Cross-platform comparison of framed topics in Twitter and Weibo: machine learning approaches to social media text mining

Yi Yang¹ · Jia-Huey Hsu² · Karl Löfgren³ · Wonhyuk Cho³

Received: 22 October 2020 / Revised: 13 July 2021 / Accepted: 13 July 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2021

Abstract

While the salience of social media platforms on modern interactive communication between diverse social actors has been demonstrated, less academic attention has been paid to comparisons between framed topics and user interactions across social media platforms, such as Twitter and Weibo. This article suggests text mining and natural language processing tools for cross-platform comparative social media studies, based on Latent Dirichlet Allocation (LDA) and social network analysis. This study illustrates how the suggested topic models and data processing algorithms can be applied to a real-life example (U.S.-China trade war discourse on social media), and experimented the methods on social media text mining data, revealing differences between user interactions on Twitter, predominantly “Western,” and Weibo, largely representing Chinese-speaking users. We discuss the strengths and weaknesses of the suggested machine learning algorithms for comparative social media studies.

Keywords Social media · Latent Dirichlet allocation · Text mining · Machine learning · Algorithm · Twitter · Weibo · Social network analysis

1 Introduction

In the past decade, the salience of social media platforms such as Twitter and Weibo as interactive communication tools has grown (Bird et al. 2011). However, while social media reduces the transaction costs of information disclosure and user interactions (McDermott 2010; Mergel 2012), its “filter bubbles” uphold an old axiom of media and communication studies: individuals who only interact with other like-minded users on social media may just perpetuate previously held beliefs (Pariser 2011). This hurdle calls for the understanding of social structures and interactions across various platforms. In addition, the nature of social media data, which stores rich text data about user interactions, can help us understand the aggregated expressions of users

about a topic or event, enabling the systematic analysis of the structure or patterns of activity between individuals or groups.

This article proposes a set of methods for cross-platform comparisons of framed topics and user engagement across two social media ecosystems: Twitter and Weibo. We selected these two platforms as the most widely used in both the English-speaking (particularly in the U.S., in which 48.35 million monthly active Twitter users represent around 22% of adults), and the Chinese-speaking world (mainland China has 290 million registered users of *Sina Weibo*, around one-fifth of the population). These two platforms also represent social media use in liberal democracies (Twitter) and single-party systems (Weibo).

The differences in user interactions and framed topics between Twitter and Weibo are still uncharted territory. Though this lack of cross-platform analysis may be due to the lack of an international incident large enough to attract enough users of both platforms, it may also follow from an absence of general analytical guidelines for applying methodological tools to social media data.

For example, the U.S-China trade war, a relevant issue in both countries, is reasonably well-reflected in both social media platforms. This conflict itself may be of great interest

✉ Wonhyuk Cho
wonhyuk.cho@vuw.ac.nz

¹ College of Public Administration, Huazhong University of Science and Technology, Wuhan, Hubei, China

² Department of International Business, Tamkang University, Taipei, Taiwan

³ Wellington School of Business and Government, Victoria University of Wellington, Wellington, New Zealand

for various disciplines: the United States constitutes the world's largest economy (based on free-market capitalist ideology) while China, the world's second biggest economy, is based on a distinctive political system that mixes single-party state control with a market economy (Wu and Ramesh 2014). Culturally, the United States represents Western individualism while China is claimed to be an exemplar of Eastern collective values (Warden et al. 2005). Social media studies of non-Western countries, and comparative studies, have been underrepresented in the existing literature. This article aims to provide methodological tools with which future studies may fill this gap in research.

This work suggests analytical procedures for a comparative analysis of user interactions and framed topics between Twitter and Weibo platforms, subjects largely missing from previous social media studies. We propose using Latent Dirichlet Allocation (hereafter LDA) topic models to analyze latent topics in tweets and posts, respectively, in a way that future studies can replicate in other social media contexts. We also propose social network analysis methods to identify opinion leaders (network influencers) on the topics of interest in tweets/posts and construct a whole-network mapping of these accounts. Little social media research has applied these analytic techniques for cross-platform comparisons.

2 Why is it important to understand social media text?

Unlike the hierarchical one-to-many traditional media, social media has facilitated many-to-many interactions, in which everyone can become a content producer (Porter 2008). The nature of social media has attracted much scholarly attention for its potential to enhance our understanding of various social phenomenon, such as political engagement and the proliferation of opinions or ideas (Castells 2007; Tumasjan et al. 2011). Among different social media platforms, Twitter has been one of the most widely used tools in domestic and international discourse due to its open, horizontal, and networked architecture (Park 2013).

As such, academic interest in the nature of user interactions and framed topics on social media has grown (Wagner et al. 2017; Banks et al. 2020). For instance, Yang and Counts (2010) emphasized Twitter's critical role in information diffusion, and Pavel (2009) identified Twitter as a mobilization tool. Williamson and Phillips (2009) and Jungherr and Jürgens (2010) have studied the use of Twitter by political parties in the United Kingdom and Germany. Multiple scholars have noted and analyzed the role of Twitter in national elections (Tumasjan et al. 2011; Yaqub et al. 2017; Linvill et al. 2019; Grimaldi 2019).

In terms of academic research, however, very little is known about Chinese social media, even though the nation

boasts the highest number of Internet users, and whether observed "Western" social media phenomena can be replicated in the Chinese-speaking world. The largest Chinese social media platform (or microblog) is *Sina Weibo* (or Weibo, for short). In terms of usability, the platform shares several characteristics with Twitter: users post 140-character messages, address others with "@" ("at") symbols for "a mention," and use hashtags, and it allows reposts and post responses (Rauchfleisch and Schäfer 2015). The 486 million monthly active users of Weibo in 2019 represented an annual net increase of about 55 million.

China's virtual communication has rapidly developed against a unique institutional background and a separate social media platform ecosystem. While we may assume that different platform features can influence the nature of user behavior, and academic interest in the distinctive characteristics of Chinese social media have grown substantially (An et al. 2018; Kim et al. 2021), few studies have compared the characteristics of social media phenomena across cultures or political contexts (Chen et al. 2011).

3 Network structure and opinion leaders (network influencer) in social media

In the social media era, opinion leaders are often called influencers (Shmargad 2018). These actors, located in the nexus of information flows, build reputations for their expertise on a specific topic, generating large and heavily engaged followings and creating a paradigm known as influencer culture. Influencers use social support and social pressure to impact their personal networks (Katz and Lazarsfeld 1955; Dubois and Gaffney 2014). Their presence facilitates the process of opinion building at the periphery, as they strategically transfer information and impact to less active and central communicators in the network (Keller and Berry 2003; Karlsen 2015).

While the term "influencer" is often used to describe one who affects buying habits by encouraging followers to purchase their promoted products or services, the concept has recently expanded beyond its commercial meaning. For example, in political activities, influencer marketing may be adopted by celebrities and other politicized influencers, changing the landscape of political communication. To accommodate the wide range of different types of social media influencers, scholars have begun to classify them by the size of their follower base (e.g., mega-, macro-, micro-, and nano-influencers) or by level of influence (e.g., celebrities, key opinion leaders such as journalists, academics, and industry experts) (Casero-Ripollés 2020). Importantly, influencers are not necessarily individuals, but may be corporations or other social, political, or media groups (Enke and Borchers 2019).

Characteristics of classical social network structure can be replicated in social media, but new digital platform relationships may redefine the attributes of influence. For example, opinion leaders, with more knowledge resources, might be better able to engage in producing content than those who are not (Yoo and Alavi 2004; Cassell et al. 2006; Welser et al. 2007). The communicative structures of social media are often determined by two overlapping and interdependent networks. The first is the long-term and relatively stable opinion follower–followee relationship (Bruns et al. 2017). The second, a relatively short-term and emergent network, based on shared interest in a topic, is coordinated by a common hashtag (Chang 2010).

Empirical studies have identified the characteristics of network influencers, such as attributes of *influence* (Diani 2003; Dubois and Gaffney 2014; Vaccari et al. 2013) and the *socio-demographic factors* of opinion leaders (e.g., Lassen and Brown 2011). Other scholars have developed methods to identify network influencers or opinion leaders (Feng 2016; see also, Cataldi et al. 2013). Research on cross-platform comparisons of network influencer characteristics, especially Twitter/Weibo network structure, is still embryonic.

4 Social media text mining and topic modeling

Text mining is defined as a multi-purpose research method to study a wide range of issues by systematically and objectively identifying characteristics of large sample data (Radovanović and Ivanović 2008; Jo et al. 2020). Social media's sheer scale has led users to share topics of interest, generating a large volume of data, specifically user-generated text, and yielding a new type of text-based discovery (Smith et al. 2012). The challenge that arises is understanding and applying relevant methodological tools for mining this data, and analyzing it to reveal i) implicit connections among users and ii) meaningful insights into the spread of information, opinions, topics, emerging issues, and trends (Leskovec 2011; Agrawal et al. 2011). Zeng et al. (2010) suggested that the rich data in social media (e.g., hashtags, subjective opinions, ratings, and user profiles) may effectively present social interactions among users, though its semantic inconsistencies/inaccuracies, misinformation, lack of meaningful structure, and dynamic nature may complicate analysis, as may the volume of the data itself.

Different disciplines offer various analytical techniques and methods to tackle these challenges in mining and analyzing social media text (Nagarajan et al. 2011). Subjectivity analysis and opinion mining, or processing social media text using specially designed computational systems, are shared objectives of various fields (Robinson et al. 2012). Text mining, the process of exploring large amounts of unstructured

text data with tools that can analyze its patterns, latent topics, keywords, and concepts, may help analyze problem-solving techniques or topic summaries (Rahmani et al. 2014), time-trends, or social networks (Salloum et al. 2017).

Recent studies have applied various social media text mining methods. Reddick et al. (2017) analyzed Facebook topics and comments to understand key people, places, organizations, and services mentioned in public posts to identify issues for improving the quality of citizen-centered public services. Kavanaugh et al. (2011) established a need for aggregation tools to sufficiently interpret vast amounts of data, in order to model the flow of information and identify changing patterns over time. Stieglitz et al. (2012) used social media text mining tools to identify trending political topics.

Machine learning-based text analysis offers various models for interpreting text data. The topic modeling method can help identify and explain general topics of interest on social media without a prepared text data set with predefined schemes; this method can find latent topics, or patterns hidden within the text data, on its own without supervision. This autonomy makes topic modeling popular with researchers, and one of its mostly widely used algorithms is Latent Dirichlet Allocation (LDA).

5 Methodology

5.1 Data pre-processing

In this section, we suggest our method for social media cross-platform comparative analysis. Mining text from free-information systems such as Twitter and Weibo is inherently noisy, generating a large number of errors (Hecht et al. 2011; Grimaldi 2019), requiring pre-processing operations (for both English and Chinese) such as *data cleaning*, *word segmentation*, and *weight calculation* (Tang et al. 2005). While these procedures can be implemented using the open-source Python scripting language, the result may still involve manual correction depending on data structure or quality or research purposes.

In preparation for text mining, we cleaned the data by filtering all the punctuations, mis-spellings, special characters (e.g., emojis and URLs), and stop words. The main purpose of this step was to ensure strict replication of previous text mining analyses of social media data (Lim and Buntine 2014; Guo et al. 2016) to discover latent topics in Twitter and Weibo. Regardless of the language used, documents

may be incorrectly segmented due to a lack of reference vocabularies (Wu et al. 2020). Word segmentation tools can resolve this problem and enhance processing. For Chinese words, we used JIEBA,¹ a popular module that segments each sequence, as applied in several previous studies (Day and Lee 2016; Li et al. 2018; Jiang et al. 2019). Natural Language Tool kit (hereafter NLTK),² a Python-based tool for processing human language data (Bird et al. 2009; Contreras et al. 2018), represents each document with comments as a vector of words, useful for word segmentation and stop-word removal in processing English text.

JIEBA has limitations in fully identifying some typical Chinese words, such as "美联储" (abbreviation for "Board of Governors of The Federal Reserve"). This may cause errors or bias in research results, requiring manual correction. For stop-word removal, we suggest using Chinese stop words from Apache Lucene's Smart Chinese Analyzer (<https://lucene.apache.org/>), which uses probabilistic knowledge to find the optimal word segmentation for (simplified) Chinese text (or mixed Chinese-English text). We converted language characters and eliminated punctuation.

We used Term Frequency-Inverse Document Frequency (hereafter, TF-IDF) to calculate the weights factors. TF-IDF is a statistical measure that analyses how important or relevant a word is to one in a collection/corpus of documents using a "Bag-of-Words" scheme (Kim et al. 2019). TF-IDF is one of the most widely used numerical statistics for transforming unstructured documents into a structured numerical vector, often in automated text analysis in machine learning algorithms (Robertson 2004), and its tools generally support language processing in both English and Chinese (Zhang et al. 2011). The method also assumes that if a word is important to a document, it should repeatedly appear in this document and not in others (Ramos 2003), which is why the TF-IDF produces statistics of term frequency (hereafter TF) and inverse document frequency (hereafter IDF). The TF-IDF is calculated as follows:

$$TF-IDF_{ij} = tf_{ij} \times \log \left(\frac{N}{df_i + 1} \right)$$

,where the parameter tf_{ij} denotes the raw count of a term in a document; word i appears in document j . The parameter df_i denotes the number of documents in which word i appears at least once. N denotes the total number of documents in the corpus. If a word i is assumed to be important for document j , it should have a large TF (tf_{ij}) and a small document frequency (df_i). We used Scikit-learn's Python libraries³ to import CountVectorizer and TfidfTransformer to calculate

TF-IDF. Since the previous data cleaning process removed the stop words (such as "a," "the," "it," "this"), those would not influence the results of TF-IDF data processing.

5.2 Latent Dirichlet allocation (LDA)

Text mining often involves the method for fitting a topic model, where each document is a mixture of topics and each topic is a mixture of words. LDA, proposed by Blei et al. (2003) as a generative probabilistic model for discrete datasets such as text sets, can be used as a topic model to discover implicit topics in text documents (Wang 2018). The large volume of user posts/comments displayed in text form, and their high dimensionality, imply various latent topics, requiring a method to extract those with relevance (Jelodar et al. 2019). We can use dimensionality reduction models to understand the nature of the text and words.

LDA can be used as a step to analyze large-scale social media entries automatically (Blei et al. 2003; Daud et al. 2010; Hagen et al. 2015; Qian, Zhang et al. Song et al. 2014; Mishler et al. 2015). Text, words, and topics in social media have been often studied in a static manner, but research on time-trends may be of significant interest for scholars across disciplines. A three-tiered Bayesian model of "document-topic-word" presents each dataset (e.g., each text) as a mixed set of unknown topics. Accordingly, each topic is modeled as a mixed probability distribution (Wang and McCallum 2006), while LDA is based on the posterior distribution framework (Girolami and Kabán 2003). Unsupervised training using LDA as the topic model text in the dataset yields the implicit topics. The computational problem of inferring the underlying topic structure from the document is to calculate the posterior distribution, which is the conditional distribution of implicit variables for a given document. (Blei 2012). The LDA analysis processing in this article is implemented using the open-source Python scripting language, generating the probability that a word belongs to a topic, which is used to sort the words according to their probability scores). The generative process of the corpus is shown in Appendix 1.

We should note that recent studies have criticized the use of LDA for analyzing social media data, which may suffer from the sparsity problem of short text classification (Alvarez-Melis and Saveski 2016; Mehrotra et al. 2013, Harrysson 2016). As social media posts are often short and may include mis-spelled words or irrelevant characters (such as emojis), social media text documents share an extremely low number of overlapping terms within a collection of posts. To address the sparsity problem, scholars have suggested alternative methods, such as LDA extension to author-topic model, and the dual LDA approach that relies on external

¹ Jieba Chinese word segmentation tool, <https://github.com/fxsjy/jieba>.

² NLTK. <http://www.nltk.org/>.

³ Scikit-learn. <https://scikit-learn.org/stable/>.

knowledge bases like Wikipedia (Atefeh, and Khreich 2015; Nugroho et al. 2020; Rafeeqe and Sendhilkumar 2011).

Alternative LDA models or extensions have their own limitations when implemented in specific social media data. For example, Nugroho et al. (2020) suggested that standard LDA performs, in fact, better than the author–topic model extension in the social media environment, and dual LDA only increases computational complexity. In addition, using external sources (content expansion) such as Wikipedia may present issues of reliability and uneven quality among documents when used in the dual LDA process, a problem which is amplified when comparing platforms such as Twitter and Weibo. In this work, we addressed the sparsity issue through extensive manual screenings on mis-spellings and emojis, and used the content merging method, by aggregating all comments generated in a certain unit of time (a week) into a single entity, as suggested by Weng et al. (2010) and Nugroho et al. (2020).

However, several methods can be used to determine the optimal number of topics for the LDA model. For example, Blei et al. (2003) suggested using the perplexity score as an indicator to select the best generalization performance. Hajjema and Latirafind (2017) specified the optimal number by looping through models with different numbers until they found the maximum log-likelihood. Zhao et al. (2015) proposed a heuristic way to estimate the most appropriate number of topics, to avoid some defects in accuracy and logic. Given the focus of our corpus on the topic of “trade war,” which has a high log-likelihood of topics when employing the LDA model, we adopted the heuristic approach suggested by Wang (2018) and Zhao et al. (2015), an adaptive method to identify the optimal number of selected topics.

5.3 k-means clustering

Clustering helps us understand social media data by grouping observations together. Widely used algorithms for clustering include k-means clustering (Likas et al., 2003), k-medoids (Park et al., 2009), hierarchical clustering (Murtagh and Contreras 2012), and ST-DBSCAN (Birant and Kut 2007). In this study, we suggest the k-means clustering method for re-grouping the LDA-produced latent topics into (fewer) topic categories. The unsupervised k-means clustering algorithm is a vector quantization method that aims to automatically organize a large volume of documents into a small (k) number of meaningful clusters, discovering underlying structures in unlabeled sets of documents (Bui

et al. 2017).⁴ As the k-means clustering is a centroid-based (or distance-based) algorithm, the method aims to minimize the sum of distances between the points in a cluster and their centroid (within-cluster variances), where each observation belongs to the cluster with the nearest mean (Li et al. 2016). The advantages of the k-means clustering algorithm include the following: (i) processing different types of attributes, (ii) discovering clusters with arbitrary shape, (iii) minimal requirements for domain knowledge to determine input parameters, (iv) dealing with noise and outliers, and (v) minimizing the dissimilarity between data (Gan et al. 2007). The squared Euclidean distance measure of the k-means algorithm is,

$$E = \sum_{i=1}^k \sum_{j \in C_i} \|x_j - c_i^2\|,$$

where k denotes the number of clusters, x_j denotes the j th data point in the i th cluster C_i , and c_i denotes the centroid of C_i . The notation $x_j - c_i^2$ stands for the distance between x_j and c_i . To achieve a representative clustering, a sum of squared error function, E , should be minimized (Kim and Gil 2019).

To address a limitation of k-means clustering, specifically that the k number of clusters needs to be pre-specified by the user/analyst (Tu et al. 2014), researchers are often required to apply a systematic method for determining the optimal number of clusters in a dataset. One simple, and popular, method would be to inspect the dendrogram produced by hierarchical clustering. However, this method is still prone to subjectivity, and users/analysts often lack sufficient domain knowledge or a priori information to properly analyze the dendrogram (Salvador and Chan 2004). In response, much literature has addressed the selection of the optimal k with multiple runs of k-means, from the trial-and-error approach (Krzanowski and Lai 1988; Hamerly and Elkan 2004; Žalik 2008).

In this article, we suggest the curvature-based “*elbow*” (or *knee*) method, which finds a cutoff point that determines the optimal number of clusters (Sugar and James 2003; Zhang et al. 2017). The *elbow* method is a heuristic used to determine the optimal number of clusters by plotting the explained variations as a function of the number of clusters and choosing the *elbow* (or *knee*) of the curve as a cutoff point (Kodinariya and Makwana 2013; Marutho et al. 2018). Zhang et al. (2017) suggested that this curvature-based⁵ method, the general *elbow* method with visual inspection, may still be ambiguous, especially when there are high levels of intermix between clusters. The method proposed by

⁴ k-means has been applied to various fields, such as web mining (Lingras and West 2004; Yadav et al. 2012), pattern recognition (Likas et al. 2003; Peng et al. 2013), and the analysis of social networks (Yang et al. 2010; Yin and Liu 2016).

⁵ The curvature is the degree to which a geometric object deviates from flatness, or, in the case of a straight line, from straightness.

Zhang et al. (2017) aims to locate the *knee* (or *elbow*) by analyzing the curvature, given that within-cluster variance $J(k)$ is:

$$J(k) = \sum_{j=1}^k \sum_{x_i \in C_j} \| \mathbf{x}_i - \bar{\mathbf{x}}_j \|^2,$$

where $\bar{\mathbf{x}}_j$ is the sample mean of class j and C_j is the set of samples that belongs to class j . The use of raw $J(k)$ can lead to ambiguity, because $J(k)$ decreases monotonically as k increases. Therefore, Zhang et al. (2017) suggested that the *elbow* (or *knee*) should correspond to the maximum curvature point, where the curvature is defined as $k = y'' / (1 + y'^2)^{3/2}$, for a curve given as $y = f(x)$, and that the researcher should choose the k with the highest peak value to be returned within the arguments of the maxima:

$$K = \operatorname{argmax}_k \frac{J''(k)}{J'(k)}$$

5.4 Social network analysis

Social network analysis is a method used to analyze theoretical constructs and concepts defined as relational processes and outcomes (Wasserman and Faust 1994). A social network consists of nodes and ties. Nodes, or actors, within a network, can represent users, organizations, or communities (Binz-Scharf et al. 2012; Provan et al. 2007; Scott 2015). In this article, which is concerned with understanding and mapping the differing natures of Weibo and Twitter influencer networks, the edges of a graph represent the actor, and the available lines between those edges represent the correlation between actors (Bezerianos et al. 2010). The interactions among users were created when they @mentioned, retweeted or reposted, or replied in specific posts (Zhang et al. 2019).

Several indicators/measurements in social network analysis show how the key actors are networked within a social structure (Kolli and Khajeheian 2020; Martin et al. 2014). The most widely used indicators are: nCloseness (Freeman 1978; Rowley 1997; Perez and Germon 2016), Farness (Hoffmann et al. 2014; Kolli and Khajeheian 2020; Martinez-Lopez et al. 2009), DegreeCent (Srinivas and Velusamy, 2015; Bródka et al. 2011), and NrmDegree (Abbasi and Altmann 2011; Leydesdorff 2007). nCloseness (normalizing closeness) is based on the notion of distance between nodes (Freeman 1978; Rowley 1997), calculated as the average shortest path length from a node to every other node in the network (Perez and Germon 2016).

Farness is equal to the sum of the distance from a node to all the other nodes (Hoffmann et al. 2014), and therefore is the reciprocal of the nCloseness (Kolli and Khajeheian

2020). Farness highlights nodes that may reach any other nodes within a few hops and nodes that may be very distant in the graph (Martinez-Lopez et al. 2009). DegreeCent (degree centrality) is defined as the number of ties incident upon a node, indicating the relative importance of a node in the network (Srinivas and Velusamy 2015)—for a given node x , it is calculated as a ratio between the number of nodes connected with node x and the total number of all nodes in the network (decreased by one) (Bródka et al. 2011). NrmDegree (normalized degree) centrality is the degree divided by the maximum possible degree expressed as %—normalized values should only be used for binary data (Abbasi and Altmann 2011; Leydesdorff 2007). Using these four indicators, we can calculate which top-ranked network influencers are included in each social media platform's social structure.

6 Application of method to twitter and Weibo cross-platform comparison

6.1 The empirical case: U.S.-China trade war discourse

This article illustrates how our suggested method can be applied to a real-life example, discourse on the U.S.-China trade war on Twitter and Weibo. We used a single rather than multiple cases, since we aimed to examine case dynamics between two cross-country platforms. For context, the role of social media in engagement on international political issues has been widely recognized (Romero et al. 2011). Social media stores users' subjective thoughts on events and provides an opportunity for politicians and political activists to engage in dialogue with supporters of different ideologies (Jenkins and Carpentier 2013).

6.2 Data collection

We collected and analyzed tweets (posts) that contained specific keywords and hashtags, where users place information on specific topics or content (Grimaldi 2019). To facilitate the time node's determination, we collected tweets/posts on Twitter and Weibo between June 29 and September 30 2019, using Python. The sample size was $N = 36,035$ tweets and $N = 17,568$ Weibo posts during the study period. Our data collection proceeded as follows: i) searched tweets/posts using keywords related to trade war on different platforms

Fig. 1 Visual inspection based on the curvature-based method

Table 1 Latent topics, categories, and probability scores identified by LDA (Twitter data)

Categories of topics	Probability scores	Latent topics (top words)
Economy recession and tariffs	0.398	Topic 0 (tariffs, recession, market, economic, maga), Topic 5 (stock, soybeans, bonds, dowjones, price) Topic 6(spy, Trumprecession, bitcoin, futures, consumer), Topic 15(trumprecession, trumpslump, businesses, escalates, trump tariffs), Topic 18 (deal, farmers, prices, money, fall), Topic 19 (currency, yuan, manipulator, bitcoin, Dollar), Topic 22 (oil, stockmarket, consumers, winning, uncertainty)
Trade and supply chain	0.296	Topic 1 (exports, investing, reuters, asia, vietnam) Topic 2 (g7summit, negotiations, trade-talks, international, google) Topic 4 (Huawei, amzn, supplychain, korea, taiwan), Topic 8(ArtificialIntelligence, Industrial, purchases, Alibaba, imf) Topic 9(Europe, soybeans, foreign, forextrading, fdi) Topic 10(exports, pork, Huawei, Geotech, Turnberry), Topic 16 (tensions, rate, imports, goods) Topic 21 (consulting, logistics, supplychains, curve, valuechains) Topic 23 (huawei, truce, manufacturing, tech, links)
Public opinion and media	0.216	Topic 12 (voices, foxnews, affect, worries, measures) Topic 17 (brexit, thinking, correction, options, opinion) Topic 20 (fed, currencywar, administration, youtube, financial-times), Topic 24 (article, watch, story, pressure)
Government and politics	0.088	Topic 3 (g20, trump, xijinpai, Osaka, agree) Topic 7(election, iran, Chinese, delegation, malaise) Topic 11(hongkong, fed, situation, federalreserve, voices) Topic 13 (hongkong, realdonaldtrump, inners, impeach, delays) Topic 14 (trump, impose, bbc, borisjohnson, announced)

(Twitter=trade war; Weibo=贸易战), ii) searched hashtags related to trade war (#TradeWar on Twitter; #贸易战 on Weibo), iii) collected text data from Twitter and Weibo at the same timeframe each day (9 am China standard time and 8 pm EST in the United States), iii) identified users' account name, content, and time (Haustein et al. 2014), and iv) collected all tweets/posts generated by users that contain one or more keywords (or hashtags), and stored in the separate dataset.⁶

⁶ The tweet (posts) datasets are available from <https://docs.google.com/spreadsheets/d/e/2PACX-1vTGen2mVRx2fWwFueWmheTli oVdAr8Cz7UY3d15fZbMLgiDrYXJgqGELFqwMvd-UILJWoz1AJ Pag-If/pubhtml>.

⁷ The LDA algorithm code is available at <https://github.com/yiyan gcpa/Cross-Platform-Comparison-of-Framed-Topics-in-Twitter-and-Weibo>.

6.3 Topic modeling results

The topic extraction processing was executed by a Python-aided method.⁷ We used scikit-learn package for LDA modeling analysis on the Twitter keywords ($N=685,131$) and Weibo keywords ($N=980,878$) after pre-processing. From the aforementioned heuristic (adaptive) approach suggested by Wang (2018) and Zhao et al. (2015), we found that the extraction seemed most efficient when the total number of classified topics was 25 on Twitter and 30 on Weibo. Our program was set to 2000 iterations to maintain accuracy and minimize the likelihood of error in the results.

As the next step to determine the number of topic categories, we employed the aforementioned k-means clustering algorithm to calculate the cluster center that shows specific topic groups, based on the Euclidean distance between the TF-IDF

Table 2 Results for each of the five days of the string-pulling puzzle

Categories of topics	Probability scores	Latent topics (top words)
US-China Politics	0.362	Topic 1(国际=international, 中美=US-China, 香港=Hong Kong, 特朗普=trump) Topic 2(关系=Relationship, 英国=UK, 美方=The United States, 人民=The people) Topic 3(商务部=Ministry of Commerce, 人民日报=People's Daily, 特朗普=Trump, 沟通=Communication), Topic 6(世界=World, 朋友=Friend, 参议员=Senator, 刘鹤=Liu He) Topic 9(联合=Alliance, 美国 商务部=US Department of Commerce, 大阪=Osaka) Topic 12(制度=Regime, 黄奇帆=Huang Qifan, 华盛顿=Washington, 副部级=Deputy level), Topic 27(美国=America, 特朗普=Trump, 文化=Culture, 暴乱=Riots) Topic 29(商务部=Ministry of Commerce, 政府采购=Government Procurement, 基础教育=Basic Education, 研讨会=Seminar)
Stock Market & Foreign Exchange Rates	0.351	Topic 4(行情=quotes, 增加=Increased, 市场=Market, 下跌=Falling) Topic 10(恒生指数=Hang seng index, 期货=Futures, 引渡=Extradition) Topic 11(上证指数=Shanghai Composite Index, 征收=Collection, 稀土=Rare Earth, 品牌=Brand), Topic 14(原油期货=Crude Oil Futures, 人民币=RMB, 高峰=Peak, 收盘=Closing Quotation) Topic 15(金融=Financial, 产业=Industry, 赢家=Winner, 供应=Supply) Topic 17(制造商 Manufacturers, 报告=Report, 材料=Materials, 个股=Individual stocks) Topic 22(指数=Index, 沈逸=Shen Yi, 推迟=Put off, 高盛=Goldman Sachs) Topic 24(上证指数=Shanghai Composite Index, 衰退=Decline, 汽车=Car, 美国公司=US company) Topic 25(赵穗生=Zhao Suisheng, 农业=Agriculture, 股指=Stock Index, 冲击=Shock) Topic 26(市场=Market, 世界=World, 全球=Global, 股市=Stock Market) Topic 28(汇率=Exchange Rate, 美元=US Dollar, 贬值=Devaluation, 同比=year-on-year)
Export Industry	0.192	Topic 0(出口=Exports, 增长=Growth, 行业=Industry, 不确定性=Uncertainty) Topic 7(达成协议=Reach an Agreement, 苹果=Apple Inc., 工厂=Factory, 亚洲=Asia) Topic 8(市场=The Market, 韩国=South Korea, 半导体=Semiconductor, 手机=Mobile Phones) Topic 18(华为=Huawei, 企业=Corporate, 产品=Products, 零件=Components) Topic 19(产业=Industry, 贸易=Trade, 农产品=Agricultural Products, 合作=Cooperation) Topic 21(华为=Huawei, 芯片=Chip, 生产=Produce, 新加坡=Singapore)
Media & Youth	0.092	Topic 5(关注=Pay Attention to, 应对=Reply, 围观=Onlookers, 紧张=Stress) Topic 13(中学校长=High School Principals, 朋友圈=Friends Circle, 年轻人=Young Man, 开学=Starting School) Topic 16(央视=CCTV, 人民日报=People's Daily, 胡锡进=Hu Xijin, 新浪=Sina.com) Topic 20(青年=The youth, 教授=Professor, 时代=Era, 复旦大学=Fudan University) Topic 23(反应=Response, 总统=President, 央视=CCTV, 新闻联播=News Broadcast)

value of the topic and the value of each cluster center—this allocates a topic to a cluster (topic category) with high levels of similarity. Figure 1 visualizes the statistics for the curvature-based *elbow* (or *knee*) method used to specify the number of clusters (topic categories), where the y-axis represents the curvature value of $J''(k)/J'(k)$, while the x-axis is the corresponding cluster number k . Through the curvature method, we specified the k with the highest peak value to be returned, which was $k = 4$ for both the Twitter and Weibo datasets.

Tables 1 and 2 show the results of LDA analysis generating latent topics, probability scores, and categories on Twitter and Weibo data, respectively. The co-occurring words were sorted from highest to lowest average fit. This article's length limited us to the top four or five topics in the list of keyword probability values. For the Twitter data analysis, we identified categories of topics and their probability values, namely *Economy Recession & Tariffs* (0.3986), *Trade & Supply Chain* (0.2961), *Public Opinion & Media*

(0.2162), and *Government & Politics* (0.0888). On Weibo, we identified topic categories and their probability values, namely *U.S.-China Politics* (0.3627), *Stock Market & Foreign Exchange Rates* (0.3519), *Export Industry* (0.1923),

and *Media & Youth* (0.0929). Time-trends, which illustrate changes in levels of posts on certain topics over time and relative focus, appear in Figs. 2, 3, and 4.

Fig. 2 The number of comments by social media platform

Fig. 3 Topic time-trend (Twitter)

Fig. 4 Topic time-trend (Weibo)

6.4 Social network analysis results: mapping the characterizes of network influencers

We conducted social network analysis to map the key network influencers (opinion leaders) on Twitter and Weibo. We first generated the relationship matrix of key network influencers, then mapped the whole-network, and finally calculated centrality measures of the top-ranked network influencers. Following the procedures used by previous studies on social media actor relationship extraction (Zhang et al. 2019; Kolli and Khajeheian 2020), we selected all accounts containing "@" (or a mention of a certain account) in the tweets/posts in the text mining data we produced for topic modeling, yielding 206 Twitter accounts and 183 Weibo accounts. The matrix data were also imported into UCINET 6.

We then mapped the network of Twitter and Weibo's key network influencer accounts. We identified these by type (i.e., media, politician, corporation, government/party, NGO/non-profit, international organization, entrepreneur, ordinary citizen, and so on) and presented them using different colors (see Figs. 5 and 6). Figure 7 shows the proportion of "mentions" for key network influencers by type for both platforms, counted in a comparison chart.

In terms of the types of network influencers in U.S.-China trade war discourse on social media, *Media* is the largest group @-mentioned in our sample on both platforms, accounting for 27.64% on Twitter, including 33 entities, such as @WSJ, @CNN, @WashingtonPost, and @FoxNews. This is followed by *Overseas Media* from the UK, China, and elsewhere (23%), including @FinancialTimes, @Reuters, @BBC, @ChinaDaily, @cctvnews, @XHNews, @the_hindu, and @ndtv.

The category *Politicians* on Twitter took up 22.11%, including members of the White House, Congress, and local councils. Incumbent politicians such as @realDonaldTrump, @POTUS, @GOP @VP, and @SecretaryRoss were found to form the core of the social network in the trade war discourse, in addition to a few non-American politicians, such as @narendramodi, @ScottMorrisonMP, and @AbeShinzo.

No individual politician's personal account appears on the Weibo platform's influencer network on this topic. *Media* accounts for 43.11% of our Weibo data, with China's state-owned media making up 51.39%, including accounts such as @People's Daily, @People.com, and @CCTV News. Chinese mainland general/private owned media accounts for 47.22% (@QQnews, @163News); another influential media account was @HongKongCommercialDaily in Hong Kong.

Fig. 5 Influencer network on Twitter (on trade war discourse)

Researchers/commentators (called experts in China) have shown some degree of centrality as network influencers on Weibo, accounting for around 17% of “mentions” in the total influencers network on the trade war topic, compared with about 4% on Twitter. These influencers include @YaoYang (the dean of the National Institute of Development at Peking University, specializing in economics and international relations), @QiuZhenhai (an expert on China's transformation and strategy and adjunct professor in Shanghai), @PoliticalCommissarCanrong (the deputy dean of the School of International Relations at Renmin University of China, whose research interests include the U.S. political system and culture and China-U.S. relations), and @YifuLin (World Bank chief economist and senior vice president and professor at Peking University). This group of influencers ranks second only to *Media* on Weibo. It is worth noting that postings from Chinese state-owned media and researchers on Weibo constitute a more significant proportion of postings than their Western/American counterparts.

Corporation on Twitter and Weibo represented 20.6% and 14.37% of our @-mention data, respectively. Twitter

accounts included 22 U.S. companies, nine Chinese companies, and four German companies, while Weibo included 20 Chinese and four U.S. companies. On Twitter, *Government/party* accounts comprised 8.04% of the total mentions, including @WhiteHouse, @GOP, @TheDemocrats, and @HouseDemocrats. By contrast, the proportion of *government/party* accounts on Weibo was 4.79%, with the major representative accounts being @CommunistYouthLeague-Central and @ChinaGovernment. NGOs and other international organizations together accounted for 3% and 2% of the Twitter sample, respectively. The Weibo influencer network contained no *NGO* accounts, but did include one interesting category distinct from Twitter with a specific impact: *Schools/research institutions*. These academic institutions together accounted for 4.79% of the total sample, including institutions such as @ChinaInstituteofInternationalStudies and @FudanUniversity. In addition, Weibo represented only @WTO and @IMF accounts as international organizations.

Some individual Chinese *Citizens* on Weibo were found be centrally located in the network. "Celebrity culture" is highly influential and impactful in Chinese social media,

Fig. 6 Influencer network on Weibo (on trade war discourse)

Fig. 7 Proportion of Twitter and Weibo key network influencer types

Table 3 Centrality scores of Twitter's network influencers (on trade war discourse)

Users	Twitter account holders	nCloseness	Farness	DegreeCent	NrmDegree
@realDonaldTrump	Donald J. Trump	22.747	466	70	66.038
@Reuters	Reuters	20.463	518	25	23.585
@CNN	CNN	20.307	522	29	27.358
@WSJ	Wall Street Journal	20.307	522	27	25.472
@CNBC	CNBC	20.229	524	27	25.472
@FinancialTimes	Financial Times	20.190	525	24	22.642
@MSNBC	MSNBC	20.076	528	23	21.698
@GOP	GOP	20.038	529	26	24.528
@business	Bloomberg	20.038	529	25	23.585
@FoxNews	Fox News	20.000	530	24	22.642
@ABC	ABC News	19.887	533	22	20.755
@ReutersBiz	Reuters Business	19.813	535	20	18.868
@BBC	BBC	19.739	537	19	17.925
@CNBusiness	CNN Business	19.703	538	18	16.981
@WSJbusiness	WSJ Business News	19.703	538	18	16.981
.....					
	Minimum	0.952	466	1	0.943
	Maximum	22.747	11,131	70	66.038

Network Centralization = 59.48%, Heterogeneity = 2.24%, Normalized = 1.32%, Actor-by-centrality matrix saved as dataset FreemanDegree

Table 4 Centrality scores of Weibo's network influencers (on trade war discourse)

Users	Weibo account holders	nCloseness	Farness	DegreeCent	NrmDegree
@央视新闻	CCTV News	12.482	697	29	33.333
@人民日报	People's Daily	12.306	707	24	27.586
@Apple支持	Apple	12.219	712	12	13.793
@新浪财经	Sina Finance	12.117	718	14	16.092
@环球时报	Global Times	12.000	725	21	24.138
@商务微新闻	Ministry of Commerce News	11.918	730	6	6.897
@胡锡进	Xi-Jin Hu	11.902	731	17	19.540
@参考消息	Cankao Xiaoxi (Reference News)	11.789	738	17	19.540
@新浪军事	Sina Military	11.757	740	15	17.241
@CCTV4	CCTV4	11.741	741	14	16.092
@宋忠平	Zhong-Ping Song	11.741	741	14	16.092
@军情直播间	Military Situation Live	11.741	741	14	16.092
@洪琳	Lin Hong	11.741	741	14	16.092
@伊朗驻华大使馆	Embassy of the Islamic Republic of Iran	11.741	741	14	16.092
@邱震海	Zhen-Hai Qiu	11.741	741	14	16.092
.....					
	Minimum	1.149	697	1	1.149
	Maximum	12.482	7569	29	33.333

Network Centralization = 26.14%, Heterogeneity = 1.88%, Normalized = 0.75%, Actor-by-centrality matrix saved as dataset FreemanDegree

often more so than traditional actors. For instance, in the trade war network, @tidaotanhuaazaimianbei and @remenbi-aoge posted opinions on international (U.S.-China) relations and politics every day, and some of these became viral. They

have 3,800,000 and 370,000 followers respectively, accounting for 5.39% of total Weibo mentions in our data.

Table 3 shows the centrality measures of Twitter accounts' top 15 network influencers, including @realDonaldTrump, @CNN, @WSJ, @CNBC, @Reuters, @

FinancialTimes, @GOP, @MSNBC, @business, and @FoxNews, which have the highest average ranking. Network measures show that Network Centralization = 59.48%, Heterogeneity = 2.24%, and Normalized = 1.32%. The results suggest that eight of the top ten are media accounts. Among the top ten Weibo accounts, 60% is news media, such as @CCTV News, @People's Daily, @Sina Finance, @Global Times, and @Cankaoxiaoxi (see Table 4).

7 Conclusion

By applying text mining and social network analysis, this article illustrates how the comparative method can be used to analyze social media interactions among English-(Twitter) and Chinese-speaking (Weibo) users, with a consistent analytical framework for comparison. Our suggested method shows how network patterns contribute to our understanding of public opinion and comment frameworks, and how the LDA-based topic probability modeling can explain latent topics in public opinion on social media. We illustrate that the social network analysis method can be used to compare the nature of social interactions between different types of network influencers on Twitter and Weibo. For example, our results on trade war discourse show that the networking concentration of Twitter was 59.4%, while Weibo's was much lower (26.14%), indicating that Twitter networking power is concentrated to a small core set of actors. Weibo is relatively loose and mainly concentrated on China's state-owned media, in which the status of every actor is relatively equal. The relationship between power distribution and major opinion leaders has long been the subject of structural social network analysis.

There are limits to our methodological approach and application. Firstly, due to the comparative nature of the methods we suggested, our experimentation did not fully employ some recent methods that complement LDA, such as hashtag-pooling (Mehrotra et al. 2013; Kant et al. 2020). Hashtag-pooling treats hashtags as topical markers, indicating the context or core idea of a tweet, which may further improve LDA modeling quality. However, for this particular study, hashtag-pooling may not evenly capture the characteristics of social media posts in each platform, due to differences in the proportion of posts that used hashtags (34.64% on Twitter vs. 19.50% on Weibo) or in the number of total posts (36,035 tweets / 12,481 posts). Nevertheless, hashtag-pooling will certainly improve the LDA method in future cross-platform comparative methods, as more and more Weibo users utilize them.

Secondly, sparsity of social media text has not fully addressed our LDA method, which remains a topic for future work. Future social media studies with LDA-based

cross-platform comparison may need to develop a reliable method, either content merging or content expansion, that does not involve computational complexity. For example, a word's specificity weighting scheme can be used in the topic assignment process, or training methods based on deep learning-based models can be employed for big(ger) data on social media platforms.

Thirdly, our methods are appropriate only for text mining from English-speaking Twitter and Chinese-speaking Weibo: individuals and organizations from countries with other languages might be underrepresented. Identifying the country of origin, mapping the public opinion network of the entire international community, and evaluating the impact of social media public comments toward different countries are of significant value in future research. Lastly, our methods relied on keywords, hashtags (#), and mentions (@), which may truncate data from texts/documents without them. Our methodology's automated procedures do not fully capture the many slang and shortened words on social media.

Appendix 1

The generative process of the corpus is as follows.

For all topics $k \in [1, K]$ do

Sample mixture components $\varphi_k \sim \text{Dir}(\beta)$

End for

For all documents $m \in [1, M]$ do

Sample mixture proportion $\theta_m \sim \text{Dir}(\alpha)$

Sample document length $N_m \sim \text{Pois}(\xi)$

For all words $n \in [1, N_m]$ do

Sample topic index $k_{m,n} \sim \text{Mult}(\theta_m)$

Sample term for word $w_{m,n} \sim \text{Mult}(\varphi_{k_{m,n}})$

End for

End for

Where, α denotes the parameter of the Dirichlet prior distribution of topics on documents in LDA, and β denotes the parameter of the Dirichlet prior distribution of words on topics in LDA. φ_k is the parameter of multinomial distribution of words on topic k . θ_m is the parameter of multinomial distribution of topics on documents. M is the collection of documents; K is the collection of topics. N_m is the length of the document. $w_{m,n}$ is the n th word of document. $k_{m,n}$ is the n th topic of the document (Zhang et al. 2019; Zhu et al. 2016).

Acknowledgements Yi Yang's contribution to this research is partially funded by National Social Science Foundation of China (grant number: 19ZDA104) and State Scholarship Foundation of China Scholarship Council (grant number: 202006160125).

References

- Abbasi A, Altmann J (2011) On the correlation between research performance and social network analysis measures applied to research collaboration networks. In: 2011 44th Hawaii international conference on system sciences, pp. 1–10. IEEE.
- Agrawal D, Budak C, El Abbadi A (2011) Information diffusion in social networks: observing and influencing societal interests. In: Proceedings of VLDB'11
- Alvarez-Melis D, Saveski M (2016). Topic modeling in twitter: Aggregating tweets by conversations. In: Proceedings of the international AAAI conference on web and social media, vol. 10, No. 1.
- An L, Yu C, Lin X, Du T, Zhou L, Li G (2018) Topical evolution patterns and temporal trends of microblogs on public health emergencies. *Online Inf Rev* 42(6):821–846
- Atefeh F, Khreich W (2015) A survey of techniques for event detection in twitter. *Comput Intell* 31(1):132–164
- Banks A, Calvo E, Karol D, Telhami S (2020) # polarizedfeeds: three experiments on polarization, framing, and social media. *The Int J Press/Politics* 26(3):609–634
- Bezerianos A, Chevalier F, Dragicevic P, Elmqvist N, Fekete JD (2010, June) Graphdice: A system for exploring multivariate social networks. In: Computer graphics forum, vol. 29, No. 3, pp. 863–872. Oxford, UK: Blackwell Publishing Ltd.
- Binz-Scharf MC, Lazer D, Mergel I (2012) Searching for answers: Networks of practice among public administrators. *The Am Rev Public Admin* 42(2):202–225
- Birant D, Kut A (2007) ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowl Eng* 60(1):208–221
- Bird S, Klein E, Loper E (2009) Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc."
- Bird D, Haynes K, Ling M, O'Brien J (2011) The use of crowd sourcing for gathering information about natural disasters. *Risk Frontiers Newsletter*, 11(2):1–4
- Blei DM (2012) Probabilistic Topic Models. *Commun ACM* 55(4):77–84
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3(Jan):993–1022
- Bródka P, Skibicki K, Kazienko P, Musiał K (2011, October) A degree centrality in multi-layered social network. In: 2011 international conference on computational aspects of social networks (CASON), pp. 237–242. IEEE.
- Bruns A, Moon B, Münch F, Sadkowsky T (2017) The Australian Twittersphere in 2016 Mapping the follower/followee network. *Soc Med Soc* 3(4):2056305117748162
- Bui QV, Sayadi K, Amor SB, Bui M (2017, April) Combining Latent Dirichlet Allocation and K-means for documents clustering: effect of probabilistic based distance measures. In: Asian conference on intelligent information and database systems, pp. 248–257. Springer, Cham.
- Casero-Ripollés A (2020) Introduction. Political influencers in the digital public sphere. *Commun Soc* 33(2):171–173
- Cassell J, Huffaker D, Tversky D, Ferriman K (2006) The language of online leadership: Gender and youth engagement on the internet. *Dev Psychol* 42(3):436
- Castells M (2007) Communication, power and counter-power in the network society. *Int J Commun* 1(1):29
- Cataldi M, Mittal N, Aufaure M-A (2013) Estimating domain-based user influence in Social networks. In: Paper presented at the proceedings of the 28th annual ACM symposium on applied computing.
- Chang HC (2010) A new perspective on twitter hashtag use: Diffusion of innovation theory. *Proc Am Soc Inf Sci Technol* 47(1):1–4
- Chen Y, Fay S, Wang Q (2011) The role of marketing in social media: How online consumer reviews evolve. *J Inter Marketing* 25(2):85–94
- Contreras JO, Hilles S, Abubakar ZB (2018) Automated essay scoring with ontology based on text mining and nltk tools. In: Paper presented at the 2018 international conference on smart computing and electronic enterprise (ICSCEE), pp. 1–6. IEEE.
- Daud A, Li J, Zhou L, Muhammad F (2010) Knowledge discovery through directed probabilistic topic models: a survey. *Front Comput Sci China* 4(2):280–301
- Day MY, Lee CC (2016) Deep learning for financial sentiment analysis on finance news providers. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp. 1127–1134. IEEE.
- Diani M (2003) Leaders or brokers? Positions and influence in social movement networks. In: Diani M, McAdam D (eds), *Social movements and networks: relational approaches to collective action*. Oxford University Press, Oxford, UK, pp 106–122
- Dubois E, Gaffney D (2014) The multiple facets of influence: Identifying political influentials and opinion leaders on twitter. *Am Behav Sci* 58(10):1260–1277
- Enke N, Borchers NS (2019) Social media influencers in strategic communication: A conceptual framework for strategic social media influencer communication. *Int J Strategic Commun* 13(4):261–277
- Feng Y (2016) Are you connected? Evaluating information cascades in online discussion about the # racetogether campaign. *Computers Human Behavior* 54:43–53
- Freeman LC (1978) Centrality in social networks conceptual clarification. *Soc Netw* 1(3):215–239
- Gan G, Ma C, Wu J (2007) Data clustering: theory, algorithms, and applications, similarity and dissimilarity measures, ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, ASA, Alexandria, VA, p67–106.
- Girolami, M., & Kabán, A. (2003, July). On an equivalence between PLSI and LDA. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 433–434).
- Grimaldi D (2019) Can we analyse political discourse using Twitter? Evidence from Spanish 2019 presidential election. *Soc Netw Anal Min* 9(1):49
- Guo L, Vargo CJ, Pan Z et al (2016) Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling[J]. *Journalism Mass Commun Quarterly* 93(2):332–359
- Hagen L, Uzuner Ö, Kotfila C, Harrison TM, Lamanna D (2015) Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. In: Paper presented at the 2015 48th Hawaii international conference on system sciences.
- Hajjem M, Latiri C (2017) Combining IR and LDA topic modeling for filtering microblogs. *Proc Comput Sci* 112:761–770
- Hamerly G, Elkan C (2004) Learning the k in k-means. *Adv Neural Inf Process Syst* 16:281–288
- Harrysson M (2016) Neural probabilistic topic modeling of short and messy text. (Master) Dissertation, KTH Royal Institute of Technology, Stockholm, Sweden.
- Haustein S, Bowman TD, Holmberg K, Peters I, Larivière V (2014) Astrophysicists on twitter: An in-depth analysis of tweeting and scientific publication behavior. *Aslib J Inf Manag* 66(3):279–296
- Hecht B, Hong L, Suh B, Chi EH (2011). Tweets from Justin: Bieber's heart. In: Proceedings of the 2011 annual conference on human factors In: computing systems-CHI'11, p 237.
- Hoffmann CP, Lutz C, Meckel M (2014, January) Impact factor 2.0: Applying social network analysis to scientific impact assessment.

- In: 2014 47th Hawaii international conference on system sciences, pp. 1576–1585. IEEE.
- Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2019) Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimed Tools Appl* 78(11):15169–15211
- Jenkins H, Carpentier N (2013) Theorizing participatory intensities: A conversation about participation and politics. *Convergence* 19(3):265–286
- Jiang Y, Xu Y, Guo J, Liu Y, Li R (2019, October) An intelligent question and answering system for dental healthcare. In: International conference on broadband communications, networks and systems, pp. 201–215. Springer, Cham.
- Jin Q, Waibel A (2000) Application of LDA to speaker recognition. In: Sixth international conference on spoken language processing.
- Jo W, Lee J, Park J, Kim Y (2020) Online information exchange and anxiety spread in the early stage of the novel coronavirus (COVID-19) outbreak in south korea: structural topic model and network analysis. *J Med Internet Res* 22(6):e19455
- Jungherr A, Jürgens P (2010) The political click: Political participation through e-petitions in germany. *Policy Internet* 2(4):131–165
- Kant G, Weisser C, Säfken B (2020) TTLocVis: A twitter topic location visualization package. *J Open Source Softw* 5(54):2507
- Karlsen R (2015) Followers are opinion leaders: The role of people in the flow of political communication on and beyond social networking sites. *Eur J Commun* 30(3):301–318
- Katz E, Lazarsfeld PF (1955) Personal influence: The part played by people in the flow of mass communications. Free Press, New York, pp 31–42
- Kavanaugh A, Fox EA, Sheetz S, Yang S, Li LT, Whalen T, Shoemaker D, Natsev P, Xie L (2011) Social media use by government: from the routine to the critical. In: Proceedings of the 12th annual international digital government research conference: digital government innovation in challenging times, College Park, Maryland, 12–15 June, 2011
- Keller E, Berry J (2003) One american in ten tells the other nine how to vote, where to eat and what to buy. They are the influentials. The Free Press, New York
- Kim SW, Gil JM (2019) Research paper classification systems based on TF-IDF and LDA schemes. *HCIS* 9(1):30
- Kim D, Seo D, Cho S, Kang P (2019) Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Inf Sci* 477:15–29
- Kim S, Sung KH, Ji Y, Xing C, Qu JG (2021) Online firestorms in social media: Comparative research between China Weibo and USA Twitter. *Public Relations Rev* 47(1):102010
- Kodinariya TM, Makwana PR (2013) Review on determining number of Cluster. *K-Means Clustering Int J* 1(6):90–95
- Kolli S, Khajeheian D (2020) How actors of social networks affect differently on the others? Addressing the critique of equal importance on actor-network theory by use of social network analysis. In: Cresswell KM, Worth A, Sheikh A (eds) Contemporary applications of actor network theory. Palgrave Macmillan, Singapore, pp 211–230.
- Krzanowski WJ, Lai YT (1988) A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 44(1):23–34
- Lassen DS, Brown AR (2011) Twitter: The electoral connection? *Soc Sci Comput Rev* 29(4):419–436
- Leskovec J (2011) Social media analytics: tracking, modeling and predicting the flow of information through networks. In: Proceedings of WWW (Companion Volume) 2011, pp 277–278
- Leydesdorff L (2007) Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *J Am Soc Inform SciTechnol* 58(9):1303–1319
- Li W, Feng Y, Li D, Yu Z (2016) Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm. *Autom Control Comput Sci* 50(4):271–277
- Li W, Guo K, Shi Y, Zhu L, Zheng Y (2018) DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. *Knowledge-Based Syst* 146:203–214
- Likas A, Vlassis N, Verbeek JJ (2003a) The global k-means clustering algorithm. *Pattern Recogn* 36(2):451–461
- Lim KW, Buntine W (2014) Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, pp. 1319–1328. New York, NY: ACM.;
- Lingras P, West C (2004) Interval set clustering of web users with rough k-means. *J Intell Inf Syst* 23(1):5–16
- Linville DL, Boatwright BC, Grant WJ, Warren PL (2019) “THE RUSSIANS ARE HACKING MY BRAIN!” investigating Russia’s internet research agency twitter tactics during the 2016 United States presidential campaign. *Comput Hum Behav* 99:292–300
- Martin T, Zhang X, Newman ME (2014) Localization and centrality in networks. *Phys Rev E* 90(5):052808
- Martinez-Lopez B, Perez AM, Sanchez-Vizcaino JM (2009) Social network analysis. Review of general concepts and use in preventive veterinary medicine. *Transbound Emerging Diseases* 56(4):109–120
- Marutho D, Handaka SH, Wijaya E (2018, September) The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In: 2018 international seminar on application for technology of information and communication, pp. 533–538. IEEE.
- McDermott P (2010) Building open government. *Gov Inf Q* 27(4):401–413
- Mehrotra R, Sanner S, Buntine W, Xie L (2013, July) Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 889–892.
- Mergel I (2012) The social media innovation challenge in the public sector. *Inf Polity* 17(3,4):281–292
- Mishler A, Crabb ES, Paletz S, Hefright B, Golonka E (2015, August) Using structural topic modeling to detect events and cluster Twitter users in the Ukrainian crisis. In: International conference on human-computer interaction, pp. 639–644. Springer, Cham
- Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplin Rev: Data Mining Knowl Discovery* 2(1):86–97
- Nagarajan M, Sheth A, Velmurugan S (2011) Citizen sensor data mining, social media analytics and development centric web applications. In: Proceedings of the 20th international conference companion on world wide web (WWW’11), pp 289–290.
- Nugroho R, Paris C, Nepal S, Yang J, Zhao W (2020) A survey of recent methods on deriving topics from Twitter: algorithm to evaluation. *Knowl Inf Syst* 62(7):2485–2519
- Pariser E (2011) The filter bubble: What the internet is hiding from you. Penguin, UK
- Park CS (2013) Does twitter motivate involvement in politics? Tweeting, opinion leadership, and political engagement. *Comput Human Behav* 29(4):1641–1648
- Pavel T (2009) The power of 140 characters: Twitter in the middle east. Tel Aviv Notes, July, 26.
- Peng X, Zhou C, Hepburn DM, Judd MD, Siew WH (2013) Application of K-Means method to pattern recognition in on-line cable partial discharge monitoring. *IEEE Trans Dielect Elect Insul* 20(3):754–761

- Perez C, Germon R (2016) Graph creation and analysis for linking actors: Application to social data. In: Layton R, Watters PA (eds) Automating open source intelligence. Elsevier, pp 103–129
- Porter J (2008) Designing for the social web, vol 1. New Riders Berkeley, CA
- Provan KG, Fish A, Sydow J (2007) Interorganizational networks at the network level: A review of the empirical literature on whole networks. *J Manag* 33(3):479–516
- Radovanović M, Ivanović M (2008) Text mining: Approaches and applications. *Novi Sad J Math* 38(3):227–234
- Rafeeq PC, Sendhilkumar S (2011, December) A survey on short text analysis in web. In: 2011 third international conference on advanced computing, pp. 365–371. IEEE.
- Rahmani A, Chen A, Sarhan A, Jida J, Rifaie M, Alhadj R (2014) Social media analysis and summarization for opinion mining: a business case study. *Soc Netw Anal Min* 4(1):1–11
- Ramos J (2003) Using tf-idf to determine word relevance in document queries. *Proc First Instruct Conf Mach Learn* 242(12):133–142
- Rauchfleisch A, Schäfer MS (2015) Multiple public spheres of weibo: A typology of forms and potentials of online public spheres in china. *Inf Commun Soc* 18(2):139–155
- Reddick CG, Chatfield AT, Ojo A (2017) A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government facebook use. *Gov Inf Q* 34(1):110–125
- Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for IDF. *J Documentation* 60(5):503–520
- Robinson R, Goh TT, Zhang R (2012) Textual factors in online product reviews: a foundation for a more influential approach to opinion mining. *Electron Commerce Res* 12(3):301–330
- Romero DM, Meeder B, Kleinberg J (2011, March) Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: Proceedings of the 20th international conference on World wide web, pp. 695–704.
- Rowley TJ (1997) Moving beyond dyadic ties: A network theory of stakeholder influences. *Acad Manag Rev* 22(4):887–910
- Salloum SA, Al-Emran M, Monem AA, Shaalan K (2017) A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J* 2(1):127–133
- Salvador S, Chan P (2004, November) Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In: 16th IEEE international conference on tools with artificial intelligence, pp. 576–584. IEEE.
- Scott M (2015) Re-theorizing social network analysis and environmental governance: Insights from human geography. *Prog Hum Geogr* 39(4):449–463
- Shmargad Y (2018) Twitter Influencers in the 2016 US Congressional Races. *J Political Mark.* <https://doi.org/10.1080/15377857.2018.1513385>.
- Smith AN, Fischer E, Yongjian C (2012) How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? *J Interact Mark* 26(2):102–113
- Song M, Kim MC, Jeong YK (2014) Analyzing the political landscape of 2012 korean presidential election in twitter. *IEEE Intell Syst* 29(2):18–26
- Srinivas A, Velusamy RL (2015, June) Identification of influential nodes from social networks based on Enhanced Degree Centrality Measure. In: 2015 IEEE international advance computing conference (IACC), pp. 1179–1184. IEEE.
- Stieglitz S, Brockmann T, Dang-Xuan L (2012, July) Usage of social media for political communication. In: PACIS (p. 22).
- Sugar CA, James GM (2003) Finding the number of clusters in a dataset: An information-theoretic approach. *J Am Stat Assoc* 98(463):750–763
- Tang J, Li H, Cao Y, Tang Z (2005, August) Email data cleaning. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 489–498
- Tu E, Cao L, Yang J, Kasabov N (2014) A novel graph-based k-means for nonlinear manifold clustering and representative selection. *Neurocomputing* 143:109–122
- Tumasjan A, Sprenger TO, Sandner PG, Welpke IM (2011) Election forecasts with twitter: How 140 characters reflect the political landscape. *Soc Sci Comput Rev* 29(4):402–418
- Vaccari C, Valeriani A, Barberá P, Bonneau R, Jost JT, Nagler J, Tucker J (2013) Social media and political communication. A survey of twitter users during the 2013 italian general election. *Rivista italiana di scienza politica* 43(3):381–410
- Wagner TF, Baccarella CV, Voigt KI (2017) Framing social media communication: Investigating the effects of brand post appeals on user interaction. *Eur Manag J* 35(5):606–616
- Wang Z (2018) Extracting latent topics from user reviews using online LDA. In: 2018 international conference on information technology and management engineering (ICITME 2018, August). Atlantis Press.
- Wang X, McCallum A (2006, August) Topics over time: a non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 424–433.
- Warden CA, Chen JF, Caskey DA (2005) Cultural values and communication online: Chinese and southeast asian students in a taiwan international mba class. *Business Commun Quarterly* 68(2):222–232
- Wasserman S, Faust K (1994) Social network analysis: Methods and applications, vol 8. Cambridge University Press, Cambridge
- Welser HT, Gleave E, Fisher D, Smith M (2007) Visualizing the signatures of social roles in online discussion groups. *J Soc Struct* 8(2):1–32
- Weng J, Lim EP, Jiang J, He Q (2010) TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM WSDM.
- Williamson A, Phillips LN (2009). Twitter: Communication tool or pointless vanity? Hansard Society.
- Wu X, Ramesh M (2014) Market imperfections, government imperfections, and policy mixes: Policy innovations in singapore. *Policy Sci* 47(3):305–320
- Wu MY, Shen CY, Wang ET, Chen AL (2020) A deep architecture for depression detection using posting, behavior, and living environment data. *J Intell Inf Syst* 54(2):225–244
- Yadav MP, Feeroz M, Yadav VK (2012, July) Mining the customer behavior using web usage mining in e-commerce. In: 2012 third international conference on computing, communication and networking technologies (ICCCNT'12), pp. 1–5. IEEE.
- Yang J, Counts S (2010) Predicting the speed, scale, and range of information diffusion in twitter. In: Paper presented at the fourth international AAAI conference on weblogs and social media.
- Yang X, Wang Y, Wu D, Ma A (2010, November) K-means based clustering on mobile usage for social network analysis purpose. In: 2010 6th international conference on advanced information management and service (IMS), pp 223–228. IEEE.
- Yaqub U, Chun SA, Atluri V, Vaidya J (2017) Analysis of political discourse on twitter in the context of the 2016 US presidential elections. *Govern Inf Quarterly* 34(4):613–626
- Yin SL, Liu J (2016) A K-means approach for mapreduce model and social network privacy protection. *J Inf Hiding Multimed Signal Process* 7(6):1215–1221
- Yoo Y, Alavi M (2004) Emergent leadership in virtual teams: What do emergent leaders do? *Inf Organiz* 14(1):27–58
- Žalik KR (2008) An efficient k'-means clustering algorithm. *Pattern Recogn Lett* 29(9):1385–1391

- Zeng D, Chen H, Lusch R, Li S (2010) Social media analytics and intelligence. *IEEE Intell Syst* 25(6):13–16
- Zhang W, Yoshida T, Tang X (2011) A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Syst Appl* 38(3):2758–2765
- Zhang Y, Mańdziuk J, Quek CH, Goh BW (2017) Curvature-based method for determining the number of clusters. *Inf Sci* 415:414–428
- Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, Zou W (2015) A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics* 16(13):1–10
- Zhu M, Zhang X, Wang H (2016, December) A LDA based model for topic evolution: Evidence from information science journals. In: 2016 international conference on modeling, simulation and optimization technologies and applications (MSOTA2016), pp. 49–54. Atlantis Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.