Taylor & Francis
Taylor & Francis Group

Journal of Experimental & Theoretical Artificial Intelligence

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/teta20

A review of approaches for topic detection in Twitter

Zeynab Mottaghinia , Mohammad-Reza Feizi-Derakhshi , Leili Farzinvash & Pedram Salehpour

To cite this article: Zeynab Mottaghinia , Mohammad-Reza Feizi-Derakhshi , Leili Farzinvash & Pedram Salehpour (2020): A review of approaches for topic detection in Twitter, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2020.1785019

To link to this article: https://doi.org/10.1080/0952813X.2020.1785019

Published online: 28 Jun 2020.

JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE
https://doi.org/10.1080/0952813X.2020.1785019

Taylor & Francis
Taylor & Francis Group

ARTICLE

A review of approaches for topic detection in Twitter

Zeynab Mottaghinia, Mohammad-Reza Feizi-Derakhshi, Leili Farzinvash and Pedram Salehpour

Faculty of Electrical and Computer Engineering,University of Tabriz, Tabriz, Iran

ABSTRACT

Online social media such as Twitter are growing so rapidly. Recently, Twitter has become one of the popular microblogging services on the Internet. It lets millions of users to communicate and interact by sending short messages of up to 140 characters. The massive amount of information over the web from Twitter requires an automatic tool that can determine the topics that people are talking about. The Topic Detection task is concentrated on discovering the main topics automatically. In this article at first, we explore different approaches to detect topics of tweets. Then, we will classify these topic detection approaches to four classes of categories, including with word embedding or without word embedding, specified or unspecified, offline (RED) or online (NED), and supervised or unsupervised. Finally, we will discuss the studied approaches in detail.

ARTICLE HISTORY
Received 11 July 2019
Accepted 3 June 2020

KEYWORDS
Topic Detection; twitter; clustering; topic Modelling; word Embedding

Introduction

The influence of microblogging service like Twitter has increased over the past few years and these social media networks facilitate communication between people across the world. The popularity of this new form of social media has also started to attract the attention of researchers. Several recent studies observed Twitter from different outlooks, including the topological characteristics of Twitter (Kwak et al., 2010), tweets as social sensors of real-time events (Sakaki et al., 2010), the estimate of box-office revenues for movies (Asur & Huberman, 2010), and topic detection in Twitter. Due to the fact that Twitter is used heavily to spread information and express opinions during events, this amount of information in microblogs motivates the needs for topic detection systems.

To extract topics from online sources such as Twitter, we need systems or tools that can automatically detect topics or trends. The task of Topic Detection and Tracking (TDT) (J. Allan et al., 2003) has been tackled in the past for well-structured documents and has attracted notable research in several communities in the last era, including machine learning (J. Allan et al., 1998; Blei et al., 2003; Steyvers & Griffiths, 2007; Yang et al., 1982); information retrieval (He et al., 2007; Mori et al., 2004, 2006; Toda & Kataoka, 2005), and social media modelling (Aggarwal & Subbian, 2012; Becker et al., 2010; Hu et al., 2012; Prabowo et al., 2008). However, it is a challenge for both humans and machines to find topics from Twitter due to the large volume of tweets, the noisy, short, and unstructured data, which negatively affect the performance of the detection algorithms.

This article provides a survey of approaches found in the literature for topic detection from Twitter streams. These approaches are classified with word embedding or without word embedding, specified or unspecified, offline or online, and supervised or unsupervised.

This article is organised as follows: In Section 2, we review basic concepts in topic detection methods; Section 3 investigates Twitter and its characteristics and syntaxes. In Section 4 topic

CONTACT Zeynab Mottaghinia, z.mottagy@gmail.com

detection techniques in traditional Medias represented and classified. Section 5 provides a general discussion, followed by a comprehensive conclusion.

Notation and terminology

In this section, we define some terms that are so popular in topic detection literature.

- **Word**: the basic unit of discrete data.
- **Document**: Given a vocabulary W, a document d is a Sequence of some words from W, i.e., $d = \{w_1, w_2, \ldots, w_{|d|}\}$ where $w_i \in W$ $(1 \leq i \leq |d|)$.
- **Corpus**: Collection of some documents.
- **Topic**: A subject discussed in one or more documents. Each topic is supposed to be represented by a multinomial distribution of words.
- **Emerging Topic**: A topic that frequently occurs in the specified time interval and has been relatively rare in the past.
- **Event**: Non-trivial thing that happens at a special date/time in a special location (Sayyadi & Raschid, 2013).
- **BOW (Bag Of Words)**: Collection of words without any ordering.
- **TF (Term Frequency)**: TF is a measure of how often a term occurs in a document.
- **IDF (Inverse Document Frequency)**: IDF for a term in a set of documents is a measure of how much information a term does provide.
- **TF-IDF**: TF-IDF is a Combination of TF and IDF that gives a high score to those terms that are frequent in one or a few documents, but infrequent in the corpus at large.
- **VSM (Vector Space Model)**: The VSM is a model that represents text algebraically. It turns a text document into a vector, where a separate term shows each dimension of the vector.

Twitter

Twitter[1] is a microblogging service founded in 2006; it has gained several hundred million users. Numbers from the first quarter of 2018 show that Twitter has an average of 321 million monthly active users worldwide.[2] Evan Williams (founder of Twitter.com) described Twitter[3]:

What we have to do is deliver to people the best and freshest most relevant information possible. We think of Twitter as it's not a social network, but it's an information network. It tells people what they care about as it is happening in the world.

Users can post messages, called tweets. They are posted by users in response to the question 'What's happening?'[4] These tweets are then shown on the Twitter homepages (or feeds) of any users following them. Any user A can follow any other user B, which makes any tweet posted by user B appear in the feed of user A. The action of the following does not need to be counteracted. It is up to user B whether or not to follow user A back when user A follows them.

Characteristics of twitter

Twitter as a social network is similar to historical diaries (Humphreys et al., 2013), where:

(A) Both of them are semi-public in nature.
(B) Both of them exist in the narrative style.
(C) Both of them are introspectively chronicling activities and trivial day-to-day logs.
(D) The entries are short.

The loss of social interaction and the inability to systematically subscribe to or 'follow' one another are the principal differences.

Twitter is one popular microblogging service that has many unique characteristics such as

(1) *Limited length*: One of the characteristic technological affordances of Twitter is the limited length of messages. A tweet permits only 140 characters, which is almost the length of a typical newspaper headline and subheading (Milstein et al., 2009).
(2) *Noisy vocabulary*: The word limit of tweets causes users to apply abbreviations, slang, and emoticons to compress information. The tweet may also contain spelling error or be grammatically incorrect, which has resulted in a noisy vocabulary.
(3) *User-generated*: since users produce, as well as consume Twitter content, there is a rich source of explicit and implicit information about the user, e.g., demographics (gender, age, location, etc.), interests, and opinions. In some cases, user-generated content is comparatively small.
(4) *Limited friends*: Twitter constrains a limit of 2000 friends for a user, but there is no limit of the number of followers a user can have.
(5) *Different fields*: Twitter includes users from different fields; consists of celebrities, national leaders, and news publishers as well as the general public.
(6) *Multilingual*: Twitter content is multilingual and less than 50% of tweets are in English, with Japanese, Spanish, Portuguese, and German also featuring prominently (Carter et al., 2013).

Twitter's syntaxes

Three prominent syntaxes were created on Twitter, and support tagging, forwarding, and communication between users. These syntaxes are *hashtags, retweets*, and *mentions*.

Hashtags

A hashtag is a single term with a '#' sign prefix in the body of a tweet, denoting the category of a tweet. Hashtags allow users to indicate the topic of a tweet. One can click on any hashtag in a tweet to show all other tweets containing that hashtag.[5] In other words, hashtags make our tweets more searchable by other users with a similar interest.

Retweets

Retweet is specified by "RT" at the start of a tweet and it is an act that users spread when they tweet to their own followers. Retweeting displays a user indirectly expressing their recognition, level of agreement and finding a tweet important or interesting (Kwak et al., 2010).

Mentions

A mention is syntax to refer special users or to address one another by using @username (or @mentions) in the body of the tweet (Boyd et al., 2010). If a user is mentioned in a tweet he will receive a notification about this on his Notifications tab. Mentions are used to reply to users' questions, get their attention, to find persons with specific interests, or to figure out other users' opinions on a special topic (Teevan et al., 2011).

Tweets' pre-processings

As we mentioned in section 3.1, the content of user-generated tweets could be unpredictably noisy. To reduce the amount of noise before the detection pre-processing is executed. The pre-processing includes some steps which have been mentioned as follow:

Tokenization

Tokenization is the first pre-process used in natural language processing (NLP) in order to form tokens known as strings of multiple characters from an input stream, and group these together where it appears that the characters are related to each other (Webster & Kit, 1992). The Tokenizer tools (O'Connor et al., 2010) have

been used to extract bags of cleaner terms from the original messages by removing stop-words and punctuation, compressing redundant character repetitions, and removing mentions, i.e., IDs or names of other users involved in the text for messaging purposes.

Normalization

Tweets with a limited amount of characters contain non-standard words as abbreviations and acronyms or disambiguation of words. These words are not typically found in a dictionary and develop through the time. Text normalization of these acronyms or abbreviations can be done by replacing the shortened word with the contextual appropriate word or word sequence (Sproat et al., 2001).

Stemming and lemmatization

Stemming is the computational process of reducing all words to their root (or stem) and is done usually by stripping each word of its suffix and derivation (Lovins, 1968), there are some stemming algorithms such as Porter stemming (Porter, 2006) and Landcaster algorithms (Hooper & Paice, 2005). Lemmatization is the process of finding the lemma, or the normalization of words such as reduce running to its base form run (Korenius et al., 2004). Stemming is an algorithmic method, while lemmatization is based on lexical analysis.

Aggregation

Topic detection approaches based on word, n-grams co-occurrences, or any other type of statistical inference; suffer in the absence of long documents. In Twitter where user-generated content is typically in the form of short posts, it is common to concatenate different messages together to produce super-documents of larger size.

Detecting topics in traditional medias

Topic Detection and Tracking (TDT) is an integral portion of the DARPA[6] Translingual Information Detection, Extraction, and Summarization (TIDES) program (J. Allan et al., 2003). Many studies in TDT have been performed (Yan, Alhawarat & Hegazi, 2018; F. Wang et al., 2018; Xu et al., 2018; X. Yan et al., 2013). TDT mostly contains two sub-tasks: Topic Detection and Topic Tracking. Topic detection aims at detecting novel topics from text corpus while topic tracking is dedicated to tracking the evolution of existing topics over the temporal dimension.

This section provides a brief overview of topic detection approaches applied to traditional medias outlets. They are commonly classified into document-pivot and feature-pivot approaches depending on whether they rely on documents or keywords. In the first method, a topic is represented by a cluster of documents, whereas in the latter, a cluster of keywords is produced instead.

Document-pivot approaches

Initial approaches for topic detection usually relied on clustering documents. Simple document-pivot approaches cluster documents by leveraging some similarity metric between them. Hierarchical clustering approaches, such as the bottom-up hierarchical agglomerative clustering (HAC) (Dubes & Jain, 1988) have been classified in document-pivot approaches. Each cluster consists of a single data point, and then on the basis of the similarity measure, nearby clusters are merged until all data points become a single cluster or some termination criteria are satisfied. Different variations of the HAC algorithm have been proposed in TDT. For instance, in (Yang et al., 1982) the group average clustering (GAC) has been employed to detect news events from accumulated news stories. In GAC authors also applied an iterative bucketing and re-clustering model (Cutting et al., 1992) to control the tradeoff between cluster quality and computational efficiency. A two-layer HAC approach has also been proposed to reduce the false positives at the cost of increased complexity based on affinity propagation (Dai et al., 2010). The traditional k-means algorithm and its variants, such as k-median and k-means++, have also been applied (Bouras & Tsogkas, 2010). For example, the

approach in (S. Li et al., 2010) is an approach that falls in this class and consists of topics representation model, K-means algorithm for text clustering and TDT evaluation method. It uses vector space model (VSM) for representing topics, and then makes use of the K-means algorithm for text clustering which documents are compared using cosine similarity. The authors study how K in K-means affects topic detection and finally, they use TDT evaluation method (Fiscus & Doddington, 2002) to evaluate system performance. In (Zhang & Li, 2011) Zhang and Li study topic detection according to the news corpus. Documents are clustered using cosine distance on vector space model (VSM) representations by the K-means algorithm. They conclude that topic detection experiments based on large-scale corpus enhance by 38.378% more than topic detection based on small-scale corpus.

Feature-pivot approaches

Feature-pivot approaches are the next generation of TDT approaches that moved the analysis from directly clustering the documents to clustering the terms or keywords. We demonstrate the overview of feature-pivot clustering approaches in Figure 1.

Latent Dirichlet Allocation or LDA (Blei et al., 2003) for short is a probabilistic method which is one of the most widely used feature-pivot techniques. LDA is a three-level hierarchical Bayesian model (Blei et al., 2003; Steyvers & Griffiths, 2007) that can identify latent information about topics in great collections of documents. Each document is represented as a finite mixture over an underlying set of latent topics, where each topic is characterized by a distribution over words. The algorithm uses the Bag Of Words (BOW) to represent the documents, which is acceptable when you are dealing with larger documents where it is possible to explore co-occurrences at the document level. With this LDA achieves a clear outline of all topics related to a document (Titov & McDonald, 2008). LDA has been a successful algorithm for topic modelling. The MALLET topic model package (McCallum, 2002) provides an implementation of LDA, which uses Gibbs sampling as the inference engine; it is labelled LDA-GS. It also includes efficient approaches for document-topic hyper-parameter optimization. Another variant of LDA has been provided, which uses Variational Expectation Maximization (VEM) for latent Dirichlet allocation (LDA-Blei)[7]; it is labelled LDA-VEM. It also performs hyper-parameter optimization.

KeyGraph (Sayyadi & Raschid, 2013) is a topic detection technique that considers keyword co-occurrence. It has consisted of three phases: building the KeyGraph; extracting topic features and assigning topics to documents. It represents a collection of documents as a keyword co-occurrence graph labelled a KeyGraph, then applies an off-the-shelf community detection algorithm to group co-occurring keywords into communities. At last, the algorithm uses the topic features to assign topics to documents. If there are multiple documents that are assigned to each pair of topics, it assumes that these are subtopics of the same parent topic, and merge these topics. In KeyGraph only terms have been used in each document as features. Another variant of KeyGraph is KeyGraph+ that

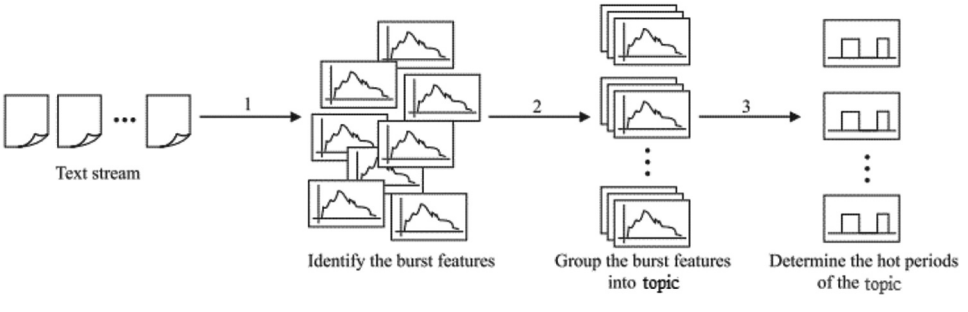

Figure 1. The Overview of Feature-Pivot Clustering Approaches.

uses terms as well as noun phrases and named entities as features in each document; KeyGraph+ increases the weight of named entities and noun phrases by doubling their term frequencies.

A hybrid relations analysis approach to integrating semantic information and co-occurrence information among terms for topic detection has been proposed in (Zhang et al., 2016). Specifically, the method fuses multiple relations into a uniform term graph by combining ID theory with the topic modelling method and detects topics from the graph using a graph analytical method. First of all, an Idea Discovery algorithm (Idea-Graph) is adopted to mine co-occurrence relations (especially latent co-occurrence relations) for converting the corpus into a term graph. After that, an LDA-based semantic relations extraction method enriches the graph with semantic information. Finally, a graph analytical method is exploited by the graph for topic detection. Two variants of this approach have been evaluated: (1) IG: it only employs Idea-Graph to generate the term graph, and (2) LDA-IG: it has equipped with Idea-Graph and LDA.

Figure 2 demonstrates the performance of some feature-pivot approaches based on Precision, Recall, and F1 score. F1 score is calculated as the harmonic mean of precision and recall. The methods have been compared on the renowned TDT4 dataset (Yang et al., 1982) which is a well-annotated and small dataset. All these approaches are equipped with pre-processing. The results in Figure 2 show that LDA-IG outperforms other methods for Recall and F1 score metrics. The improvement of LDA-IG in F1 score is over IG, it is expected that LDA-IG can leverage semantic information for topic detection. The advantage of LDA-IG is benefited from taking hybrid term relations when detecting topics. However, the other approaches used either co-occurrence relations or semantic relations, which led to the detection of incomplete information. Also, from Figure 2, we note that using noun phrases and named entities as features in KeyGraph+ can slightly improve the performance of this method over KeyGraph.

Differences between document pivot and feature-pivot approaches

It is not easy to conclude on which class, Document pivot or Feature pivot, produces the best results. Both Document-pivot and Feature-pivot approaches have advantages and disadvantages (Fung et al., 2005). Document-pivot approach disadvantages are sensitivity to noise and cluster fragmentation and, in a streaming context, they depend on arbitrary thresholds for the inclusion of a new

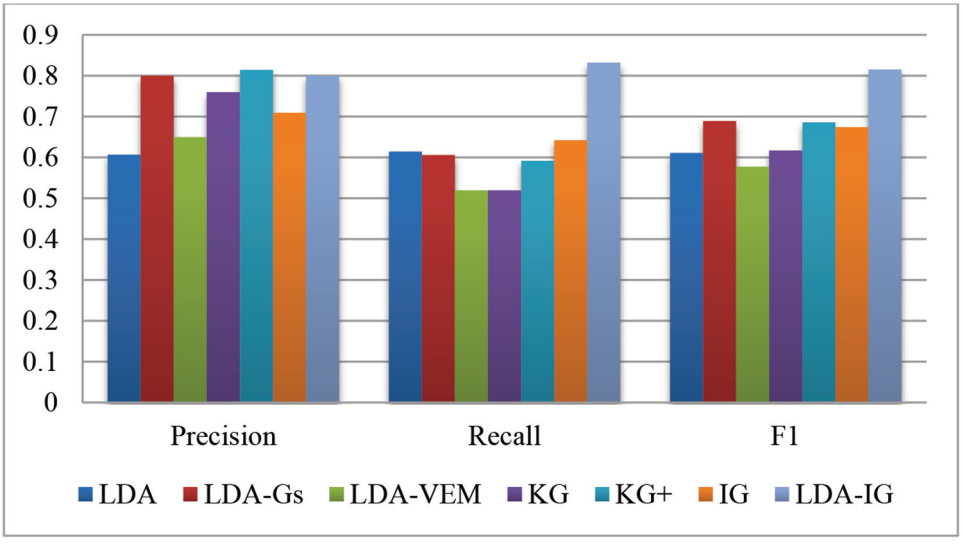

Figure 2. The performance of some feature-pivot approaches.

document to an existing topic. On the other hand, feature-pivot approaches are commonly based on the analysis of associations between terms and capture misleading term correlations. Overall, the two approaches can be considered complementary and, depending on the application, one may be more suitable than the other.

Detecting topics in twitter

Automatic topic detection presents an interesting task for social media streams. Many approaches for topic detection in Twitter service have been proposed recently. They are mainly classified into several classes based on short text representation, type of events, detection method and detection task. These classes are approaches without word embeddings versus with word embeddings, specified versus unspecified, supervised versus unsupervised (or combination of them), and online versus offline topic detection approaches.

Short text representation

Tweets are short texts that have become a fashionable form of information on the Internet. These texts are limited contexts and have data sparsity. In this section, we review topic representations on short texts as shown in columns 2 and 3 of Table 1.

Topic detection approaches for short texts without word embeddings

Word representation is crucial to natural language processing. In the past decades, several models are proposed, such as Bag-of-Words (BOW) (Harris, 1954), and TF-IDF (Salton & McGill, 1983). The default approach of representing words as discrete and distinct symbols is insufficient for many tasks, and suffers from poor generalization. Also, most of the conventional topic detection approaches rely on word co-occurrences to obtain topics from a collection of documents. Due to the short length of tweets (short texts), they are sparse in terms of word co-occurrences. It impacts the efficiency of conventional topic detection approaches. The approaches described in the following attempted to address these challenges.

TwitterStand (Sankaranarayanan et al., 2009) focuses on detecting news topics around specific geographical area. It employs a naive Bayes classifier to separate news from irrelevant information and an online clustering algorithm is used, which assigns the news-related tweets to the closest cluster if the distance to this cluster is smaller than a given threshold. Then, a new cluster with this tweet as the only member is created. The distance between a cluster and a tweet is based on the words in the tweet and the time at which the tweet was posted. The obtained clusters are considered as specific topics. Finally, for each obtained topic, additional relevant tweets are searched using the hashtags present in the tweets of its corresponding cluster. The system uses TF-IDF weightings to decrease the effect of popular terms. TwitterStand does not require a query; its topic detection performance depends on the pre-selected seeder list. Fragmentation of clusters and sensitivity to noise are some disadvantages of this method.

The work by Phuvipadawat and Murata (Phuvipadawat & Murata, 2010) provides a technique for breaking news detection in Twitter. The task has been presented in three phases: sampling, indexing, and grouping. First, tweets retrieved using targeted queries or hashtags. Then, keywords' content has been indexed with Apache Lucene.[8] At last, similar messages grouped together to form a news story based on boosted tf-idf, emphasising significant entities such as names of countries or public figures by using the Standard Named Entity Recognizer (NER) (Finkel et al., 2005). Tweets are incrementally merged by considering the textual similarity between incoming tweets and existing clusters. A software that enables a user to see the storyline of each topic based on the proposed method is called Hotstream.

Petrovic et al. (Petrović et al., 2010) use the streaming First Story Detection (FSD) model to analyse Twitter in real-time. The authors presented an algorithm based on locality sensitive hashing (LSH) (Indyk & Motwani, 1998) to reach constant processing time. LSH is a randomized technique that

Table 1. Classification of Topic Detection Approaches.

	Short text representation		Type of events		Detection method		Detection task	
References	Without embedding	With embedding	Specified	Unspecified	Supervised	Unsupervised	NED	RED
Sankaranarayanan et al. (Sankaranarayanan et al., 2009)	*			*	*	*	*	
Phuvipadawat and Murata (Phuvipadawat & Murata, 2010)	*			*		*	*	
Petrović et al. (Petrović et al., 2010)	*			*		*	*	
Sriram et al. (Sriram et al., 2010)	*		*		*			*
Rakesh et al. (Rakesh et al., 2013)	*		*			*		*
Cataldi et al. (Cataldi et al., 2010)	*		*		*	*		*
Chen et al. (Chen et al., 2013)	*		*		*		*	
Lee et al. (Lee et al., 2011)	*			*	*			*
Tembhurnikar and Patil (Tembhurnikar & Patil, 2015)	*			*		*		*
Ishikawa et al. (Ishikawa et al., 2012)	*		*			*	*	
Kim et al. (Kim et al., 2013)	*		*			*		*
Rill et al. (Rill et al., 2014)	*		*		*		*	
Benny and Philip (Benny & Philip, 2015)	*		*			*		*
Xie et al. (Xie et al., 2016)	*		*		*		*	
Yin et al. (Yin et al., 2013)	*		*			*		*
Nur'aini et al. (Nur'Aini et al., 2015)	*			*		*		*
Cigarrán et al. (Cigarrán et al., 2016)	*			*		*	*	
Geng et al. (Geng et al., 2017)	*			*		*		*
Hasan et al. (Hasan et al., 2019)	*		*			*	*	
Yan et al. (D. Yan et al., 2016)		*				*	*	
Kumar and Sridhar (Sridhar, 2015)		*	*			*		*
Nguyen et al. (Nguyen et al., 2015)		*	*		*			*
Hu et al. (Hu et al., 2018)		*		*	*			*
Zhao et al. (Zhao et al., 2018)		*		*	*			*
Ma (Ma, 2018)		*	*		*			*
Gencoglu (Gencoglu, 2018)		*	*			*		*

reduces the time needed to find the nearest neighbour in vector space and keeps the memory use constant to save space. In experiments, they did not consider replies, retweets, and hashtags. The paper showed that the method is able to overcome the limitations of traditional approaches and could reach significant speedup. Also, the results have shown that ranking according to the number of users is better than ranking according to the number of tweets and the amount of spam messages in output has been decreased by considering entropy of the messages.

Sriram et al. (Sriram et al., 2010) propose a tweet classification algorithm. This algorithm automatically classifies incoming tweets to a predefined set of topics such as News, Events, Opinions, Deals, and Private Messages using Twitter related metadata about the authors (e.g., name, information in the profile). As the experimental results show, the authorship plays an important role in classification. Each author usually has a special tweeting pattern and most of his tweets tend to be in a limited set of topics. Also, the classification accuracy is high even without meta-information.

Location-Specific Tweet Detection and Topic Summarization in Twitter proposed in (Rakesh et al., 2013). It identifies and summarizes tweets that talk about geographical location. A weighting

scheme called Location Centric Word Co-occurrence (LCWC) has been suggested that uses Mutual Information (MI) score of tweet bi-grams; the term frequency (TF) of tweets; the tweet's inverse document frequency (IDF), and the user's network score to define the location-specific tweets. It then uses a topic model based on Latent Dirichlet Allocation (LDA) to predict the topics from location-specific tweets and present them using a web-based interface. This paper has three important findings: (1) top trending tweets from a location cannot be a good feature for predicting location-specific tweets, (2) ranking tweets purely based on geo-location of users' cannot determine the location specificity of tweets tweeted by that users, and (3) users' network information plays a significant role in determining the location-specific characteristics of the tweets.

A paper in (Cataldi et al., 2010) provided an approach to detect the real-time emerging topics that relied on five stages. First, it extracts and formalizes the content (set of terms) of tweets as vectors of terms with their relative frequencies; second, it describes a directed author-based graph based on active authors social relationships, and estimate their authority based on the Page Rank algorithm (Page et al., 1999); third, the life cycle of each term is modelled according to the content aging theory that leverages the users' authority to study its usage in a specified time interval; forth, set of emerging terms are selected by ranking the keywords depending on their life status; and finally, it creates a navigable topic graph that connects the emerging terms with their co-occurrent ones in order to detect a set of emerging topics under user-specified time restriction. In this system, a topic is defined as a coherent set of semantically related terms that express a single argument.

(Chen et al., 2013) is a framework to detect user-defined hot topics from microblogging posts which includes three steps: Automatic Keywords Expansion (AKE), User-Defined Microblogging Filter and Hot Topic Detection Algorithm. User-defined hot topics mean specific topics which satisfies user's personalized preferences; these preferences are predefined by users and represented by several keywords. For automatic keywords expansion, authors presented an algorithm that effectively extracts relevant words by fusing importance, relevance, and penalty factors. The second step filters topic-related microblogging through the expanded terms by using the logistic regression classifier. If microblogging contains topic word or expansion term, it is in the positive document set and vice versa. At last, hot topic detection algorithm consist of two subtasks: Heat Words Detection and Heat Words Cluster. In this step, the heat value of a word calculated based on word frequency and change along the time, when the heat value of a word at a certain time is greater than the predefined threshold, it is more likely to be a heat word, and contrariwise. It then groups the heat words in different events by a classical clustering algorithm, and Affinity Propagation (AP) (Frey & Dueck, 2007; Givoni et al., 2012).

'Twitter Trending Topic Classification' was proposed by Lee et al. (Lee et al., 2011). The approach classifies Twitter trending topics into 18 general and predefined classes such as sports, politics, technology, etc. This system consists of four phases: (1) *Data Collection* that collected data such as tweets, trending topic, and topic definition; (2) *Labelling* that labelled topics manually into 18 classes; (3) *Data Modelling* which has two methods (i) *Text-based Modelling* that constructs word vectors with trending topic definition and tweets and converts to tokens with tf-idf weights (ii) *Network-based Modelling* that identifies five most similar topics for each trending topic based on the number of common influential users; (4) *Machine Learning* which applied various classification methods such as Naive Bayes Multinomial, C5.0 (an improved version of C4.5) decision tree learner (Quinlan, 1996), k-Nearest Neighbour (KNN) (Aha, Kibler et al. 1991), Support Vector Machine (SVM) (Cristianini & Shawe-Taylor, 2000), Logistic Regression (Le Cessie & Van Houwelingen, 1992), and ZeroR (the baseline classifier) using 10-fold cross validation to find the best classifier. Experiments show that classification accuracy of up to 65% and 70% can be reached using text-based and network-based classification modelling respectively.

In (Tembhurnikar & Patil, 2015) the authors proposed topic detection and sentiment analysing method using BNgram. The components of the method are Data Preprocessing, BNgram, and Sentiment Analysis. In the first step, different preprocessing techniques such as tokenization, slang words translation, stemming, URL removal, and aggregation have been used. The results show that

preprocessing of the data affecting the quality of detected topics. After that, BNgram topic detection method takes into account the simultaneous co-occurrences between more than two terms by using n-grams and the changing frequency of terms over time as a useful source of information to detect emerging topics by using df-idft metric which introduces time to the classic tf-idf score. This approach indexes all keywords from the posts of the collection by using Lucene which are organised into different time slots. Also, the boost factor is considered to raise the importance of proper nouns (persons, locations, and organisations) using a standard named entity recognizer. As a consequence of this process, a ranking of n-grams is created based on their df-idft scores. A single n-gram is often not very informative, but a group of them often offers interesting details of a story. So, it uses a clustering algorithm to group the most representative n-grams into clusters that each representing a single topic. Every n-gram has been assigned to its own particular cluster, and then follows a standard 'group average' hierarchical clustering algorithm (Murtagh, 1983) to iteratively find and merge the closest pair of clusters. The procedure of clustering is repeated until the similarity between the nearest unmerged clusters drops below a fixed threshold value. Finally, the clusters are ranked based on the highest df-idft score of the n-grams contained in the cluster. In the sentiment analysis phase, the polarity of the sentiments has been calculated and they have been classified as positive, negative, and neutral. From this, we can consider the views of persons related to particular topics. This method also has been implemented for detecting trending topics on Indonesian tweets (Winarko & Pulungan, 2019). The results show that the use of trigram obtains the highest quality.

The approach employed by Ishikawa et al. (Ishikawa et al., 2012) uses Twitter and Wikipedia as the information source and detects hot topics in local areas and during a particular period. The basic approach is to classify tweets into topics and to select top topics ranked according to the number of the topics' tweets. Due to the semantic fluctuation of tweets, this classification has some problems. For instance, different words that used with the same events will be categorised as different topics. In this approach, some fluctuations, such as spelling, spatial, and temporal have been identified and the clustering method to manage these fluctuations has been proposed. The presented hot-topics detection approach described as follows. Tweets with geolocation information are collected and unwanted tweets are eliminated, and then morphological analyses are performed and only nouns and verbs are focused. Afterwards, the hot topics detection method is done. The detection method has three procedures. (1) Building a relationship among a set of words by means of weight function, (2) Classifying the words into topics using a clustering algorithm, this clustering algorithm presents a similarity degree of association between words and topics. The authors adopt an incremental clustering method such as that proposed in Allan et al., 1998, rather than recreating the sequential clustering, and (3) detecting hot topics using a burst detection method that has been proposed in (Kleinberg, 2003). This method detects whether the interval of the arriving messages is denser than that in a normal condition through comparison with other document streams such as current news articles and bulletin board threads. The experiments illustrate that spatial words often appeared in all data range, while temporary words related to some events. Also, the result illustrates that tweet with some topics only bursts temporarily so that hot topic could be collected.

Kim et al. (Kim et al., 2013) have studied the geographic clustering analysis based on social topics across the United States (US). They presented a method that extracts social hot topics of every day which are not tweeted in the prior day by means of word frequency's ratio. In this method, the non-topic keywords like small talks or emotional words (e.g., lol, love, oh, etc.) are repressed, while identifying topic words. The results show that unlike daily life topics, the frequency of emotional words is relatively constant over time. In addition, the authors exploited the concept that the words with the same frequencies during a period of time can be grouped together, so, the geographic communities for topics are detected. This approach needs a long period of the observation and also finds general topics like Easter, sports, etc.

Politwi system described in (Rill et al., 2014) is designed to detect political topics emerging (Top Topics) in Twitter. The system has the modules Data Selection, Preprocessing, Analysis, and Presentation. In data selection module tweets are collected and stored in a raw database. The

preprocessing module consists of source analyser, hashtag analyser, sentiment hashtag analyser, emotion analyser, and analyse database. The source analyser excludes the automatically generated tweets, and the hashtag analyser extracts all hashtags contained in the tweets. Also, sentiment hashtag analysis is a component that detects the polarity of topics marked by hashtags. Moreover, the emotion analyser extracts information about emotions inside of the tweets. The results of preprocessing are stored in analyse database and used in topic detection later. In the analysis module, the topic detection is done. In this project, hashtags are used as a candidate for Top Topics and if each Top Topic have a Topic Value (TV) greater than 0.0 are selected. A Topic Value is calculated for each hashtag occurring in the tweets by comparing the current number of tweets with hashtag H to the number of tweets of the previous period. In the module presentation, Top Topics presented in the wider public such as Twitter channel, website, and smartphone apps. The results show that new topics appearing in Twitter can be detected right after their occurrence.

A pattern-based method to extract the topics related to certain keywords is proposed in (Benny & Philip, 2015). For the first step, the proposed method collects tweets using a specific keyword. The next step is the identification of frequent patterns by using tf-idf. After that, it evaluates the AGF (Associative Gravity Force) values between each pattern pairs. AGF represents the attraction between each pair. If the AGF is large, that means, the attraction between each pair is very high, i.e., the chances of occurrences of each pair together are very high. So, in this paper, two algorithms TDA (Topic Detection using AGF), and TCTR (Topic Clustering and Tweet Retrieval) are used to extract topics using AGF values. The goal of the TDA algorithm is generating the topics by clustering the frequent patterns. But these clusters may also suffer from the wrong correlation problem of patterns. The method used to avoid this problem is TCTR. The TCTR clusters the topics and returns the corresponding tweets as final outputs.

Xie et al. (Xie et al., 2016) proposed TopicSketch, Real-time Bursty Topic Detection method. TopicSketch consists of two principal techniques, a sketch-based topic model and a hashing-based dimension reduction technique. The first technique has two phases; firstly, it maintains the acceleration (arriving rate) of each pair of words and each triple of words as a sketch of the data, which are indicators of popularity surge as early as probable and can be updated efficiently at a low cost, making real-time detection possible; secondly, the method learns the bursty topics by tensor decomposition algorithm (Anandkumar et al., 2014), based on the data sketch. A dimension reduction technique has been proposed to perform the detection efficiently in large-scale real-time setting; this technique is based on hashing (Schubert et al., 2014) which provides scalability with preserving the topics' quality. The experiments show that, due to the single topic assumption, TopicSketch is not appropriate for a stream of documents with multiple topics.

In (Yin et al., 2013) the authors differentiate between stable topics and temporal topics due to their different natures. Stable topics are on users' regular interests and their routine discussions, which remain stable popularity over time. Temporal topics are on popular real-life events or hot spots, such as emergencies, politics, and public events. Yin et al. present a unified mixture method to detect both stable and temporal topics simultaneously from social media data. For stable topics, they focus on its user who generates it, and for a temporal topic, they focus on when it is generated. Then, to enhance the method they propose a regularization framework that exploits both spatial and temporal information and a burst weighted smoothing scheme. The enhancements use two insights: the user topic distribution and the temporal topic distribution change smoothly along the spatial and the temporal dimensions, correspondingly.

In (Nur'Aini et al., 2015) a method to detect topics from tweets is proposed that combines the singular value decomposition (SVD) and k-means clustering methods. After the preprocessing phase, firstly it reduces the dimension of the word-tweet matrix by truncated SVD; secondly, the tweets in the reduced matrix, that is a latent semantic tweet matrix, clusters using K-means clustering algorithm; thirdly, the discovered centroids transform into the original dimension of tweets, that is, the words, because the latent semantic tweet matrix is not displaying the relationship between

words and tweets, consequently, the centroid of each cluster cannot be directly interpreted as a topic; and finally, a specific topic has been shown by most heavily weighted words in each centroid.

Cigarrán et al. (Cigarrán et al., 2016) proposed a novel approach for Topic Detection based on mathematical theory named formal concept analysis (FCA) (Ganter & Wille, 2012; Wille, 1992, 2009). In this approach, they take tweets as objects, and their terms as attribute; and they consider formal concepts as topics. The proposed system does the following steps: (a) it pre-processes the data collection, (b) for each entity, a formal context is generated, (c) the system reduced formal context by term selection strategy, it selects the smallest set of features which describes the largest number of objects, (d) the reduced formal context used for FCA computation and a concept lattice for each entity is gained, (e) from the concept lattice a set of topics which have a stability value (Kuznetsov, 2007; Kuzuetsov, 1990) higher than the selected threshold will be selected, a high stability value means that the concept represents a proper topic because it represents a cohesive set of tweets, and (f) the selected topics will be used to make the final result list. If a tweet is contained in different topics, different selection strategies are applied. The wide analysis of the results shows that there is a direct correlation between the precision of the approach and information; the system has adaptability when new topics appear; the approach integrates previous knowledge of the prior topics, and the results are the best when the Reliability and Sensitivity values tend to be equal. The authors in the other paper (Castellanos et al., 2017) evaluated this method from different aspects. The evaluation carried out by taking into account the internal clustering validity metrics to analyse the quality of the topics detected. This evaluation shows that this approach creates cohesive clusters, which are less subject to changes in cluster granularity.

A three-layer hybrid clustering algorithm (K-H-K) (Geng et al., 2017) has been proposed for topic detection in microblogs. In the first layer, it uses the K-means algorithm to group the large-scale microblog texts into many smaller and purer clusters at a high speed. In the second layer, it uses the AGNES (agglomerative nesting) algorithm to merge the small clusters consisting of texts of the same topic. In the last layer, it reassigns the texts assigned to the wrong cluster by the K-means. This method does not need to set the cluster numbers and it has high-quality efficiency.

A real-time event detection system called TwitterNews+ has been proposed in (Hasan et al., 2019). The two components of TwitterNews+ are the Search Module and the EventCluster Module. The search Module includes detecting a burst in the number of tweets discussing an event. The tweet decided as 'not unique' shows that similar tweets have been visited before and this tweet is sent to EventCluster Module. The eventCluster Module includes clustering the tweets that discuss the same events and tracking these events (Hasan et al., 2016). Furthermore, the authors did a wide parameter sensitivity analysis on the different parameters to fine-tune the parameters used in TwitterNews+ to improve its performance in terms of recall and precision.

Topic detection approaches for short texts with word embedding

Alternative distributed word representation, namely word embedding or continuous space representation of words is a set of language modeling and learning techniques in NLP where words can be mapped to vectors in a low dimensional space (Suzuki & Nagata, 2015) and so they are simple to work with. Word embedding is introduced in (Elman, 1991; Rumelhart et al., 1988) and plays an important role in learning continuous word vectors based on their contexts in a large corpus by using the word2vec model (Mikolov, Mikolov, Chen et al., 2013; Mikolov, Sutskever et al., 2013) or Glove[9] Model (Pennington et al., 2014). The typical characteristic of word vectors is that semantically similar words would be similar to each other by calculating cosine similarity among the vectors of words. However traditional bag-of-words or bag-of-ngrams hardly capture the semantics of words or the distances between words. This kind of word embeddings named classic word embeddings. The major drawback of classic word embeddings is that the same word will always have the same representation regardless of the context where it occurs. And so, contextualized word embeddings such as ELMO,[10] FlairEmbeddings and BERT[10] (Akbik et al., 2018; Devlin et al., 2018; Gardner et al., 2018; Peters et al., 2018) came into the picture and helped overwhelmed the limitation of the classic word embedding methods like Word2vec and Glove. These new word embeddings methods take into consideration the context of the word and can be seen as dynamic word embeddings methods, most of which make use of some language model to help modeling the representation of a word. The approaches detailed later exploit word embedding.

Paper (D. Yan et al., 2016) focuses on improved Single-Pass topic detection and tracking of Chinese Microblog called MC-TSP to reduce the disadvantages of traditional Single-Pass methods. Traditional Single-Pass is appropriate for processing time stream data and does not need to specify the number of clusters compared with other clustering algorithms, but it has some shortcomings: (1) the sequence of input text has an influence on the clustering performance of Single-Pass algorithm and (2) the cost of calculation is high, and it is time-consuming. In this paper, at first, the work prior to clustering, including pre-processing, feature selection, feature vector representation of short text and similarity calculation, has been done. In feature selection, an Incremental TF-IWF-IDF of terms part-of-speech and position weight calculation method have been proposed; TF and IDF are defined in section 2, and IWF denotes word frequency at a certain time. Also, the paper proposes a new feature vector representation method to consider the semantic and context of terms based on TF-IWF-IDF. The authors use Cosine similarity to compute the similarity between two microblogs. MC-TSP is an improved online TDT method that consists of two steps to overcome the problems of the traditional Single-Pass algorithm. In order to avoid the influence of sequence of text input, MC-TSP introduces the time window; thus, each time the data is clustered according to the batch. It also introduces the second Single-Pass for clustering to overcome the second shortcoming. First, the initial clustering is carried out to form a plurality of micro-clusters. It can reduce the number of similarity calculation with all the text in the existing clusters. Second, the center vector of the cluster has been extracted to denote all text in clusters, which can decrease the cost of calculations. Overall, MC-TSP has two stages: (1) Initial clustering and (2) Merging clustering. In the first stage, the Single-Pass algorithm has been used to do clustering initially with the micro-blogs in the time window and the micro-clusters with multiple topic centers can be created. In the second stage, the similarity of the micro-clusters and the cenetr vector of the existing clusters are calculated. According to the threshold value, the micro-cluster is merged into the existing historical topic to get the final clustering result. The merging clustering improves the quality of initially clustering. Experimental results display that for microblog dataset, the MC-TSP algorithm has better performance than the unimproved Single-Pass algorithm by about 10% in online topic detection and tracking.

Sridhar (Sridhar, 2015) presented an unsupervised, language-agnostic and scalable topic model for short texts like Twitter that utilizes soft clustering over distributed representations of words. In contrast with conventional topic modelling methods that need aggregation of short messages to avoid data sparsity in short documents, the proposed method works on large amounts of raw short texts (billions of tweets) without any aggregation scheme. This method uses the Continuous Bag-Of-Words (CBOW) log-linear model (Mikolov, Mikolov, Chen et al., 2013) to obtain the distributed word representations and Gaussian Mixture Models (GMMs) to parameterize the vector space represented by the distributed representations. The K components of the GMMs can be considered as the latent topics that are captured by the model; that is, GMMs can learn the latent topics by clustering over the distributed representations that are trained with a semantic similarity objective (contextual and positional similarity). This method works well on Twitter data for English, Spanish, French, Portuguese, and Russian.

The authors in (Nguyen et al., 2015) proposed two probabilistic topic models, LF-LDA and LF-DMM, that merge a latent feature vector model with two Dirichlet multinomial topic models: LDA (Blei et al., 2003) and a one-topic-per-document DMM[11] (Nigam et al., 2000) model. In these new methods, the topic-to-word Dirichlet multinomial component of LDA and DMM that generates words from topics has been replaced with a two-component mixture of a topic-to-word Dirichlet multinomial component and a latent feature component. Latent feature vectors are based on deep neural networks and have been learned by predicting words given their window-based context (Collobert & Weston, 2008; Liu et al., 2015; Mikolov, Sutskever et al., 2013; Pennington et al., 2014). These models use two pre-trained word vectors, Google word vectors,[12] and Stanford vectors,[13] to improve the word-topic mapping learned on a smaller corpus. Experimental results show that the latent feature word vectors improve topic modelling performance. A disadvantage of this approach is that it is slow on very large corpora.

In (Hu et al., 2018) SVMCNN method has been presented for short text classification by combining Convolutional Neural Networks (CNN) and Support Vector Machine. In this model, CNN has three convolutional layers that extract features of short texts. The input of CNN is the matrix of concatenated word vectors. For producing different feature maps, different filters are applied to a window of word vectors. The parameters are reduced by the max-pooling layer and input into the fully connected layer for which output is the feature vector. After feature extraction by CNN, every short text can be represented as a 384-dimensional vector. These vectors can be input into the SVM classifier and classify the short text features.

SW-CNN and WW-CNN classification methods (Zhao et al., 2018) have been proposed by combining the Word2vec trained word vector with the Convolutional Neural Network (CNN) model. W-Word2vec (WW) and Seq-Word2vec (SW) algorithms overcome the lack of word order and position considerations in the Word2vec and then use CNN to classify text. Experimental results show that the accuracy of WW-CNN is faintly higher than that of the SW-CNN. Also, SW-CNN and WW-CNN methods have increased the accuracy of 2.7% and 3.3% compared with that of traditional CNN on Multi-label classification of short texts in social media.

In (Ma, 2018), Ma applied deep learning techniques for Tweets classification in the Field of Tweets Classification. In particular, BERT which has 12 transformer layers, 12 self-attention heads, and with a hidden size 768 is used for transfer learning. Results show that BERT-based classifiers could attain better performance compared with the baseline bidirectional LSTM with pre-trained Glove Twitter embeddings. Subjectivity and ambiguity affect the performance of BERT-based models.

In (Gencoglu, 2018) Gencoglu proposed deep 2D convolutional autoencoders (CAEs) based feature extraction, i.e., representation learning, for clustering of health tweets. CAEs consists of two parts, the encoder, and the decoder. The encoder consists of three 2D convolutional layers with 64, 32, and 1 filters. The decoder has the same symmetry with three convolutional layers with 1, 32, and 64 filters, and an output convolutional layer of a single filter. CAEs enhanced the performance of three clustering algorithms, namely, k-means, Ward and spectral clustering when compared to conventional tweet representation schemes including bag-of-words, term frequency-inverse document frequency, Latent Dirichlet Allocation, and Non-negative Matrix Factorization. The experimental results showed that the clustering performance using the proposed representation learning method outperforms conventional methods.

Specified versus unspecified topic detection approaches

Depending on the type of topics, topic detection approaches are classified into two classes, specified and unspecified, as shown in columns 4 and 5 of Table 1. The specified topic detection approaches rely on specific information such as location, time and etc. that are provided by users. On the other hand, unspecified approaches have not prior information.

Supervised versus unsupervised topic detection approaches

In general, machine-learning tasks are typically classified into two categories; these are supervised learning and unsupervised learning. Supervised learning is the task of inferring a function from labeled training data. On the other hand, unsupervised machine learning is the task of inferring a function to describe hidden structure from unlabeled data. In this section, we categorize the topic detection approaches from Twitter into supervised and unsupervised learning or combination of them according to the detection method (columns 6 and 7 of Table 1). In supervised approaches to filtering out the noisy data, several classifiers are utilized such as support vector machines, naive Bays (Sankaranarayanan et al., 2009) and logistic regression (Chen et al., 2013). The proposed supervised topic detection approaches usually assume a static environment. A single classifier is typically trained offline on a small batch of Twitter data over a few weeks or months and then filtered and labeled manually. The classifier is then deployed for detecting topics. Unlike supervised detection, detection is unsupervised; that is there are no training documents or queries and they do not require labeled data. In addition, these techniques have the advantages of not depending on a training corpus and, therefore, are robust and likely to be applicable to real-world settings.

Online versus offline topic detection approaches

TDT usually falls into two distinct categories based on detection task and target application, Offline or Retrospective Event Detection (RED) and Online or New Event Detection (NED). As shown in columns 8 and 9 of Table 1, Offline or RED is focused on discovering previously unidentified topics from the accumulated historical collections (Yang et al., 1982), while Online or NED algorithms deal with documents processed online or in other words before looking at any subsequent documents. On the other hand, in offline approaches, all the data are collected and then all the collected data are analysed entirely at once. It is not surprising that Yang et al. (Yang et al., 1982) demonstrate that the consequences of retrospective detection are much well than the online one, as more information is available from a retrospective way. In these approaches, you have definite liberties in how to perform the analysis due to the availability of all data at the start. Because the data set size is known, you are able to tweak size-dependent parameters in your analysis tool to better fit the data set. Furthermore, offline approaches enable you to make various passes of the data to improve the accuracy and precision of the analysis. The disadvantage of these approaches is that as you need to gather all the data you wish to process in advance, it is unsuited for processing streams of data arriving in real-time. In online approaches, the data as it arrives from the source are analysed (Babcock et al., 2002; H.-F. Li et al., 2004). This means that these approaches are able to discover the topic from real-time streams.

Discussion

In this section, we discuss the approaches studied above (Tables 2 and 3). Table 2 provides an overview of detection techniques and feature representations. We begin by presenting a detailed description of the criteria used:

a) Detection Techniques: Four detection techniques have been used in the discussed methods that are categorised into Classification Based, Clustering Based, Topic Modelling Based, and FCA Based. In machine learning, classification is the problem of identifying a new observation belongs to which of a set of categories, on the basis of a training set of data containing observations or instances whose category membership is known. The dataset may have various data, such as image (Yu et al., 2018; M. Zhang et al., 2018; Zhang et al., 2018; Zhou et al., 2019), video (Karpathy et al., 2014; L. Wang et al., 2018), sound (Piczak, 2015, 2019; Salamon & Bello, 2017), and text. Classification techniques base their operation on the analysis of training (annotated) dataset to learn a classification function. In these methods, there is a series of predefined topics or classes (e.g., sports, economics, education, etc.) which have to be known in advance. Classification-based methods have a high performance in classifying content according to the features seen in the training set. But, if a new topic including new features, unseen in the training set, appears, the system will not be able to classify it correctly. On the other hand, in Clustering-Based methods a topic is represented as a cluster of related tweets, keywords, hashtags, phrases, or segments. Clustering-based methods do not need to be trained to compute the categorisation, thus detecting the topic will not be restricted to the training dataset and therefore they will be robust and likely to be applicable to real-world settings. Most of these algorithms need to predefine the number of clusters or a similarity threshold to create a final cluster partition. Topic modelling can be defined as a method for discovering a group of words (i.e., topic) from a collection of documents that best represents the information in the collection; these documents can be images (Bahrehdar & Purves, 2018) or texts (Marwick, 2013). In Topic Modelling methods, it is common to use a probability distribution over words to represent a topic. These methods act under the assumption that some latent topics exist in the tweets that are processed. Topic modelling methods tend to over-generalize; that is, to generate a quite generic topic (Chemudugunta et al., 2007). Other limitations are: How many topics should be generated? (Guo et al., 2013), and How can the methods take prior knowledge of topics into account? The applications of probabilistic techniques such as (Sridhar, 2015), known as Probabilistic Topic Modelling (PTM), and they are a subset of Topic Modelling methods. Also, in the context of unsupervised approaches, there

Table 2. Overview of detection techniques and feature representation.

References	Detection techniques	Representation	Twitter specific features
Sankaranarayanan et al. (Sankaranarayanan et al., 2009)	Naive Bayes classifier and online clustering	Term vector	Time Hashtags
Phuvipadawat and Murata (Phuvipadawat & Murata, 2010)	Online clustering	Term vector	Time Hashtags Retweets
Petrović et al. (Petrović et al., 2010)	Online clustering (based on locality sensitive hashing)	___	___
Sriram et al. (Sriram et al., 2010)	Naive Bayes classifier	Raw data	Time Username
Rakesh et al. (Rakesh et al., 2013)	weighting scheme called Location Centric Word Co-occurrence (LCWC) and Latent Dirichlet Allocation (LDA)	Undefined	Retweets Hashtags Links to external news sources
Cataldi et al. (Cataldi et al., 2010)	Topic graph model	Term vector	Time Username Retweets
Chen et al. (Chen et al., 2013)	Classical cluster algorithm -Affinity Propagation(AP)-	Keyword vector	___
Lee et al. (Lee et al., 2011)	Different classifiers such as Naïve Bayes Multinomial, SVM, ZeroR baseline, C5.0 decision tree, KNN, logistic Regression	Bag of Words	Time Friend-follower relationship
Tembhurnikar and Patil (Tembhurnikar & Patil, 2015)	Clustering (based on n-grams and named entity boosting)	Raw data	URLs Time
Ishikawa et al. (Ishikawa et al., 2012)	Incremental clustering and burst detection	Set of words	URLs Location information
Kim et al. (Kim et al., 2013)	Geographic Clustering	– -	Time Location information
Rill et al. (Rill et al., 2014)	Concept level sentiment analysis based where web ontology and semantic networks are used as knowledge base	– -	Hashtags Users' language location
Benny and Philip (Benny & Philip, 2015)	Clustering using AGF	– -	– –
Xie et al. (Xie et al., 2016)	Topic modelling based	Word vector	Time
Yin et al. (Yin et al., 2013)	Topic modelling based	bag of words	Time
Nur'aini et al. (Nur'Aini et al., 2015)	Clustering and Singular Value Decomposition (SVD)	SVD	– –
Cigarrán et al. (Cigarrán et al., 2016)	FCA-Based Method	Continuous stream of data	Time Hashtags
Geng et al. (Geng et al., 2017)	Hybrid clustering (based on K-means and AGNES)	– –	– –
Hasan et al. (Hasan et al., 2019)	incremental clustering	– –	Time
Yan et al. (D. Yan et al., 2016)	Single pass clustering algorithm	word embedding	Time
Kumar and Sridhar (Sridhar, 2015)	soft clustering over distributed representations of words	Distributed Representations of Words	Language Usernames Hashtags web addresses
Nguyen et al. (Nguyen et al., 2015)	Probabilistic topic model	word embedding	– -
Hu et al. (Hu et al., 2018)	CNN and SVM classifier	pre-trained word embedding	User's Sentiment
Zhao et al. (Zhao et al., 2018)	CNN	Word vectors	– -
Ma (Ma, 2018)	BERT-based Classifier	Contextual word embedding	Language
Gencoglu (Gencoglu, 2018)	K-means, ward and spectral clustering	Contextual word embedding	Language Time URL Hashtag

are matrix factorization methods. In this regard, Non-Negative Matrix Factorization (NMF) has been applied to the Topic Detection task (Arora et al., 2012) where a document-term matrix is approximately factorized into term-feature and feature-document matrices. NMF is a method of topic

Table 3. Advantages and disadvantages of detection techniques

Detection techniques	Advantages	Disadvantages
Classification Based	• Their topic representation is accurate. • They have a high performance in classifying content seen in the training set.	• They are supervised and the algorithms need to be trained in order to classify new inputs and they limit their implementation to these training data. • They will not be correctly classified the new topics including new features which unseen in the training set. • They limit the number of topics to a series of predefined classes (e.g., sports, politics, etc.) which have to be known in advance. • They propose flat topic representations that do not capture the topic-inherent hierarchy.
Clustering Based	• They are unsupervised and they do not need a training process, and the topic will not be restricted to the training data.	• They need to set a predefined number of clusters, so they limit the number of topics to be detected. • They have to be parameterized and they are highly dependent on this parameterization, or require the application of supervised methods as a previous step. • Clustering-based topics are difficult to describe, and then they need clustering labelling methods. • They propose flat topic representations that do not capture the topic-inherent hierarchy.
Topic Modelling Based	• They are unsupervised and they do not need a training process, and the topic will not be restricted to the training data.	• They limit the number of topics to be detected. • They have to be parameterized and they are highly dependent on this parameterization, or require the application of supervised methods as a previous step. • They tend to over-generalize (i.e., they generate quite generic topics). • There is trouble in dealing with topics that have complex generalization-specialization relationships. • They generate topics according to the relationships between terms in a latent semantic space and they are difficult to be transferred as topics in the real world. • They propose flat topic representations that do not capture the topic-inherent hierarchy.
FCA Based	• They are fully unsupervised, so they do not require any training process or any parameter to be implemented. • They behave well in the adaptation to new topics or the selection of new features. • They do not need to know a priori the number of topics. • The number of generated topics is dependent on the features of data, and then they will be adaptable to different data. • They labelled the topics automatically. • They do not need prior knowledge about the data. • The hierarchical construction of the concept lattice allows the fine- and coarse-grained representation of topics. • They use formal concepts as topics that make them clear interpretable.	• They have high computational complexity, especially when the number of attributes is very large.

modelling on classic literature, and we put matrix factorization methods in the topic modelling category. Finally, FCA-based approaches are mathematical frameworks that model the contents (tweets) according to their attributes (terminology) in a lattice structure with no prior knowledge about the data. The FCA does not need to set the number of topics since it depends only on the data features and it does not limit the number of generated topics. The disadvantage of FCA is the high number of concepts (possible topics) has been generated and the algorithm computation time has been increased. We summarize the advantages and disadvantages of detection techniques in Table 3.

b) **Representation**: it is a format of a tweet's transformation that is easier to realise.

c) **Twitter-Specific Features**: Content and Context are the two tweets' features (Figure 3). Textual content is 140 unstructured that consists of terms and sentiments. The Twitter syntaxes that we discussed in section 3.2 have been created for spreading information between users and promoting tweet searchability (Liao et al., 2012). These Twitter-specific features can be used to infer further insights to improve content understanding and assess tweet quality (Naveed et al., 2011).

There is additional information in a tweet's metadata (username, following and followers, user verification and user location) for several purposes. Username can be utilized to eliminate topics that are dominated by single user, which are prone to be spam. The following and followers relationship can be used to verify user fame, and estimate the information diffusion. User verification can be used to filter and rank tweets (Uysal & Croft, 2011), and user location can show interesting topics of a specific geographic (Sankaranarayanan et al., 2009).

Performance evaluation of different approaches is the main issue facing topic detection in Twitter. Table 4 lists existing works on topic detection approaches for Twitter and summarizes what methods were used to evaluate them. A detailed description of the columns' content is presented at first:

(a) **Measures**: The Measures column summarizes the various measures that were proposed to evaluate the different approaches. Precision, recall, and F-scores are common performance metrics. Precision is the fraction of retrieved tweets that are relevant to the topic; recall is the fraction of relevant documents that have been retrieved. F-scores are weighted harmonic means of precision and recall. Apart from these well-known measures, some novel measures were defined such as missing rate, NMI (Normalized Mutual Information), NPMI (Normalized Pointwise Mutual Information) scores, etc. Precision and recall can be calculated using the following formulas (Equations 1 and 2).

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

TP (True positive) indicates the number of tweets the system correctly identifies as relevant; FP (False Positive) is the number of tweets the system identifies as relevant. FN is the number of relevant tweets the system fails to identify and TN (True Negative) is the number of tweets that the system correctly identifies as irrelevant. Recall is difficult to compute in large and noisy data sets, since manual enumeration of all relevant topics that exist in a given Twitter stream is time-consuming for small sets and infeasible for larger ones. Therefore, as illustrated in Table 4, a few of the work surveyed in this article focused on recall measure.

(b) **Tweets**: The Tweets column lists the number of tweets used in the evaluation.
(c) **Temporal Scope**: In this column, we show the period of time that a collection of Twitter data gathered in each method.
(d) **Data Collection**: In the data collection column, we list the different applications that are used to collect the data for the evaluation. As illustrated in Table 4, most of the works are based on the Streaming API (Application Programming Interface). Twitter streaming API provides filtering by keywords, location, author, and others. Streaming API has different levels or restrictions, such as Filter API, Spritzer access level, and Gardenhose level. With Filter API, it is possible to obtain the data in a streaming fashion and to pre-define filter queries based on geographical locations or keywords. This API is the most popular choice. The Spritzer access level provides a uniform random 1% stream of the public timeline and is freely available to everyone. The Gardenhose level provides elevated access to a 10% stream but needs some special authorisation.

Table 4. Overview of topic detection methods' datasets and evaluation metrics.

References	Measures	Tweets	Temporal scope	Data collection
Sankaranarayanan et al. (Sankaranarayanan et al., 2009)	Qualitative	– –	– –	GardenHose BirdDog
Phuvipadawat and Murata (Phuvipadawat & Murata, 2010)	Qualitative	154,000	– –	Streaming API (selected users who use #breakingnews in their messages.)
Petrović et al. (Petrović et al., 2010)	Average precision	163.5 m	6 months	Spritzer
Sriram et al. (Sriram et al., 2010)	Accuracy	5407	– –	streaming API
Rakesh et al. (Rakesh et al., 2013)	Precision	10,000	– –	streaming API (tweets that pertain to US) Filter
Cataldi et al. (Cataldi et al., 2010)	Qualitative	3 m	15 days	Spritzer
Chen et al. (Chen et al., 2013)	event posts distribution	22,360,000	1 month	Sina-Weibo (Twitter-like microblogging system in China provided by Sina)
Lee et al. (Lee et al., 2011)	Accuracy	– -	– -	Twitter API
Tembhurnikar and Patil (Tembhurnikar & Patil, 2015)	Qualitative	– -	3 days	3 big dataset (cover the domain of sport, disease and bill) collected from public Twitter API
Ishikawa et al. (Ishikawa et al., 2012)	– -	– -	9 days	API follows a Zipfian distribution
Kim et al. (Kim et al., 2013)	Ratio of word frequency	18,720,902	9 days	Twitter Stream API(geo-tagged public statuses tweeted in North America regions) Filter
Rill et al. (Rill et al., 2014)	Correlation Probability	4 m	5 months	streaming API (predefined search terms targeting political tweets) Filter
Benny and Philip (Benny & Philip, 2015)	Purity cluster entropy class entropy	530	– –	Twitter API (tweets contain the keyword "India") Filter
Xie et al. (Xie et al., 2016)	Throughput Inference Time Precision Recall KL-divergence	30 m	– –	two different Twitter data sets (tweets from Singapore and San Francisco)
Yin et al. (Yin et al., 2013)	Accuracy Normalized Frequency	9,884,640	6 months	Two real-life data sets from Twitter and Del.icio.us.
Nur'aini et al. (Nur'Aini et al., 2015)	Accuracy Topic Recall Keyword Precision Keyword recall computing time	– –	– –	three Twitter datasets focused on three popular real-world events, The US Super Tuesday, The US Presidential Elections, and The English FA Cup
Cigarrán et al. (Cigarrán et al., 2016)	Reliability Sensitivity F-Measure	134,200	7 months	Twitter API
Geng et al. (Geng et al., 2017)	F1 Accuracy Recall Missing rate False detection rate Model index norm (Cadet) (J. Allan et al., 2005) Time	12,000	– –	Sina-Weibo

(*Continued*)

Table 4. (Continued).

References	Measures	Tweets	Temporal scope	Data collection
Hasan et al. (Hasan et al., 2019)	Precision Recall Number of detected ground truth events	17,000,000	3 days	Events2012 corpus (McMinn et al., 2013)
Yan et al. (D. Yan et al., 2016)	Missing rate False detection rate Error identification cost Time cost	2 million microblog	4 months	Sina-Weibo
Kumar and Sridhar (Sridhar, 2015)	Topic Coherence Score	305,989,257 senteces	2 weeks	10% random sample of Twitter firehose data across all languages
Nguyen et al. (Nguyen et al., 2015)	NPMI scores Qualitative Purity and NMI F1 scores	2,520	____	TagMyNews news dataset and the Sanders Twitter corpus
Hu et al. (Hu et al., 2018)	Precision Recall F1-measure Accuracy	3,169	____	____
Zhao et al. (Zhao et al., 2018)	Accuracy	10,539	____	____
Ma (Ma, 2018)	Accuracy Matthews Correlation Coefficient Precision Recall F1-Score	74,346	____	CrisisLexT26 (Olteanu et al., 2014) CrisisNLP (Imran et al., 2013, Imran et al., 2016, Alam et al., 2018)
Gencoglu (Gencoglu, 2018)	Calinski-Harabasz (Caliński & Harabasz, 1974)	63,326	Approximately 4 years	Twitter API

As shown in Table 4, the approaches have been evaluated on different data sets, and we cannot compare them in that condition. Hence, Representative Data sets and publicly available testbeds are highly required for the comprehensive evaluation of performance and objective comparison of different detection approaches that are proposed for topic detection from Twitter. Therefore, sharing (and if possible merging) the labeled data sets online (such as those presented in Table 4), as well as using crowd sourcing services for larger scale labeling, would provide more representative data for evaluation.

We achieved the following results by analysing the tables:

- In the context of topic detection, most of the works are unsupervised and rely on clustering.
- Most unspecified approaches rely on clustering approaches.
- The majority of the NED and unspecified methods are unsupervised clustering of new tweets.
- The NED and specified techniques mainly rely on supervised learning approaches. Although manually labelling a large number of tweets is a time-consuming and labor-intensive task, it is more feasible for specified techniques than for unspecified. In specified methods, filtering according to specific information, such as location, time, or, keywords would reduce the number of Tweets that must be processed and allow the detection algorithm to focus on a limited set of Tweets.
- Clustering detection technique is suitable for unspecified and NED approaches.
- Clustering techniques follow an unsupervised methodology. Thus, they do not require labelled data for training, so detecting the topic will not be restricted to the training data.

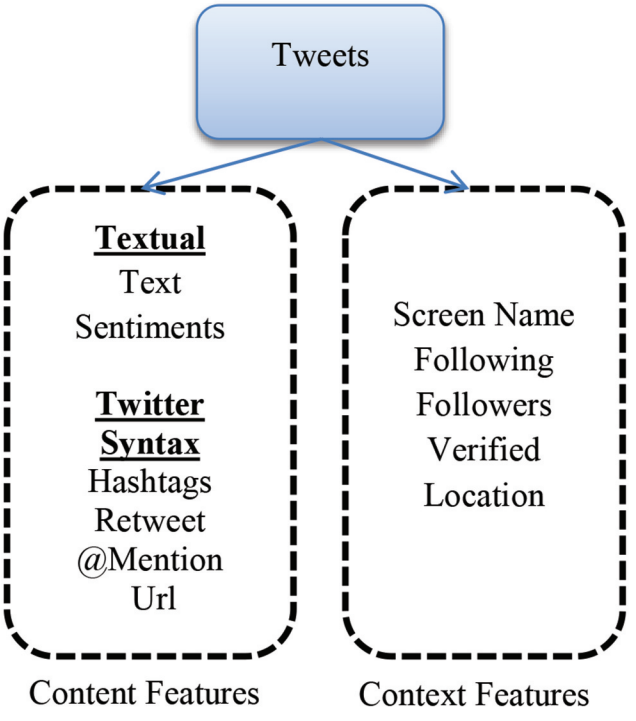

Figure 3. Type of features in a tweet.

Table 5. Characteristic of each category of topic detection approaches.

Category		Characteristic
Short text representation	**Without Word Embedding**	• representing words as discrete and distinct symbols. • hardly capture the semantics of words or the distances between words.
	With Word Embedding	• Words can be mapped to vectors in low dimensional space and so they are simple to work with. • Semantically similar words would be similar to each other by calculating cosine similarity among the vectors of words.
Type of events	**Specified**	• They rely on specific information such as location, time that is provided by users.
	Unspecified	• They have not prior information.
Detection method	**Supervised**	• They need labelled training data. • They usually assume a static environment.
	Unsupervised	• there are no training documents or queries and they do not require labelled data.
Detection task	**NED**	• They deal with documents processed online or in other words before looking at any subsequent documents. • The data as it arrives from the source are analysed so they are able to discover topics from real-time streams.
	RED	• They focused on discovering previously unidentified topics from the accumulated historical collections. • All the data are collected and then all the collected data are analysed entirely at once. • You have definite liberties in how to perform the analysis due to the availability of all data at the start. Because the dataset size is known, you are able to tweak size-dependent parameters in your analysis tool to better fit the data set. • As you need to gather all the data you wish to process in advance, it is unsuited for processing streams of data arriving in real-time.

- Most topic modelling-based methods fall into RED category. Detecting the topics just in time as they are taking place is challenging due to the high computational complexity inherent in the topic models.
- The major clustering-based methods are NED approaches.
- The FCA-based methods are fully unsupervised methods; this aspect is significant in TDT tasks, where the development of training datasets is expensive and difficult.
- Most of the systems that have top performance, apply unsupervised methodologies.
- Most approaches concentrated on English language.
- Most of the approaches work on content features, context features are less used.
- Most of the topic detection approaches have traditional word representation and a few methods use word embeddings.

At last, we can summarize the characteristic of each category of topic detection approaches in Table 5.

Conclusion and future work

In recent years, the popularity of microblogs such as Twitter is growing unprecedentedly. Many numbers of users used Twitter and exchanged and told their last thoughts, moods, or activities by tweets in some words. Topic detection is the solution for monitoring and summarizing information generates from Twitter. The topic detection methods for traditional Medias do not suit Twitter due to the nature of Twitter. So, in this study, firstly, we discuss the characteristics of Twitter. Next, we review the topic detection approaches from Twitter. These methods are classified to with word embedding or without word embedding, specified or unspecified, offline or online and supervised or unsupervised. Thirdly, we categorised the detection techniques and summarized their advantages and disadvantages. Lastly, we discussed the categories of methods, extracted characteristics of each category, and their limitations identified. A more in-depth discussion appears in section 6. Generally, depending on the application, one of the categories may be more suitable than the other. In terms of perspectives, we will try to take advantage of the approaches in each category studied in this paper, to propose a new method for detecting topics in twitter in real-world. Finally, this article highlights the need for publicly available testbeds for inclusive evaluation of performance and objective comparison of different detection approaches.

Notes

1. http://www.twitter.com/
2. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
3. http://blogs.cornell.edu/newmediaandsociety2010/2010/02/24/twitter-i-dont-care-about-yourdaily-minutiae/
4. http://blog.twitter.com/2009/11/whats-happening
5. https://support.twitter.com/articles/49,309
6. Defence Advanced Research Projects Agency
7. C implementation of variational expectation maximization for latent Dirichlet allocation (LDA)., from http://www.cs.princeton.edu/~blei/lda-c/index.html.
8. http://lucene.apache.org/core/
9. Global vectors for word representation
10. Embeddings from Language Models
11. Bidirectional Encoder Representations from Transformers
12. Dirichlet Multinomial Mixture
13. https://code.google.com/p/word2vec/
14. http://www-nlp.stanford.edu/projects/glove/

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Aggarwal, C. C., & Subbian, K. (2012). Event detection in social streams. *Proceedings of the 2012 SIAM international conference on data mining*, SIAM. Anaheim, California, USA.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66. https://doi.org/10.1007/BF00153759

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA.

Alam, F., Ofli, F., Imran, M., & Aupetit, M. (2018). A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. *arXiv Preprint*. https://arxiv.org/abs/1805.05144

Alhawarat, M., & Hegazi, M. (2018). Revisiting K-means and topic modeling, a comparison study to cluster arabic documents. *IEEE Access*, *6*(1), 42740–42749. https://doi.org/10.1109/ACCESS.2018.2852648

Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., & Yang, Y. (2003). Topic detection and tracking pilot study final report. Carnegie Mellon University. https://doi.org/10.1184/R1/6626252.v1

Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., & Amstutz, P. (2005). Taking topic detection from evaluation to practice. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, : Big Island, Hawaii, USA: IEEE.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 37–45). Citeseer.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., & Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, *15*(1), 2773–2832. https://escholarship.org/uc/item/87b1x7b6

Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models–going beyond SVD. *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, IEEE, New Brunswick, NJ, USA.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* Toronto, Ontario, Canada: IEEE Computer Society.

Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and issues in data stream systems. *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. Madison, Wisconsin, USA: ACM.

Bahrehdar, A. R., & Purves, R. S. (2018). Description and characterization of place properties using topic modeling on georeferenced tags. *Geo-spatial Information Science*, *21*(3), 173–184. https://doi.org/10.1080/10095020.2018.1493238

Becker, H., Naaman, M., & Gravano, L. (2010). Learning similarity metrics for event identification in social media. *Proceedings of the third ACM international conference on Web search and data mining*. New York, USA: ACM.

Benny, A., & Philip, M. (2015). Keyword based tweet extraction and detection of related topics. *Procedia Computer Science*, *46*(4), 364–371. https://doi.org/10.1016/j.procs.2015.02.032

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993

Bouras, C., & Tsogkas, V. (2010). Assigning web news to clusters. *2010 Fifth International Conference on Internet and Web Applications and Services*. Barcelona, Spain: IEEE.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. *2010 43rd Hawaii International Conference on System Sciences*. Koloa, Kauai, Hawaii: IEEE.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Carter, S., Weerkamp, W., & Tsagkias, M. (2013). Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, *47*(1), 195–215. https://doi.org/10.1007/s10579-012-9195-y

Castellanos, A., Cigarrán, J., & García-Serrano, A. (2017). Formal concept analysis for topic detection: A clustering quality experimental analysis. *Information Systems*, *66*(4), 24–42. https://doi.org/10.1016/j.is.2017.01.008

Cataldi, M., Di Caro, L., & Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. *Proceedings of the tenth international workshop on multimedia data mining*. Washington, DC, USA: ACM.

Chemudugunta, C., Smyth, P., & Steyvers, M. (2007). *Modeling general and specific aspects of documents with a probabilistic topic model*. Advances in neural information processing systems.

Chen, Y., Xu, B., Hao, H., Zhou, S., & Cao, J. (2013). User-defined hot topic detection in microblogging. *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*. Huangshan, China: ACM.

Cigarrán, J., Castellanos, Á., & García-Serrano, A. (2016). A step forward for topic detection in twitter: An FCA-based approach. *Expert Systems with Applications*, *57*(15), 21–36. https://doi.org/10.1016/j.eswa.2016.03.011

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*. Helsinki, Finland: ACM.
Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. Copenhagen, Denmark: ACM.
Dai, X.-Y., Chen, Q.-C., Wang, X.-L., & Xu, J. (2010). Online topic detection and tracking of financial news based on hierarchical clustering. *2010 International Conference on Machine Learning and Cybernetics*. Qingdao, China: IEEE.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805
Dubes, R. C., & Jain, A. K. (1988). *Algorithms for clustering data*. Prentice hall Englewood Cliffs.
Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2–3), 195–225. https://doi.org/10.1007/BF00114844
Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd annual meeting on association for computational linguistics*, Association for Computational Linguistics, Ann Arbor, Michigan.
Fiscus, J. G., & Doddington, G. R. (2002). Topic detection and tracking evaluation overview. In J. Allan (Ed.), *Topic detection and tracking* (pp. 17–31). Springer.
Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *science*, *315*(5814), 972–976. https://doi.org/10.1126/science.1136800
Fung, G. P. C., Yu, J. X., Yu, P. S., & Lu, H. (2005). Parameter free bursty events detection in text streams. *Proceedings of the 31st international conference on Very large data bases*, VLDB Endowment, Trondheim, Norway.
Ganter, B., & Wille, R. (2012). *Formal concept analysis: Mathematical foundations*. Springer Science & Business Media.
Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. *arXiv Preprint*. https://arxiv.org/abs/1803.07640v2
Gencoglu, O. (2018). Deep representation learning for clustering of health tweets. https://arxiv.org/abs/1901.00439
Geng, X., Zhang, Y., Jiao, Y., & Mei, Y. (2017). A novel hybrid clustering algorithm for microblog topic detection. *AIP Conference Proceedings*. Hubei Province, China: AIP Publishing.
Givoni, I., Chung, C., & Frey, B. J. (2012). Hierarchical affinity propagation. https://arxiv.org/abs/1202.3722
Guo, X., Xiang, Y., Chen, Q., Huang, Z., & Hao, Y. (2013). LDA-based online topic detection using tensor factorization. *Journal of Information Science*, *39*(4), 459–469. https://doi.org/10.1177/0165551512473066
Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520
Hasan, M., Orgun, M. A., & Schwitter, R. (2016). TwitterNews±: A framework for real time event detection from the Twitter data stream. *International Conference on Social Informatics*. Bellevue, USA: Springer.
Hasan, M., Orgun, M. A., & Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management*, *56*(3), 1146–1165. https://doi.org/10.1016/j.ipm.2018.03.001
He, Q., Chang, K., & Lim, E.-P. (2007). Analyzing feature trajectories for event detection. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. Amsterdam, Netherlands: ACM.
Hooper, R., & Paice, C. (2005). *The Lancaster stemming algorithm*. University of Lancester.
Hu, Y., John, A., Seligmann, D. D., & Wang, F. (2012). What were the tweets about? Topical associations between public events and twitter feeds. *Sixth International AAAI Conference on Weblogs and Social Media*. Dublin, Ireland
Hu, Y., Yi, Y., Yang, T., & Pan, Q. (2018). Short text classification with a convolutional neural networks based method. *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Marina Bay Sands Expo and Convention Centre, Singapore, IEEE.
Humphreys, L., Gill, P., Krishnamurthy, B., & Newbury, E. (2013). Historicizing new media: A content analysis of Twitter. *Journal of Communication*, *63*(3), 413–431. https://doi.org/10.1111/jcom.12030
Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). *Extracting information nuggets from disaster-related messages in social media*. Iscram.
Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. https://arxiv.org/abs/1605.05894
Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. Dallas, Texas, USA: ACM.
Ishikawa, S., Arakawa, Y., Tagashira, S., & Fukuda, A. (2012). Hot topic detection in local areas using Twitter and Wikipedia. *ARCS 2012*. Muenchen, Germany: IEEE.
Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. Columbus, Ohio.

Kim, H.-G., Lee, S., & Kyeong, S. (2013). Discovering hot topics using Twitter streaming data social topic detection and geographic clustering. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. Niagara Ontario, Canada: IEEE.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*(4), 373–397. https://doi.org/10.1023/A:1024940629314

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of Finnish text documents. *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. Washington D.C., USA: ACM.

Kuznetsov, S. O. (2007). On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, *49*(1–4), 101–115. https://doi.org/10.1007/s10472-007-9053-6

Kuzuetsov, S. (1990). Stability as an estimate of thie degree of substantiation of hypotheses derived on the basis of operational, *Nauchno-Tekhnicheskaya Informatslya, Seriya 2, 24*(12), 21–29. https://www.researchgate.net/publication/265673866_Stability_as_an_estimate_of_the_degree_of_substantiation_of_hypotheses_derived_on_the_basis_of_operational_similarity

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*. Raleigh, North Carolina, USA: AcM.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, *41*(1), 191–201. https://www.jstor.org/stable/2347628

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011). Twitter trending topic classification. *2011 IEEE 11th International Conference on Data Mining Workshops*. Vancouver, Canada: IEEE.

Li, H.-F., Lee, S.-Y., & Shan, M.-K. (2004). An efficient algorithm for mining frequent itemsets over the entire history of data streams. *Proc. of First International Workshop on Knowledge Discovery in Data Streams*. Pisa, Italy

Li, S., Lv, X., Wang, T., & Shi, S. (2010). The key technology of topic detection based on K-means. *2010 International Conference on Future Information Technology and Management Engineering*. Changzhou, China: IEEE.

Liao, Y., Moshtaghi, M., Han, B., Karunasekera, S., Kotagiri, R., Baldwin, T., Harwood, A., & Pattison, P. (2012). Mining microblogs: Opportunities and challenges. In Abraham A. (Ed.), *Computational social networks* (pp. 129–159). Springer.

Liu, Y., Liu, Z., Chua, T.-S., & Sun, M. (2015). Topical word embeddings. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, Texas, USA.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanics Translations & Computer. Linguistics*, *11*(1–2), 22–31. http://chuvyr.ru/MT-1968-Lovins.pdf

Ma, G. (2018). *Tweets classification with BERT in the field of disaster management*. Stanford, CA 94305, Department of Civil Engineering Stanford University.

Marwick, B. (2013). Discovery of emergent issues and controversies in anthropology using text mining, topic modeling, and social network analysis of microblog content. In Y. Zhao & Y. Cen (Eds.), *Data Mining Applications with R* (pp. 63–93). Elsevier.

McCallum, A. K. (2002). "*Mallet: A machine learning for language toolkit*." http://mallet.cs.umass.edu

McMinn, A. J., Moshfeghi, Y., & Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. San Francisco, California, USA: ACM.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. Advances in neural information processing systems.

Milstein, S., Lorica, B., Magoulas, R., Hochmuth, G., Chowdhury, A., & O'Reilly, T. (2009). *Twitter and the micro-messaging revolution: Communication, connections, and immediacy–140 characters at a time*. O'Reilly Media.

Mori, M., Miura, T., & Shioya, I. (2004). Extracting events from web pages. *AISTA'04: Proceedings of the International Conference on Advances in Intelligent Systems-Theory and Applications (AISTA)*. Centre de Recherche Public Henri Tudor, Luxembourg-Kirchberg, Luxembourg

Mori, M., Miura, T., & Shioya, I. (2006). Topic detection and tracking for news web pages. *Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence*, IEEE Computer Society.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, *26*(4), 354–359. https://doi.org/10.1093/comjnl/26.4.354

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Searching microblogs: Coping with sparsity and document quality. *Proceedings of the 20th ACM international conference on Information and knowledge management*. Hong Kong: ACM.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, *3*(1), 299–313. https://doi.org/10.1162/tacl_a_00140

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*(2–3), 103–134. https://doi.org/10.1023/A:1007692713085

Nur'Aini, K., Najahaty, I., Hidayati, L., Murfi, H., & Nurrohmah, S. (2015). Combination of singular value decomposition and K-means clustering methods for topic detection on Twitter. *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. Depok, Indonesia: IEEE.

O'Connor, B., Krieger, M., & Ahn, D. (2010). Tweetmotif: Exploratory search and topic summarization for twitter. *Fourth International AAAI Conference on Weblogs and Social Media*, George Washington University in Washington, DC, USA.

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. *Eighth International AAAI Conference on Weblogs and Social Media,* Ann Arbor, Michigan, USA

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. https://arxiv.org/abs/1802.05365.

Petrović, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. Human language technologies: The 2010 *Annual conference of the north american chapter of the association for computational linguistics*, Association for Computational Linguistics, Los Angeles, California, USA.

Phuvipadawat, S., & Murata, T. (2010). Breaking news detection and tracking in Twitter. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Los Alamitos, CA, US: IEEE.

Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Boston, USA: IEEE.

Piczak, K. J. (2019). *Sound classification with convolutional neural networks*. The Institute of Computer Science.

Porter, M. F. (2006). An algorithm for suffix stripping. *Program*, *40*(3), 211–218. https://doi.org/10.1108/00330330610681286

Prabowo, R., Thelwall, M., Hellsten, I., & Scharnhorst, A. (2008). Evolving debates in online communication: A graph analytical approach. *Internet Research*, *18*(5), 520–540. https://doi.org/10.1108/10662240810912765

Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research*, *4*(1), 77–90. https://doi.org/10.1613/jair.279

Rakesh, V., Reddy, C. K., & Singh, D. (2013). Location-specific tweet detection and topic summarization in twitter. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Niagara Ontario, Canada: ACM.

Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V. (2014). PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, *69*(15), 24–33. https://doi.org/10.1016/j.knosys.2014.05.008

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, *5*(3), 1. https://doi.org/10.1038/323533a0

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web*. Raleigh, North Carolina, USA: ACM.

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. https://doi.org/10.1109/LSP.2017.2657381

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. mcgraw-hill.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009). Twitterstand: News in tweets. *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*. Seattle, USA: ACM.

Sayyadi, H., & Raschid, L. (2013). A graph analytical approach for topic detection. *ACM Transactions on Internet Technology (TOIT)*, *13*(2), 4. https://doi.org/10.1145/2542214.2542215

Schubert, E., Weiler, M., & Kriegel, H.-P. (2014). Signitrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, USA: ACM.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. *Computer Speech & Language*, *15*(3), 287–333. https://doi.org/10.1006/csla.2001.0169

Sridhar, V. K. R. (2015). Unsupervised topic modeling for short texts using distributed representations of words. *Proceedings of the 1st workshop on vector space modeling for natural language processing,* Denver, Colorado, USA

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. Geneva, Switzerland: ACM.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*, *427*(7), 424–440. http://173.236.226.255/tom/papers/SteyversGriffiths.pdf

Suzuki, J., & Nagata, M. (2015). A unified learning framework of skip-grams and global vectors. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China.

Teevan, J., Ramage, D., & Morris, M. R. (2011). # TwitterSearch: A comparison of microblog search and web search. *Proceedings of the fourth ACM international conference on Web search and data mining*. Hong Kong: ACM.

Tembhurnikar, S. D., & Patil, N. N. (2015). Topic detection using BNgram method and sentiment analysis on twitter dataset. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)*. Noida, India: IEEE.

Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. *Proceedings of the 17th international conference on World Wide Web*. Beijing, China: ACM.

Toda, H., & Kataoka, R. (2005). A search result clustering method using informatively named entities. *Proceedings of the 7th annual ACM international workshop on Web information and data management*. Bremen, Germany: ACM.

Uysal, I., & Croft, W. B. (2011). User oriented tweet ranking: A filtering approach to microblogs. *Proceedings of the 20th ACM international conference on Information and knowledge management*. Glasgow Scotland, UK: ACM.

Wang, F., Orton, K., Wagenseller, P., & Xu, K. (2018). Towards understanding community interests with topic modeling. *IEEE Access*, *6*(1), 24660–24668. https://doi.org/10.1109/ACCESS.2018.2815904

Wang, L., Qiao, Y., Li, W., Xu, C., & Tang, X. (2018). *Video classification method and apparatus, US Patent App*. 15/167,388. Google Patents

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. *COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics*, Nantes, France..

Wille, R. (1992). Concept lattices and conceptual knowledge systems. *Computers & Mathematics with Applications*, *23* (6–9), 493–515. https://doi.org/10.1016/0898-1221(92)90120-7

Wille, R. (2009). Restructuring lattice theory: An approach based on hierarchies of concepts. *International Conference on Formal Concept Analysis*. Darmstadt, Germany: Springer.

Winarko, E., & Pulungan, R. (2019). Trending topics detection of Indonesian tweets using BN-grams and Doc-p. *Journal of King Saud University-Computer and Information Sciences*, *31*(2), 266–274. https://doi.org/10.1016/j.jksuci.2018.01.005

Xie, W., Zhu, F., Jiang, J., Lim, E.-P., & Wang, K. (2016). Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2216–2229. https://doi.org/10.1109/TKDE.2016.2556661

Xu, G., Yu, Z., & Qi, Q. (2018). Efficient sensitive information classification and topic tracking based on tibetan Web pages. *IEEE Access*, *6*(1), 55643–55652. https://doi.org/10.1109/ACCESS.2018.2870122

Yan, D., Hua, E., & Hu, B. (2016). An improved single-pass algorithm for Chinese microblog topic detection and tracking. *2016 IEEE International Congress on Big Data (BigData Congress)*. San Francisco, CA, USA: IEEE.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the 22nd international conference on World Wide Web*. Rio de Janeiro, Brazil: ACM.

Yang, Y., Pierce, T., & Carbonell, J. G. (1982). A study on retrospective and on-line event detection. *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information*, August 1998, 28–36. https://doi.org/10.1145/290941.290953

Yin, H., Cui, B., Lu, H., Huang, Y., & Yao, J. (2013). A unified model for stable and temporal topic detection from social media data. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. Brisbane, QLD, Australia: IEEE.

Yu, Y., Li, M., & Fu, Y. (2018). Forest type identification by random forest classification combined with SPOT and multitemporal SAR data. *Journal of Forestry Research*, *29*(5), 1407–1414. https://doi.org/10.1007/s11676-017-0530-4

Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., & Atkinson, P. M. (2018). A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *140*(6), 133–144. https://doi.org/10.1016/j.isprsjprs.2017.07.014

Zhang, C., Wang, H., Cao, L., Wang, W., & Xu, F. (2016). A hybrid term–term relations analysis approach for topic detection. *Knowledge-Based Systems*, *93*(3), 109–120. https://doi.org/10.1016/j.knosys.2015.11.006

Zhang, D., & Li, S. (2011). Topic detection based on K-means. *2011 International Conference on Electronics, Communications and Control (ICECC)*. Ningbo, China: IEEE.

Zhang, M., Li, W., & Du, Q. (2018). Diverse region-based CNN for hyperspectral image classification. *IEEE Transactions on Image Processing*, *27*(6), 2623–2634. https://doi.org/10.1109/TIP.2018.2809606

Zhao, D., Chang, Z., Du, N., & Guo, S. (2018). Classification for social media short text based on word distributed representation. *International Conference on Web Information Systems and Applications*. Taiyuan, China: Springer.

Zhou, P., Han, J., Cheng, G., & Zhang, B. (2019). Learning compact and discriminative stacked autoencoder for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 57*(7), 4823–4833. doi: 10.1109/TGRS.2019.2893180.