "A Hierarchical Dirichlet Language Model," *Natural Language Engineering* (1:3), pp. 289-308.

Mcauliffe, J., and Blei, D. 2008. "Supervised Topic Models," *Advances in Neural Information Processing (20)*, pp. 121-128.

McCallum, A., Nigam, K., and Ungar, L. H. 2000. "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching," in *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 169-178.

McCallum, A., Wang, X., and Corrada-Emmanuel, A. 2005. "Topic and role discovery in social networks," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 786-791.

McCord, M. and Chuah, M. 2011. "Spam Detection on Twitter Using Traditional Classiffers," in *Autonomic and Trusted Computing, volume 6906 of Lecture Notes in Computer Science*, J. Calero, L. Yang, F. Mármol, L. García Villalb, A. Li, and Y. Wang (eds.), Berlin: Springer-Verlag, Heidelberg, pp. 175-186.

Mei, Q., and Zhai, C. 2005. "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 198-207.

Mehmood, R., Maurer, H., and Afzal, M. T. 2013. "Knowledge discovery in hashtags#," in *Proceedings of 2013 IEEE 9th International Conference on Emerging Technologies (ICET)*, pp. 1-6.

Mehrotra, R., Sanner, S., Buntine, W., and 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, SIGIR '13*, New York, NY, USA, pp. 889-892.

Mei, Q., Cai, D., Zhang, D., and Zhai, C. 2008. "Topic Modeling with Network Regularization," in *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, New York, NY, USA, pp. 101-110.

Meske, C. and Stieglitz, S. 2013. "Adoption and Use of Social Media in Small and Medium-sized Enterprises," in *Proceedings of the 6th Practice-Driven Research on Enterprise Transformation (PRET)*, Lecture Notes in Business Information Processing (LNBIP), pp. 61-75.

Milgram, S. 1967. "The small world problem," *Psychology today*, (2:1), pp. 60-67.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., and Wang, A. H. 2014. "Twitter spammer detection using data stream clustering," Information Sciences, (260:0), pp. 64-73.

Newman, M. E. J. 2003. "The structure and function of complex networks," *SIAM review*, (45:2), pp. 167-256.

Ni, X., Quan, X., Lu, Z., Wenyin, L., and Hua, B. 2011. "Short text clustering by finding core terms," *Knowledge and Information Systems*, (27:3), pp. 345-365.

Nurwidyantoro, A., Winarko, E. 2013. "Event detection in social media: A survey," in *Proceedings of 2013 International Conference on ICT for Smart Society (ICISS),* pp. 1-5.

Petkos, G., Papadopoulos, S., and Kompatsiaris, Y. 2014. "Two-level Message Clustering for Topic Detection in Twitter," *SNOW-DC@ WWW*, pp. 49-56.

Popescul, A., Flake, G. W., Lawrence, S., Ungar, L.H., and Giles, C. L. 2000. "Clustering and Identifying Temporal Trends in Document Databases," in *Proceedings of IEEE on Advances in Digital Libraries*, pp. 173-182.

Ramage, D., Dumais, S. T., and Liebling, D. J. 2010. "Characterizing Microblogs with Topic Models," in *Proceedings of ICWSM*, 10:1.

Seo, Y.-W., Sycara, K. 2004. "Text Clustering for Topic Detection," *Tech. Rep.*, Robotics Institute, Carnegie Mellon University.

Shahnaz, F., Berry, M. W., Pauca, V., and Plemmons, R. J. 2006. "Document Clustering Using Nonnegative Matrix Factorization," *Information Processing and Management* (42:2), pp. 373-386.

Slutsky, A., Hu, X., and An, Y. 2014. "Hash-Based Stream LDA: Topic Modeling in Social Streams," in *Advances in Knowledge Discovery and Data Mining, volume 8443 of Lecture Notes in Computer Science*, V. Tseng, T. Ho, Z.-H. Zhou, A. Chen, and H.-Y. Kao (eds.), Springer International Publishing, pp. 151-162.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. 2010. "Short Text Classification in Twitter to Improve Information Filtering," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10*, New York, NY, USA, pp. 841-842.

Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T. 2004. "Probabilistic Author-Topic Models for Information Discovery," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 306-315.

Stieglitz, S., Dang-Xuan, L. 2013. "Social media and political communication: a social media analytics framework," *Social Network Analysis and Mining*, (3:4), pp. 1277-1291.

Stieglitz, S., Schallenmüller, S. and Meske, C. 2013. "Adoption of Social Media for Internal Usage in a Global Enterprise," in *Proceedings of the IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*, pp. 1483- 1488.

Twitter 2014a. "Twitter Developers," https://dev.twitter.com/. Accessed 21-July-2014.

Twitter 2014b. "Twitter Usage," https://about.twitter.com/company. Accessed 21-July-2014.

Wang, A. H. 2010. "Don't follow me: Spam detection in Twitter," in *Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT),* pp. 1-10.

Wang, X., Mohanty, N., and McCallum, A. 2005. "Group and Topic Discovery from Relations and Text*," in Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD '05*, New York, NY, USA, pp. 28-35.

Wang, Y., Agichtein, E., and 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, KDD '12*, New York, NY, USA. Pp. 123-131.

Wasserman, S. 1994. *Social network analysis: Methods and applications, volume 8*, Cambridge university press.

Weng, J. and Lee, B.-S. 2011. "Event Detection in Twitter," in *Proceedings of ICWSM*, pp. 401-408.

Weng, J., Lim, E.-P., Jiang, J., and He, Q. 2010. "TwitterRank: Finding Topic-sensitive Influential Twitterers," *in Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10*, New York, NY, USA, pp. 261-270.

Yao, L., Mimno, D., and McCallum, A. 2009. "Efficient Methods for Topic Model Inference on Streaming Document Collections," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09*, New York, NY, USA, pp. 937-946.

Zeng, D., Chen, H., Lusch, R., and Li, S.-H. 2010. "Social Media Analytics and Intelligence" in *Proceedings of IEEE on Intelligent Systems*, (25:6), pp. 13-16.

Zhang, L., Sun, X., and Zhuge, H. 2013. "Location-Driven Geographical Topic Discovery," in *Proceedings of 2013 Nineth International Conference on Semantics, Knowledge and Grids (SKG)*, pp. 210-213.

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. 2011. "Comparing Twitter and Traditional Media Using Topic Models," in *Advances in Information Retrieval, volume 6611 of Lecture Notes in Computer Science*, P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, and V. Mudoch (eds.), Berlin, Heidelberg: Springer-Verlag, pp. 338-349.