BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 02/06 pp259–279
Volume 3, Number 4, December 2020
DOI: 10.26599/BDMA.2020.9020006

Survey on Data Analysis in Social Media: A Practical Application Aspect

Qixuan Hou, Meng Han*, and Zhipeng Cai*

Abstract: Social media has more than three billion users sharing events, comments, and feelings throughout the world. It serves as a critical information source with large volumes, high velocity, and a wide variety of data. The previous studies on information spreading, relationship analyzing, and individual modeling, etc., have been heavily conducted to explore the tremendous social and commercial values of social media data. This survey studies the previous literature and the existing applications from a practical perspective. We outline a commonly used pipeline in building social media-based applications and focus on discussing available analysis techniques, such as topic analysis, time series analysis, sentiment analysis, and network analysis. After that, we present the impacts of such applications in three different areas, including disaster management, healthcare, and business. Finally, we list existing challenges and suggest promising future research directions in terms of data privacy, 5G wireless network, and multilingual support.

Key words: social media; topic analysis; time series analysis; sentiment analysis; network analysis; disaster management; bio-surveillance; business intelligence

1 Introduction

The rise of social media has made it faster and easier to share and access information. Social media is serving as a tool in raising awareness, especially during emergencies. When the Amazon rainforests had alarming clusters of burning wildfires, the first post about the wildfires on Twitter was published on August 6, 2019, two weeks before the cable news reported the incident. Many studies are focusing on extracting abnormal events, such as disease outbreaks, natural disasters, and infrastructure breakdowns, from social media data. Additionally, social media connects individuals and helps break the barriers of communication. When earthquake and tsunami hit Japan in March 2011, millions of people around the globe used social media to search for family and friends. Furthermore, social media provides an opportunity for everyone to tell his/her stories and share his/her opinions. The businesses apply text mining and sentiment analysis on social media data to understand the public's opinions and comments on their products or services so that they can better understand the markets and prepare for the future.

Social media can be categorized based on what kind of content it generates. As shown in Fig. 1, wiki, like Wikipedia, allows collaborative editing; blogging, like WordPress, regularly posts articles about events and topics; microblogging, like Twitter, is a shorter version of blogging; opinion and reviews, such as Yelp, share reviews of restaurants or services; question answering, like StackOverflow, serves as platforms for users to

• Qixuan Hou is with the College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA. E-mail: qhou6@gatech.edu.
• Meng Han is with the Data-driven Intelligence Research (DIR) Lab of College of Computing and Software Engineering, Kennesaw State University, Kennesaw, GA 30060, USA. E-mail: mhan9@kennesaw.edu.
• Zhipeng Cai is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA. E-mail: zcai@gsu.edu.
* To whom correspondence should be addressed.
Manuscript received: 2020-05-21; accepted: 2020-06-08

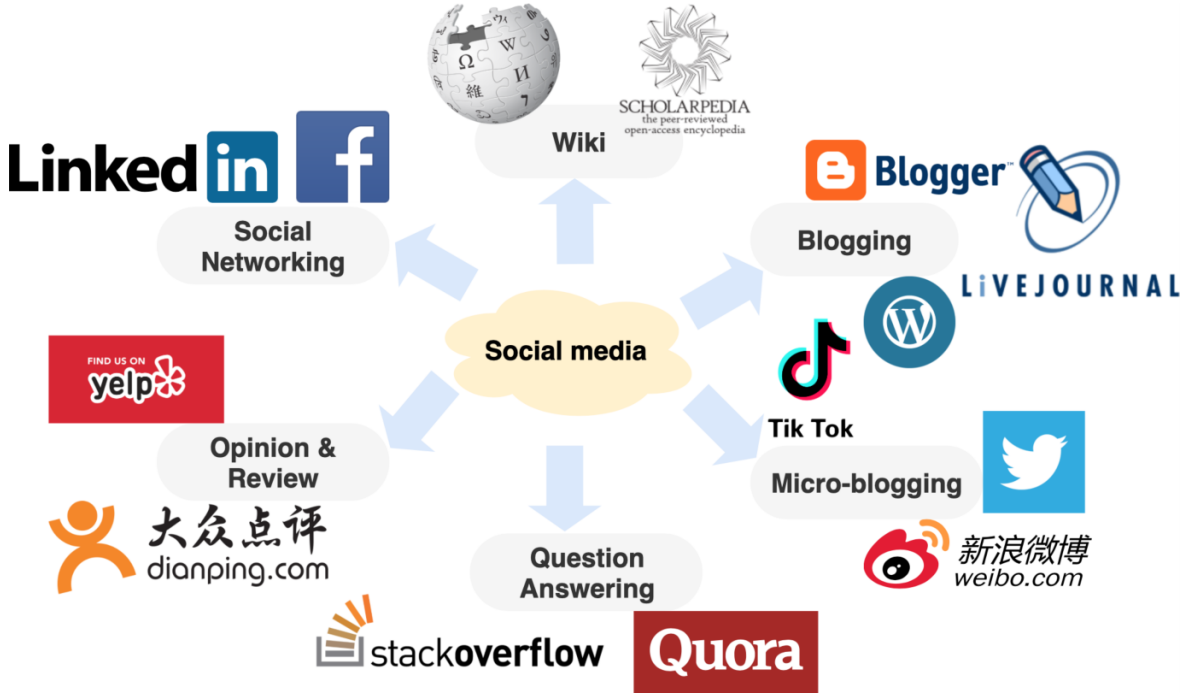


Fig. 1 Social media's categories and their examples.

exchange questions and answers; social networking, like LinkedIn and Facebook, connects users.

The variety of available social media platforms makes it challenging to offer one accurate and exact definition of "social media". In 2010, Kaplan and Haenlein[1] defined "social media" as a group of Internet-based applications that build on the ideological and technological foundations of Web2.0 and that allow the creation and exchange of User Generated Content. In 2012, Howard and Parks[2] commented that "social media" consist three parts: (1) the information infrastructure and tools used to produce and distribute content; (2) the content that takes the digital form of personal messages, news, ideas, and cultural products; and (3) the people, organizations, and industries that produce and consume digital content. In 2015, Carr and Hayes[3] defined "social media" as Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others. The previously proposed definitions all agree that "social media" allows users to generate content and create online communities to share user-generated content.

The number of users on social media platforms is huge. Domo's Data Never Sleeps 5.0 report shows every minute of the day in 2017, there were 456 000 tweets sent on Twitter, there were 46 740 photos posted on Instagram, and there were 527 760 photos shared on Snapchat[4]. According to Our World in Data, there are 7.6 billion people in the world, with at least 4 billion of us online and 3 billion of us using social media, which means social media platforms are used by one-in-three people in the world, as shown in Fig. 2. There has been a dramatic usage increase since 2006. Figure 3 presents the history of social media usage. In 2019, there were more than 2 billion Facebook users, which means one-in-four people in the world were using Facebook[5].

With the huge number of users and the gigantic amount of content generated by the users, social media is regarded as a valuable data source in both academia and industry. A large number of applications have been built, which consume and analyze social media data, to provide values and insights in diverse fields. However,

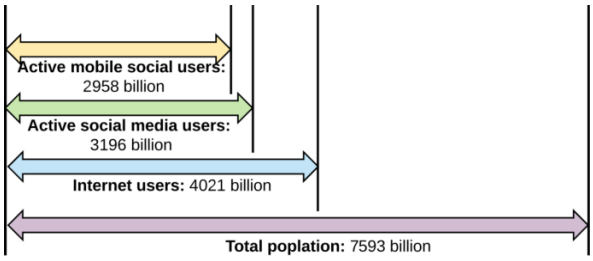


Fig. 2 Number of social media users in the world.

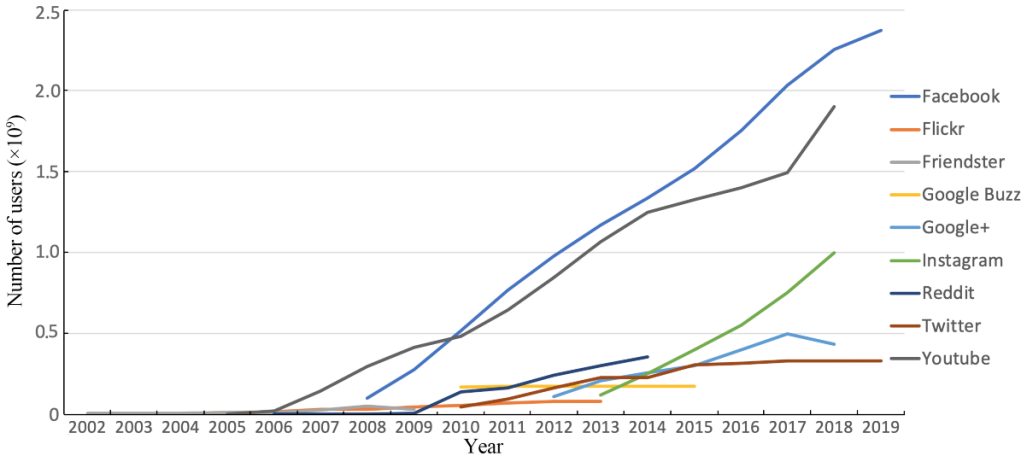


Fig. 3 Number of users using social media platforms from 2002 to 2019.

challenges exist, due to the four major phenomena of social media: volume, velocity, variety, and veracity. Different from a traditional data source which typically has a centralized content-generating agent, social media allows users to generate content themselves. Because of the large number of users, social media produces data with large volume, high velocity, and a wide variety. Additionally, due to the lack of quality controls, the information flowing on social media cannot receive the same credibility and integrity as the official news sources.

Researchers have been exploring analysis methods to overcome the four challenges in designing and building social media-based applications, as shown in Fig. 4.

The remaining parts of this paper are organized as follows. We summarize a commonly used pipeline in building such applications in Section 2. Section 3 presents four major analysis techniques used to extract values and insights from social media data. In Section 4, we discuss applications' impacts in three fields: healthcare, disaster management, and business. Additional challenges and potential future opportunities are presented in Section 5. Section 6 concludes the survey.

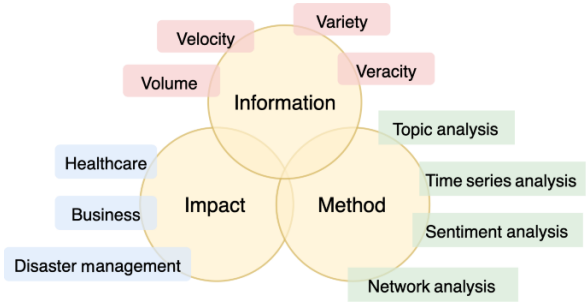


Fig. 4 Four challenges, four analysis techniques, and three impacts of social media-based applications.

The survey focuses on the practical aspects of building such applications. Below, we summarize our major contributions.

- We systematically compare the existing applications and categorize those based on their analysis techniques and impacting areas, hoping to provide a comprehensive and in-depth survey of social media-based applications.
- We outline a four-stage pipeline which is commonly used in building such applications. Four stages are data collection, data storage, data analysis, and data visualization, as shown in Fig. 5.
- We demonstrate the four most popular analysis techniques, which are topic analysis, time series analysis, sentiment analysis, and network analysis. Important solutions and major applications of each type are presented with examples from the previous literature.
- We not only investigate a more extensive set of previous work but also present a summary of the insights behind each technique. A thorough review of available

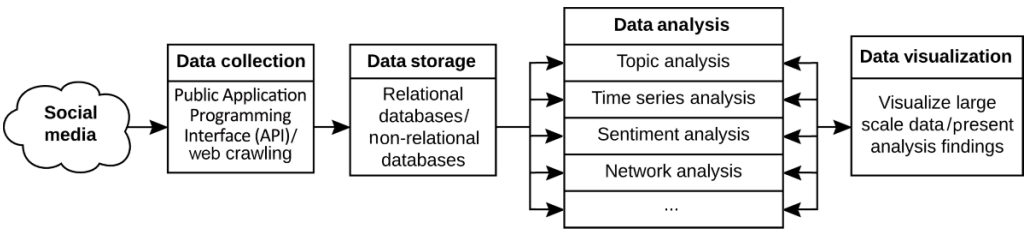


Fig. 5 Common application pipeline.

analysis techniques can serve as a guide book for future research.

• We cover three areas that have been significantly impacted by social media-based applications. We hope to explore such applications and their impacts more broadly, thereby bringing new perspectives to future researchers.

• Developing social media-based applications without compromising the rights of data subjects can be challenging. We would like to raise awareness of privacy issues and promote best practices of data protection.

• We outlook the future of social media in the context of the rapid development of 5G wireless network technology. We suggest that video and image processing are becoming a crucial building block of social media-based applications.

• We suggest promising future research opportunities in terms of solutions and their applications. For each direction, we discuss its disadvantages in current work and propose future research directions.

2 Common Application Pipeline

After reviewing more than a hundred of social media-based applications, the common application pipeline is summarized as follows: (1) collecting data; (2) cleaning and storing data; (3) analyzing data; and (4) visualizing results, as presented in Fig. 5.

2.1 Data collection

Most social media platforms, such as Twitter and Facebook, provide robust official API for developers to collect data. Some API, like the Twitter Streaming API, even enables streaming real-time data with concurrency, low latency, and high throughput. It is not a challenge to collect a large number of data from social media platforms. With real-time streaming service, it is also achievable to stream live data for real-time event detection.

Web crawling is used to gather data when public API is not available. Web crawling, also known as web scraping or spider, is a computer program used to scrape information from websites. However, web crawling might have a negative connotation when scraping private or classified information, disregarding websites' privacy policy and terms of service, or scraping without owners' permission. It is important to check policies and regulations before crawling.

With the implementation of privacy regulations, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), data privacy is becoming a concern when collecting an immense amount of data from social media platforms. Section 5.1 focuses on the impacts of privacy regulations on social media-based applications.

2.2 Data storage

Deliberate considerations are needed to determine appropriate technology to store and process large datasets. Relational databases (such as Microsoft SQL Server, MySQL) and non-relational databases (such as Cassandra, Redis) are commonly used to store social media data in current research. Despite the proliferation research of NoSQL-based databases, there is still a drawback that NoSQL-based solutions do not usually allow queries. To solve the problem, some applications add an in-memory caching layer using Redis. The survey does not focus on studying data storage solutions but extensive studies in this field can be found as follows. Stieglitz et al.[6] shared available solutions for data storage. Moalla et al.[7] provided a thorough review of data warehouse design for social media data. Cao et al.[8] discussed data warehousing and online analytical processing technology in the field of business intelligence.

2.3 Data analysis

A major class of applications that consume social media data is dedicated to detecting events using topic analysis technology. The extracted events range from natural disasters[9, 10] to disease outbreaks[11, 12]. Hashtags and searching queries are topics of interest in detecting events. Besides, topic modeling, such as Latent Dirichlet Allocation (LDA), is used to discover abstract topics in a collection of documents.

Some applications are designed to study time series information on social media[13]. The information retrieved from social media can be organized in chronological order according to its timestamp. Time series data provide a good resource for the study of past behaviors, the analysis of current trends, the identification of anomalies or bursts, the forecast of future, and the analysis of seasonal variations.

Social media provides a rich source of user-generated content. Sentiment analysis has been popular especially in the business and marketing fields, because countless people share their thoughts and feelings about products, events, and news on social media.

Some applications investigate the networks in the virtual community of social media. They would like

to identify influencers or to study information diffusion patterns. Such applications analyze the relationships between users, and such relationships include "follow", "like", and "repost".

Data analysis is the most interesting block in the pipeline. In social media-based applications, we identify four leading analytical techniques: topic analysis, time series analysis, sentiment analysis, and network analysis. Section 3 discusses the techniques and their applications in detail.

2.4 Data visualization

Data visualization is to present information in a graph, chart, or other visual formats. Well designed data visualization better communicates relationships, trends, and patterns in data. As social media data are always in a large scale and it is challenging to interpret increasingly large batches of data, data visualization has become an indispensable tool. Applications are providing innovative visualizations to better present the information from social media. By combining ThemeRiver with storyline style visualization, Xu et al.[14] designed a visualization that includes both topical and social aspects of the information diffusion process. Dou et al.[15] proposed an interactive visual analysis system, LeadLine, to automatically identify meaningful events in the news and social media data. The system also supports event exploration. Chen et al.[16] presented an interactive visual analysis system to support the exploration of sparsely sampled trajectory data from social media, hoping to provide rich text information and motion information at the same time to facilitate the understanding of people's movements. Marcus et al.[17] showed TwitInfo, which allows users to browse a large collection of tweets using a timeline-based display that highlights the peak of high tweet activity. When users drill down to sub-events, they can explore more information about geographic location, sentiment, and commonly used Uniform Resource Locators (URLs).

Data visualization is not only a tool to help researchers better understand information during data analysis, but also serves as a demonstration to communicate the analysis results to the audience.

3 Primary Analytic Technique

Summary of previous work in topic analysis, time series analysis, sentiment analysis, and network analysis can be found in Tables 1–4.

3.1 Topic analysis

Social media has become an important news source. Evan Williams, a founder of Twitter, defined the service as "What we have to do is deliver to people the best and freshest most relevant information possible. We think of Twitter as it is not a social network, but it is an information network. It tells people what they care about as it is happening in the world".[63] Following this strategy, around 2010, Twitter changed the question, which asks users for status updates, from "What are you doing?" to "What's happening?" Cataldi et al.[18] mentioned that "Twitter defines a low-level information news flashes portal". Even if it cannot be considered

Table 1 Summary of previous works in topic analysis.

Year	Reference	Method	Application
2010	Cataldi et al.[18]	Queries	Create a navigable topic graph with emerging topics over time
2010	Sizov[19]	Topic modeling	GeoFolk to discover latent topics
2010	Song et al.[20]	Topic modeling	Explore spatio-temporal framework for related topic search
2012	Han and Kang[21]	Queries	Identify the personalized relevance of social issues to targets
2012	Song et al.[22]	Topic modeling	Identify the personalized relevance of social issues to targets
2012	Hu et al.[23]	Topic modeling	Propose a topic modeling with user features
2013	Kamath et al.[24]	Hashtags	Study the spatio-temporal dynamics of Twitter hashtags
2013	Ma et al.[25]	Topic modeling	Propose Tag-Latent Dirichlet Allocation (TLDA) to bridge hash tags and topics
2013	Bogdanov et al.[26]	Topic modeling	Identify the personalized relevance of social issues to targets
2015	Jang and Myaeng[27]	Topic modeling	Analyze spatially oriented topic versatility
2015	Yao et al.[28]	Topic modeling	Analyze news trends in Twitter
2016	Qian et al.[29]	Topic modeling	Analyze multi-model event topic model
2016	Musaev and Hou[30]	Queries	Detect landslides with Twitter data
2016	Rohani et al.[31]	Topic modeling	Explore an unsupervised topic modeling approach
2018	Argyrou et al.[32]	Hashtags	Prepare training images for Automatic Image Annotation (AIA)
2018	Ejaz et al.[33]	Topic modeling	Analyze news using ontology

Table 2 Summary of previous works in time series analysis.

Year	Reference	Method	Application
2007	Fu et al.[34]	Visualization tool	Visualize consumer-generated media
2010	Mathioudakis and Koudas[35]	Visualization tool	TwitterMonitor: Trend detection
2011	Yang and Leskoves[36]	Clustering	Detect patterns of temporal variation
2012	Zhao et al.[37]	Smoothing	Identify event-related bursts
2012	Chae et al.[38]	Seasonal-trend decomposition	Detect abnormal event
2014	Ahn and Spangler[39]	Prediction	Predict sales
2014	Kucher et al.[40]	Visualization tool	Visualize stance markers in online social media
2015	Healy et al.[41]	Peak detection	Detect events
2015	Tsuboi et al.[42]	Prediction	Predict product purchases
2015	Nusratullah et al.[43]	Analyze variance	Detect changes in email notwork
2015	Johnson and Ni[44]	Prediction	Infer dynamic consumer valuations
2015	Zhao et al.[45]	Prediction	Predict the reposting number of micro-blog messages
2017	Ni et al.[46]	Prediction	Forecast the subway passenger flow
2017	Dahouei[47]	Smoothing	Identify hot topics
2017	Comito et al.[48]	Peak detection	Identify deviation from user normal behavior

Table 3 Summary of previous works in sentiment analysis.

Year	Reference	Method	Application
2012	Rui and Whinston[49]	Learning-based	Design business intelligence system
2013	Yuan et al.[50]	Lexicon-based & learning-based	Sentiment classify Chinese micorblogs
2013	Yu et al.[51]	Learning-based	Analyze firm equity value
2014	Wang et al.[52]	Learning-based	Enhance sentiment analysis technique
2016	Wang et al.[53]	Lexicon-based	Build finer-grained sentiment analysis

Table 4 Summary of previous works in network analysis.

Year	Reference	Method	Application
2003	Popescul and Ungar[54]	Classification	Predict links
2007	Noweell and Kleinberg[55]	Classification	Predict links
2010	Cataldi et al.[18]	Page rank	Detect emerging topics
2011	Rossetti et al.[56]	Classification	Predict links
2012	Michalski et al.[57]	Time series	Predict links
2013	Smith[58]	Visualization	NodeXl: Visualizae network
2016	Musaev and Hou[30]	Page rank	Detect landslides with Twitter data
2016	Erlandsson et al.[59]	Association rule learning	Identify influencers
2016	Li et al.[60]	Location-based social networks	Identify influencers
2017	Zhao et al.[61]	Degree/betweenness/closeness centrality	Identify influencers
2019	Tang et al.[62]	The second-order Independent Cascade (IC) model	Identify influencers

as an alternative to the authoritative news source, social media can provide real-time information that sometimes predates newspapers to inform the online community of emerging topics. Authoritative information media always requires a certain amount of time to report an incident, but the public can easily post what is happening on Twitter. Therefore, researchers are interested in developing methods to extract meaningful topics from millions of data points. In the previous work, three major solutions are proposed.

Hashtags. A hashtag is defined as a word or a short term prefixed with the symbol "#". It is widely used in social media including Twitter and Instagram. Hashtags are regarded as important metadata to categorize posts and to propagate ideas and topics. The use of hashtags has become an inseparable part of social media. Kamath et al.[24] and Argyrou et al.[32] used hashtags as the main source for identifying topics from social media.

Queries. Searching queries are another important source for topic detection, as discussed in Refs. [18, 21]. For instance, searching on Google with the keyword "Georgia Tech" provides us a list of articles about

Georgia Tech, so "Georgia Tech" can be regarded as the topic for the set of articles. Many event detection platforms collect social media data with searching queries. Musaev and Hou[30] built a landslide detection system with Twitter data. To feed the system, tweets are collected with queries, "landslides" and "mudslides". However, noises exist as searching keywords might have multiple meanings. "Landslides" can also be used to describe an election in which the victor wins by an overwhelming margin. Additional technologies, such as classification models, are needed to further improve the accuracy of topic detection in such a scenario.

Topic modeling. A topic model is defined as a type of model for discovering abstract topics in a collection of documents. Topic modeling is frequently used as a text-mining tool to discover hidden semantic structures in a text body. LDA is a popular topic modeling technique. In the hope to improve the accuracy, researchers are extending LDA by incorporating additional information, such as hashtags, user profiles, and location information[31]. Ma et al.[25] proposed TLDA to bridge hashtags and topics. TLDA extends LDA by adding observed hashtags and mapping hashtags to a mix of shared topics. To discover latent topics, GeoFolk integrates topic-specific normal distributions that describe the location (latitude and longitude). GeoFolk also adds sampling-based parameter estimation based on the Monte Carlo Markov Chain method (MCMC)[19]. Previous works utilized spatial information to improve topic extraction because they believe the relevance and similarity of resources can be directly estimated by the distance between resource locations[20,24,27]. Others took user features and social networks into account to improve the accuracy of topic models[18,21–23].

Topic analysis generates values in diverse aspects. Topic versatility captures its characteristics across different regions. Identifying versatile topics can help understand regional interests and thus provide region-dependent services. To understand the events and their evolutionary trends, Cataldi et al.[18] created a navigable topic graph to present a set of emerging topics over time. Qian et al.[29] proposed a novel multi-modal Event Topic Model (mmETM) to obtain information on social events over time. News collected from different sources can be compared to identify potential biases and to find different focal points between news agencies and social media[28,33]. The extracted topics can also be used to identify the personalized relevance of social issues to targets[21,22,26]. AIA is a process of assigning tags to digital images without human intervention. Most of the approaches require large training examples. Argyrou et al.[32] applied topic modeling, LDA, on Instagram hashtags to obtain the subject of images and then utilized the images with the extracted topics to train AIA methods.

3.2 Time series analysis

A time series is a series of data points indexed in chronological order. Time series carries profound importance in studying the past behavior, comparing current trends with the past, identifying anomalies or bursts, forecasting future trends, analyzing seasonal variations, etc. Time series analysis has become another important analysis technique in social media research and two characteristics of social media make it more interesting and challenging[37]. Firstly, social media involves various types of events that may be temporally correlated with one another, but they may also capture different aspects of the event. Secondly, there are irregular, unpredictable, and spurious noises in social media data.

Three steps of time series analysis are shown as follows[43]:

Step one: *data selection and preprocessing.* Data are filtered based on research questions. Tsuboi et al.[42] predicted the purchases of digital cameras and personal computers, so they collected tweets written by users who explicitly admitted they purchased the products. Comito et al.[48] wanted to capture relevant events in a geographic area, so they collected data from one location.

Step two: *data extraction and transformation.* Time series analysis requires data points arranged in chronological order. Each data point can be a frequency of certain events, texts, or locations. Social media data carry a lot of information, including hashtags, geographic information, texts, and user profiles. Interesting features can be extracted from the data, such as frequency of hashtags, user connections, and reposting behaviors. Nusratullah et al.[43] studied email networks, so they extracted the frequencies of communication between accounts and used them as features for further analysis. Zhao et al.[37] and Dahouei[47] identified hot topics from social media and they used the frequencies of hashtags as features. Chae et al.[38] utilized LDA to extract and rank major topics and the ranking of topics was used in time series analysis.

Step three: *time series analysis.* Based on research purposes, different techniques are applied to the extracted features.

Exploratory analysis is a method that usually uses visual methods to summarize the main features of the dataset. Some researches design and build tools to help others explore time series data. TwitterMonitor groups information based on keywords and visualizes the trends[35]. Kucher et al.[40] presented a visual analytic tool to investigate stance phenomena on time series data from social media.

Smoothing techniques are usually used to smooth irregular roughness to see a clearer signal in time series. Dahouei[47] analyzed the distribution of hashtags from hot topics and attempted to find patterns beneath the distribution. In other words, they wanted to smooth the distribution by identifying the critical points. Zhao et al.[37] were interested in the local state smoothness and correlation across multiple streams to detect event bursts. Fu et al.[34] employed the Perceptual Import Point (PIP) technique to extract only the useful data and to smooth time series trends.

Clustering with time series is to partition data into groups based on similarity or distance, so that time series patterns in the same cluster are similar. Yang and Leskovec[36] studied temporal patterns associated with online content and how the content's popularity grows and fades over time. They tackled the problem with a time series clustering technique, the K-Spectral Centroid (K-SC) clustering algorithm. They found most press agency news exhibit a very rapid rise followed by relatively slow decay, and whereas, bloggers play a very important role in determining the longevity of news on the web. The results can be used for better placing content to maximize click-through rates and also for finding influential blogs and tweets.

Peak detection is to identify sudden surges, which are inherently interesting, as they suggest the occurrence of an influential event in social media. In a business context, events that generate enough interests to result in flurries of social media activities are meaningful to business analysts. For social goods, surges may be natural disasters, disease outbreaks, or infrastructure breakdowns, and peak detection can be used to realize emergency detection. There are a number of peak detection techniques, including Palshikar's peak detection algorithm[64], Du et al.'s continuous wavelet transform algorithm[65], and Lehmann et al.'s peak detection algorithm[66]. Healy et al.[41] evaluated the algorithms for social media event detection. Comito et al.[48] extracted space-time features from social media and applied peak detection to identify deviation from user normal behavior, in the hope to uncover meaningful events.

Prediction is to forecast the future based on historical observations. Zhao et al.[45] predicted the reposting counts of micro-blog messages on Sina Weibo using curve fitting optimized by empirical model. Ni et al.[46] forecast the subway passenger flow with social media data. Prediction with social media data is also popular in the business context. A large number of traditional time-series models to predict future sales of a product sometimes provide unpromising results, because they have majorly relied on the historical sales and ignored the dynamic impact of recent events. Customers' current reviews and comments about companies and products would largely affect buying behaviors[67]. Social media is a good source for such information. Therefore, Ahn and Spangler[39] proposed to combine social media prediction and existing analytical forecasting models. Tsuboi et al.[42], Johnson and Ni[44], and Desai and Han[68] emphasized on extracting sentiment from social media for sales prediction.

3.3 Sentiment analysis

Social media is one of the richest sources of publicly generated text. Sentiment analysis has become an important area of work to understand the feelings and emotions of certain groups because so many people share their thoughts and feelings towards products, events, and news on social media. Business owners would like to study how their products or service make people feel; consumers refer to online reviews when making purchase decisions; politicians gauge public reactions to their policies; short-term stock prices are influenced by recent public opinion on companies; and the revenue of movies, concerts, and performances are driven by audience reviews. Rui and Whinston[49] used real-time data from Twitter to build Business Intelligence (BI) systems to forecast sales and revenues. Yu et al.[51] investigated the effect of social media on short term firm stock market performances by analyzing sentiment in rich texts.

Sentiment analysis aims to determine sentiment polarity. Most existing analysis techniques focus on differentiating between positive and negative feelings, or some might have neutral as an additional category. Wang et al.[53] and He et al.[69] attempted to achieve

finer-grained sentiment analysis, which can yield more specific and more actionable results with detailed negative emotion subcategories, such as anger, sadness and anxiety, and positive emotion subcategories, such as happiness and excitements.

There are two major approaches discussed in the previous literature.

Lexicon-based approaches are popular in the early stages. They identify the dominant polarity of text by searching for emotion indicators based on lexicons. The performance heavily depends on the quality of sentiment lexicons. Furthermore, accuracy is limited due to semantic ambiguity. The widely-used English sentiment lexicons include Linguistic Inquiry and Word Count (LIWC)[70], General Inquirer Lexicon[71], and SentiWordNet1[72]. To analyze other languages, sentiment lexicons in the corresponding languages have to be built, or translating targeted texts into English or translating English lexicons into corresponding languages is needed. Yuan et al.[50] used Chinese Emotion Word Ontology to conduct a Simple Sentiment Word Count Method (SSWCM) on Weibo, a Chinese social media platform.

Learning-based approaches, including supervised and unsupervised learning approaches, have become very successful in the field. The approaches derive sentiment polarity from the relationship between the features of the targeted text. Earlier works focus on bag-of-words and linguistic features. Later works start exploring metadata, including reposts, hashtags, likes, and counts. The main blockers of building machine learning models are learning speed and effectiveness. It is difficult to train a model with too many features on a huge training dataset. Therefore, feature selection is the key of building effective sentiment classification models. Wang et al.[52] proposed to use Chi-Square feature selection to select significant features and remove irrelevant, redundant, and noisy data to improve the performance. They also discussed how to handle negation and emoticon to further improve the accuracy of machine learning-based approaches.

Perikos and Hatzilygeroudis[73] introduced the ensemble classifier schema that combines lexicon-based and learning-based approaches to analyze large social data and model emotions in social media. The lexicon-based model does not require training and it can be used directly and effectively for large amounts of texts. The learning-based model can further boost accuracy. The goal of the ensemble classifier schema is to efficiently leverage the advantages of the base classifiers, aiming to increase overall efficiency and accuracy.

3.4 Network analysis

Social media creates a virtual community where users are grouped according to their preferences. Information is transmitted within and between groups. Previous applications are dedicating to studying the relationships in this virtual community and to understanding how information flows and who contributes to the dissemination of information. There are five common types of user interactions on social media: (1) being tagged on a post; (2) commenting on a post; (3) sharing a post; (4) liking a post; and (5) following others. Generally, in social network analysis, we present participating users as vertices or nodes, and we define edges or links as connections between users. Some studies consider direction, while others do not. Most of them assume that any link between two vertices indicates their similarity. The three main purposes of social network analysis are described below.

Identify influencers. Generally, influencers are individuals or organizations that have established credibility in a specific industry. Social media influencers can reach a large audience and convince others with their authenticity and influence. Influencers are rare in social networks, but their influence can quickly spread to most nodes in the network. Identifying influencers allows us to accelerate information propagation, conduct successful e-commerce advertisements, etc. Page Rank is a popular algorithm used to calculate users' influences. Musaev and Hou[30] and Han et al.[74] considered user influence to improve landslide detection with Twitter data. They introduced the concepts of relevant and irrelevant virtual communities based on whether the posted messages are relevant to landslides or not and then applied the Page Rank algorithm to identify influential nodes in each community. Cataldi et al.[18] used Page Rank to calculate users' authority to better detect emerging topics on Twitter. From graph perspectives, various centrality measures have been presented to identify vital nodes, such as Degree Centrality, Betweenness Centrality, and Closeness Centrality. Additionally, there are new algorithms proposed to further improve the measurements by considering both the topological connections among neighbors and the number of neighbors[61,75]. The IC model is a widely used information diffusion model. Tang et al.[62] obtained influential nodes by extending

the IC model to the second-order IC model. The second-order IC model considers the influences of direct neighbors and also takes the previous influence into consideration. Erlandsson et al.[59] proposed to use association rule learning to detect the relationships between users. Association rule learning is a machine learning technique that aims to find out how one item affects another by analyzing how frequently certain items appear together. Li et al.[60] used data from the Location-Based Social Networks (LBSNs) and proposed a novel network model and an influence propagation model, which study influence propagation in both online social networks and the physical world.

Predict links. Link prediction in social networks is defined as the task of predicting whether a link between two specific nodes will appear in the future. Link prediction can provide insights regarding how the network will grow or collapse and how many connected components will exist in the network. A graph can be expressed as an adjacency matrix. Link prediction can be viewed as a binary classification problem, classifying the elements in the adjacency matrix as zero (linked) or one (linked). To build a classification model, previous work has explored different features, including graph structural properties[55,56] and attributes of the nodes[54]. However, due to the large scale of social networks, the classification approach is time-consuming. Michalski et al.[57] proposed an efficient solution of predicting key network measure values with time series forecasting.

Visualize networks. Network analysis is important and powerful, but it is also complicated. Analysis of large networks requires specialized domain knowledge. Some researches focus on developing tools to help non-programmers understand and analyze complex networks. They hope to enable more extensive network analysis research and help researchers analyze the networks with less effort. NodeXL is such a tool that allows users to explore networks and to discover insights with an easy-to-use interface[58].

4 Application Impact

Social media-based applications have a wide range of impacts. Our survey focuses on their influences in the fields of healthcare, disaster management, and business.

4.1 Healthcare

Traditionally, public health surveillance is supported by the reports of specific diseases from health care providers to public health officials. The approach is classified as indicator-based surveillance. Because health care providers are instructed to report cases that involve laboratory confirmations or meet certain definitions, the reports are usually credible. However, the limitations of the approach are coverage and timeliness[76]. The reports only come from the places where there are health infrastructures and health providers are willing to report the cases. Also, the report can only occur after sick persons seek medical attention.

The rapid growth of social media usage and the innovation of big data analysis provide an opportunity to identify health-related events through unstructured and non-standardized or subjective reports, blogs, stories, tweets, and other information. The approach is classified as event-based surveillance. It is cheap and flexible, thus avoiding the limitations of the traditional approach. The approach can be used anywhere, and can even detect an outbreak early before the patient seeks medical treatment. However, another set of concerns exists. Since information is not always monitored by professionals, the trustworthiness of the results has been questioned. The acceptability among public health practitioners and policymakers is another issue.

Social media is widely used in public health practice. A large number of extensive and heterogeneous social media-based prototypes have been developed to achieve disease surveillance and outbreak management. Some applications have been adopted by health authorities throughout the world. We identified thirteen examples as representatives. Twitter is a major information source. Most systems are funded by governments, universities, and hospitals, such as Global Public Health Intelligence Network (GPHIN) funded by the Canadian government[77] and Healthmap funded by Children's Hospital Boston and Harvard University[78]. Building multilingual applications is crucial to cover global events. GPHIN, built in 1997, can support 9 languages[77] and MedISys, built in 2008, can support 25 languages[79].

With the further development of data analysis techniques, outbreak detection has become more accurate and efficient. The acceptance among public health practitioners and policymakers has dramatically increased. The World Health Organization (WHO) uses the Canadian-based GPHIN in its global alert and response activities[77]. The European Centre for Disease Prevention and Control utilizes MedISys, a web-based tool for epidemic intelligence activities[80]. The US Centers for Disease Control and Prevention (CDC) has a special program, the Global Disease Detection

(GDD) Operations Center, dedicated to detecting and monitoring global public health events via Event-Based Surveillance (EBS)[81]. The Department of Health in Chicago and New York use Twitter and Yelp to identify foodborne illness[11, 12].

Some applications focus on one type of diseases, such as Foodborne New York and Foodborne Chicago[11, 12]. Others tend to cover multiple diseases, like GPHIN and Healthmap[77, 78]. However, almost all applications have a predetermined set of disease types and collect social media data based on these disease types. In such a way, the applications cannot identify outbreaks of new disease types. For instance, in December 2019, the new coronavirus, COVID-19, was first reported. An application that collects social media data with a predetermined disease name will not be able to detect this outbreak because this is a new disease type. However, COVID-19 symptoms include fever and difficulty in breathing, which are similar to many other respiratory diseases. Thus, the coverage of the applications can be improved by collecting data based on symptoms.

With the help of rich data from social media and advanced text mining technology, social media-based bio-surveillance systems can be used as early warning systems for upcoming public health emergencies. The systems document the impact of interventions, track progress towards specific goals, monitor and clarify the epidemiology of health problems, set priorities, and provide information for public health policies and strategies.

4.2 Disaster management

Dedicated physical sensors are traditionally required to detect natural disasters, such as floods, earthquakes, hurricanes, and wildfires. One type of physical sensors can typically detect one type of events and the maintenance costs are usually high. According to the National Severe Storms Laboratory, the WSR-88D radars graphically display the detected precipitation on a map, which helps to issue flash flood declarations, watches, or warnings[82]. However, floods can happen during heavy rains, but they can also happen when waves come on shore, when the snow melts too fast, or when dams or levees break.

When detecting natural disasters, we need to know what will happen, earthquakes, floods, or landslides; where and when it will happen; and who will be affected. As social media is becoming a real-time information dissemination platform, researchers have studied the use of social media in natural disaster detection.

F. Cheong and C. Cheong[83] and Yates and Paquette[84] provided case studies of tweets during Australian floods and Haitian Earthquakes. Imran et al.[85] and MacEachren et al.[86] proposed methods to raise situational awareness during natural disasters. Musaev and Hou[30] built a Landslide Information System (LITMUS) with multiple data sources, including Twitter, Facebook, and Instagram. They show that the pipeline of LITMUS can be easily generalized for other event types, such as infrastructure breakdowns[87].

Social media provides a new way to explore decentralized communication during disasters[88]. As citizens on the ubiquitous Web, social media users act as sensors and share their observations and views online[89]. Social media enhances situation awareness, especially during disaster management. During the Indian Ocean tsunami in 2004 and Hurricane Katarina in 2005, social media launched online disaster response communities[90]. Some communities reported the latest infrastructure conditions and others helped contact missing family members.

Smartphones have many built-in sensors, including the gyroscope, magnetometer, Global Positioning System (GPS), proximity sensor, ambient light sensor, microphone, touchscreen sensor, fingerprint sensor, pedometer, barometer, heart rate sensor, thermometer, and air humidity sensor. If future social media can enable sharing such sensor data, citizen sensors would have a greater impact on disaster responses.

4.3 Business

Social media helps disseminate information to a wider audience at a lower cost in a shorter period. From business and marketing perspectives, the proliferation of social media enables an innovative means of consumer engagement. The influence of social media on the relationship between companies and consumers cannot be underestimated. A survey of 2700 professionals in Europe revealed that a huge proportion of high-growth companies (81%) are using social tools to facilitate expansion and improve how teams collaborate and share knowledge[91].

Multi-way communication among consumers and providers is established on social media. With free of charge, social media platforms provide businesses an inexpensive way of advertising to a niche market. Viral marketing is becoming cheaper, more efficient,

and widespread. On the other hand, social media is a platform for consumers to share their comments. Such user-generated content furnishes companies with insights into their products and services and helps them advance and promote new product development.

The existing applications have studied social media data to assess the effectiveness of digital marketing and brand equity. They would like to effectively compose advertisements to better communicate their products and services with potential consumers. Desai and Han[92] used data from Instagram to explore the impact of textual comments on university brand equity. This is a crucial discovery because it is necessary to promote the university to a diverse group of people on social media, considering that the university recruits students from all over the world.

Some studies focus on helping businesses better monitor and understand their customers' feedbacks and comments on social media. Rui and Whinston[49] proposed a framework of a social-broadcasting-based Business Intelligence system that utilizes real-time information extracted from social media with text mining techniques. Business Intelligence (BI) platforms are built to acquire, interpret, collect, analyze, and explore information to assist business functions. Traditionally, BI analyzes internal data sources, such as operational data. The explosive growth of user-generated content has offered BI a new perspective.

Businesses can predict sales, revenues, and even firm equity, with the knowledge of public attitudes toward their products and services. Most traditional sales forecast time series models relaying on historical and seasonal sales, and they fail to capture the dynamic impact of recent events. The mixture of social media prediction models and traditional forecasting models is worthwhile in any field where consumer behavior is an important component. Ahn and Spangler[39] presented a predictive model of monthly automobile sales using social media data. Choudhery and Leung[93] presented a social data mining system to predict box office revenue of movies. Yu et al.[51] investigated the effect of social media on short term firm stock market performances. They applied the advanced sentiment analysis and used stock return & risk as the indicators of companies' short-term performances. Their findings suggest that when implementing social media marketing strategies, companies must carefully evaluate the importance of various media outlets.

However, multi-way communication among customers and providers on social media also poses some challenges for businesses. Traditionally, promotion only allows companies to talk to customers, while social media platforms allow customers to talk with each other directly. The content, time, and frequency of social media-based conversations among consumers cannot be controlled by businesses. Therefore, monitoring and understanding customers' comments promptly are extremely important for businesses. Additionally, businesses must learn to guide consumer discussions in a direction consistent with the organizations' mission and performance goals[94].

5 Challenge & Opportunity

5.1 Data privacy

"Knowledge is power, and in the internet age, knowledge is derived from data. Our personal data is what powers today's data-driven economy and the wealth it generates. It is time we had control over the use of our personal data. That includes keeping it private", said Attorney General Becerra[95]. Since 2018, many countries around the world have been in the process of reviewing and discussing privacy legislation bills. The General Data Protection Regulation (GDPR), enforceable beginning May 25th, 2018, regulates data privacy and protection of data subjects. The GDPR gives people greater control over their personal information collected by the businesses. In America, the California Consumer Privacy Act (CCPA) came into force on January 1st, 2020. It is the legislation that provides consumers with new rights relating to the access to, deletion of, and sharing of their personal information that is collected by businesses. South Korea is updating its regulations in the hope to achieve adequacy. The Brazilian General Data Protection Act will enter into force on August 15th, 2020.

The enforcement of privacy regulations, such as the GDPR and the CCPA, and other incoming privacy laws increase the public's awareness of data privacy, especially in the organizations which collect and process personal data. There are also researchers highlighting the necessity and the complexity of researching public interest while protecting users' privacy. Privacy laws from different regions may differ but their goals are the same, to protect users' data privacy. This section uses the GDPR as an example and focuses on discussing the consequences of GDPR on social media-based applications and how to develop GDPR-compliant

applications.

Data privacy pertains to the use of personal data or personally identifiable information. Article 4(1) in GDPR defines "personal data" as any information relating to an identified or identifiable natural person (data subject)[96]. Information shared on social media platforms can be categorized as personal data because it relates to an identified or identifiable natural person. The personal data from social media are not owned by the platforms but are owned by a data subject. In the GDPR, there is a significant difference between collecting data from a data subject and collecting data from the third parties, because the scale of the collection from the third parties tends to be tremendously larger than the scale of the collection from the data subject himself. The applications discussed in the survey are collecting and processing data from social media platforms which are the third parties.

The GDPR concerns when data subjects should be informed regarding the third parties' data processing. The principle of transparency is put into action. Transparency addresses the right of data subject to know and understand how the data are being used. It seems relatively easy to inform data subjects when collecting data directly from them, but this can become difficult when data are collected from the third parties, such as collecting data via Twitter stream API. There are some exemptions embedded in the GDPR, codifying and specifying research conduct. The exemptions assume that social media platforms have already informed their users through appropriate Terms of Services that their data will be shared with the third parties (e.g., through API) or assume that the large scale of collected data will require a disproportionate effort to inform all affected data subjects.

Once we have obtained the personal information lawfully, we must assess whether we have lawful bases for data processing. Article 6 in GDPR lists six lawful bases and three of them are relevant to our scenario: (1) the consent of data subject, (2) the task carried out in the public interest, and (3) the legitimate interests of the controllers. The consent is not needed if the processing is for the public interest. Some regions, such as Sweden, grant a significant level of freedom to research done by public institutions. However, such freedom is not available to commercial organizations, who must rely on consent or legitimate interests.

While processing data, suitable safeguards are required, especially for sensitive data, such as ethnicity and sexual preferences. Also, the GDPR emphasizes on the concept of profiling. Profiling is defined as automated analysis to identify correlations and apply the correlations to an individual to identify characteristics of his present or future behaviors. Given the definition above, most of the researches presented in previous sections can be categorized as profiling. Such researches may require the performance of a Data Protection Impact Assessment (DPIA) typically conducted by the data protection officer.

Data storage is another concern of the GDPR. If we want to store the collected data after processing, we need to provide the accessibility of the stored data to the data subject and also allow data subject to withdraw the data, as a data subject has the right of access (Article 15 in GDPR), the right to rectification (Article 16 in GDPR), the right to be forgotten (Article 17 in GDPR), and the right to restriction of processing (Article 18 in GDPR).

With the legal obligations emerging from the GDPR, we shall consider the restrictions while collecting and processing personal data. We have to build the GDPR-compliant applications and respect the data subject's rights, acknowledging that data subject must be treated with dignity and respect regardless of research aims. In the meantime, future opportunities are presented. Notification systems need to be designed and developed to enable large scale data subject notification for the third parties' data collection and data processing. Standard and automatic tools are expected to fulfill data subject requests, including withdrawing or modifying his data. Designing such tools, which can communicate between social media users who are data subjects, and researchers who process their data will reduce the burdens of data privacy compliance in future research.

Additionally, privacy-preserving mechanisms should also be implemented by social media platforms to protect users' privacy. Neglectfully publishing content online can lead to severe disclosure of sensitive information. Han et al.[97] and Zheng et al.[98] proposed a privacy preservation framework that regulates the data publishing behaviors to hide sensitive physical profiles. They further improved the framework to guarantee high performance while preserving privacy and guaranteeing fairness among users[99, 100]. Cai et al.[101] stated that an attacker can utilize user profiles and social relationships to predict sensitive information from social networks. To protect against this, they proposed a data sanitization method which collectively manipulates user profile and social relationships to reduce adversary's prediction

accuracy on sensitive information.

5.2 Rumor & fake news

With billions of users worldwide, social media is becoming one of the major information sources. However, being under fewer controls and restrictions, comparing to official news sources, social media carries a lot of false claims and, additionally, its popularity makes the spread of rumors very easy and fast. It is a big concern as it may lead to economic or political turmoil. Nshetri and Voas[79] stated that false information has two forms: misinformation (incorrect information) or disinformation (information used to deceive its audience). A rumor on social media is a piece of information that is not verified for its truthfulness at the time of posting. A rumor can spread misinformation or disinformation. Fake news consist of information that is intentionally and verifiably false with a motive to mislead readers. Many researchers have been focusing on identifying rumors and fake news on social media. There are five common approaches.

Classification-based approach. Rumor detection can be considered as a binary classification problem. Supervised learning approaches, such as Support Vector Machine and Logistics Regression, are used to classify whether one trending topic is fake or not. Researchers improve the performance of the classification models for rumor detection with feature engineering. Castillo et al.[102] used content-based, user-based, behavior-based, and propagation-based features. Sun et al.[103] and Yang et al.[104] extracted multimedia-based and location-based features. Kwon et al.[105] proposed the model based on temporal, structural, and linguistic features of rumor propagation.

Pattern-based approach. This approach summarizes rumor-related patterns and uses matching techniques to highlight disputed posts from social media[106]. Zhao et al.[107] identified rumors based on inquiry phrase patterns.

Simulation-based approach. A micro-scale generative model for context-driven viral activity is presented by Sathanur et al.[108] The work leverages the stochastic discrete-event agent-based simulator PhySense to simulate viral activity cascades on synthesized network topologies and then examine strategies to control such activity cascades via appropriate metrics.

Time series-based approach. Kotteti et al.[109] proposed a quick detection of rumors by using the temporal properties of tweets. Without many features, the training time and the prediction time can be significantly reduced.

Crowdsourcing-based approach. Lin et al.[110] proposed a rumor control framework, called Crowdblocking, allowing users to implement control schemes in a collaborative and distributed way. They designed two incentive mechanisms to motivate more users to actively participate in rumor blocking activities.

The previous work in rumor detection provides a solid foundation to further advance social media-based event detection applications. We have studied the applications built to realize natural disaster detection and bio-surveillance. Future works can integrate steps to identify rumors in event detection applications and systematically evaluate the results.

5.3 Image & video

With the rapid development of the internet speed, images and videos are becoming significant parts of social media, such as Instagram (images) and YouTube (videos). Image analysis is maturing as well. Machine learning and neural networks are available to automatically annotate images and to segment certain objects in images.

Images on social media have been used as training datasets in computer vision projects. It is easier to understand the content of images from social media because there are texts and geographic tags attached. Qin et al.[111] proposed to use convolutional neural networks and fully convolutional networks to classify high resolution urban remote sensing images and the labeled images from social media are used as training datasets to reduce the cost of manual labeling. Hoffmann et al.[112] regarded social media images as valuable information to assess building usages in neighborhoods. A logistic regression classifier is trained to distinguish among five different building usage classes.

Future works can navigate through social media by focusing on textual and visual aspects. With the existing social media-based applications, we can design ways to add image analysis to further improve the performance. For example, in addition to text analysis, an emergency management system can be boosted by incorporating image classifiers[113].

5.4 Multilingual support

Due to economic globalization, most social media channels, including Facebook and Twitter, are open

to all languages and people throughout the world. Although English has been adopted by many as a de facto standard international language, it is common for people from each country to exchange information in their languages. On Twitter, a study in 2013 found that just 34% tweets were written in English, as shown in Fig. 6. Consequently, applications using social media as a data source might need to support languages other than English. Disaster detection applications would achieve better event coverage if they can analyze data in different languages because discussions about an event that happened in a country are more likely to have postings in the local language in addition to English.

Designers of social media applications with global coverage often face a difficult trade-off between (1) supporting only English (without local information) and (2) laboriously localizing the applications in each language of interest[114]. Attempting to integrate social media information from many languages is an even more significant challenge. Programmers and translators familiar with the native languages are required to extend one application to support an additional language. For machine learning models, training datasets in corresponding languages are required to train models for a new language. Hou et al.[114] proposed to use Online Machine Translation tools, such as Google and Bing translators, as the promising cheaper and faster alternatives. They conducted a systematic comparative study to explore the design alternatives that adopt an increasing number of manually developed filtering stages. The experiment results show that automated translation can achieve comparable or better results. Xia et al.[115] also explored the use of machine translation

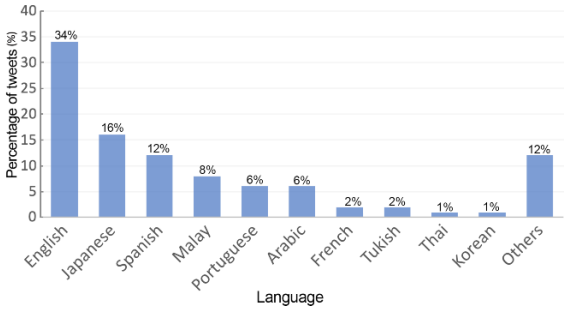


Fig. 6 Most used languages on Twitter in September 2013.

to assist Wikipedia translation activities. To achieve multilingual sentiment classification, Fuadvy and Ibrahim[116] produced a comparable Malay dataset from the English dataset with Google Machine Translation and then combined English and Malay datasets into a single dataset to build a deep learning model with multilingual word embedding.

Zhou et al.[117] aimed to develop cross-lingual sentiment classification so that the resources in a resource-rich language (such as English) can be used to classify the sentiment polarity of texts in a resource-poor language (such as Japanese). The biggest challenge is the vocabulary gap between languages, and they propose a cross-lingual representation learning model BiDRL which simultaneously learns words and documents in both languages. Xie et al.[118] built MuSES, a multilingual sentiment identification system with a novel zero-effort labeling approach that leverages knowledge bases like Wikipedia and labels word-level sentiment for non-English words.

Even more interestingly, in multi-ethnic and multi-cultural countries, such as Malaysia, with a population made up of Malays, Chinese, Indians, and other migrants, people sometimes mix several languages to express their opinions. It is becoming an urgent need to support multiple languages in social media-based applications, especially those for healthcare and disaster management purposes. Previous work has studied the potential benefits of adopting machine translators or building multilingual word embeddings, but more researches have to be done to further improve multilingual supports to handle large scale data on time.

5.5 5G wireless network

The 5th generation mobile communication assures almost unlimited access to information at any time[119]. Cellular generation evolution is shown in Fig. 7. Wireless networks were originally built for sharing voice and text. Around 2000, 3G enabled mobile broadband for people to surf the web and it also empowered the beginning of social media. With the 4G enhanced mobile broadband capabilities, social media users dramatically increase, and the content of social media is no longer limited to texts. Now, 5G will

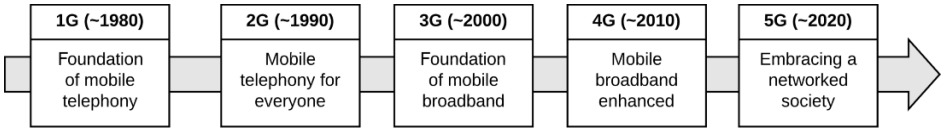


Fig. 7 Cellular generation evolution.

further increase performance levels. Global mobile video traffic keeps increasing from 2017 according to Cisco, as shown in Fig. 8[120]. Additionally, Ericsson Mobility Report forecasts mobile video traffic will grow around 50% annually through 2022 and social networking will grow by 39% annually through 2020. The use of embedded video in social media continues to grow, fueled by larger device screens, higher resolution, and new platforms supporting live stream[121]. We are expecting extraordinary opportunities in studying video processing and incorporating video content in social media applications.

Furthermore, social media live streaming has made the leap from novelty to necessity. Twitter, Facebook, Instagram, and LinkedIn have integrated live video to their platforms. In 2018, an estimated 47% of global consumers reported an increase in their live streaming from the previous year[122]. Streaming content is high-volume real-time data and it is challenging to process and analyze. There will be many potential benefits in understanding live streaming content. For instance, during or after disasters, live streaming content can be a valuable source for first responders to monitor the situations and identify the needs in the affected areas. The applications which consume social media data expect an increase in video content. Future studies are needed to process videos and incorporate processed information into existing social media-based applications.

6 Conclusion

Different from the traditional information source, social media provides user-generated content with large volume, high velocity, and a wide variety. In this survey, we conduct a comprehensive review of the previous literature which discusses extracting values and insights from social media data. We systematically compare the existing applications and categorize those based on their analysis techniques and impacting areas. From practical perspectives, we outline a commonly used pipeline in building such applications and focus on discussing popular analysis techniques.

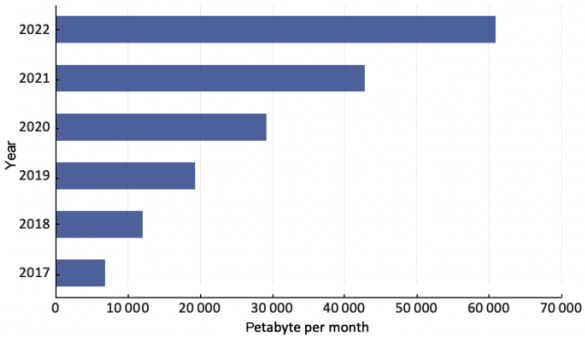


Fig. 8 Global mobile video traffic from 2017 to 2022[120].

After that, we emphasize three fields in terms of their impacts, which are healthcare, disaster management, and business, in the hope to provide a broader exploration of such applications. Finally, we point out the existing challenges and outlook the future research directions in the field of social media-based applications in terms of data privacy, the 5G wireless network, and multilingual support.

References

[1] A. M. Kaplan and M. Haenlein, Users of the world, unite! The challenges and opportunities of social media, *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[2] P. N. Howard and M. R. Parks, Social media and political change: Capacity, constraint, and consequence, *Journal of Communication*, vol. 62, no. 2, pp. 359–362, 2012.

[3] C. T. Carr and R. A. Hayes, Social media: Defining, developing, and divining, *Atlantic Journal of Communication*, vol. 23, no. 1, pp. 46–65, 2015.

[4] Data never sleeps 5.0, https://www.domo.com/learn/data-never-sleeps-5, 2020.

[5] E. Ortiz-Ospina, The rise of social media, https://ourworldindata.org/rise-of-social-media, 2019.

[6] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, Social media analytics—Challenges in topic discovery, data collection, and data preparation, *International Journal of Information Management*, vol. 39, pp. 156–168, 2018.

[7] I. Moalla, A. Nabli, L. Bouzguenda, and M. Hammami, Data warehouse design approaches from social media: Review and comparison, *Social Network Analysis and Mining*, vol. 7, no. 1, p. 5, 2017.

[8] J. W. Cao, K. Basoglu, H. Sheng, and P. B. Lowry, A systematic review of social networking research in information systems, *Communications of the Association for Information Systems*, vol. 36, no. 1, 2015.

[9] S. Karimi, J. Yin, and C. Paris, Classifying microblogs for disasters, in *Proc. 18th Australasian Document Computing Symp.*, Brisbane, Australia, 2013, pp. 26–33.

[10] T. Sakaki, M. Okazaki, and Y. Matsuo, Earthquake shakes twitter users: Real-time event detection by social sensors, in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, 2010, pp. 851–860.

[11] J. K. Harris, R. Mansour, B. Choucair, J. Olson, C. Nissen, and J. Bhatt, Health department use of social media to identify foodborne illness, Chicago, IL, USA, 2013–2014, *Morbidity and Mortality Weekly Report*, vol. 63, no. 32, pp. 681–685, 2014.

[12] E. O. Nsoesie, S. A. Kluberg, and J. S. Brownstein, Online

reports of foodborne illness capture foods implicated in official foodborne outbreak reports, *Preventive Medicine*, vol. 67, pp. 264–269, 2014.

[13] M. Han, M. Y. Yan, Z. P. Cai, and Y. S. Li, An exploration of broader influence maximization in timeliness networks with opportunistic selection, *Journal of Network and Computer Applications*, vol. 63, pp. 39–49, 2016.

[14] P. P. Xu, Y. C. Wu, E. X. Wei, T. Q. Peng, S. X. Liu, J. J. H. Zhu, and H. M. Qu, Visual analysis of topic competition on social media, *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2012–2021, 2013.

[15] W. W. Dou, X. Y. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, LeadLine: Interactive visual analysis of text data through event identification and exploration, presented at 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), Seattle, WA, USA, 2012, pp. 93–102.

[16] S. M. Chen, X. R. Yuan, Z. H. Wang, C. Guo, J. Liang, Z. C. Wang, X. L. Zhang, and J. W. Zhang, Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data, *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 270–279, 2016.

[17] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, Twitinfo: Aggregating and visualizing microblogs for event exploration, in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, Vancouver, Canada, 2011, pp. 227–236.

[18] M. Cataldi, L. Di Caro, and C. Schifanella, Emerging topic detection on twitter based on temporal and social terms evaluation, in *Proc. 10^{th} Int. Workshop on Multimedia Data Mining*, Washington, DC, USA, 2010, pp. 1–10.

[19] S. Sizov, GeoFolk: Latent spatial semantics in web 2.0 social media, in *Proc. 3^{rd} ACM Int.l Conf. Web Search and Data Mining*, New York, NY, USA, 2010, pp. 281–290.

[20] S. Y. Song, Q. D. Li, and N. Zheng, A spatio-temporal framework for related topic search in micro-blogging, presented at International Conference on Active Media Technology, A. An, P. Lingras, S. Petty, and R. Huang, eds. Berlin, Germany: Springer, 2010, pp. 63–73.

[21] S. C. Han and B. H. Kang, Identifying the relevance of social issues to a target, presented at 2012 IEEE 19th Int. Conf. Web Services, Honolulu, HI, USA, 2012, pp. 666–667.

[22] S. Y. Song, Q. D. Li, and X. L. Zheng, Detecting popular topics in micro-blogging based on a user interest-based model, presented at 2012 Int. Joint Conf. Neural Networks (IJCNN), Brisbane, Australia, 2012, pp. 1–8.

[23] B. Hu, Z. Song, and M. Ester, User features and social networks for topic modeling in online social media, presented at 2012 IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining, Istanbul, Turkey, 2012, pp. 202–209.

[24] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Y. Cheng, Spatio-temporal dynamics of online memes: A study of geo-tagged tweets, in *Proc. 22^{nd} Int. Conf. World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 667–678.

[25] Z. Q. Ma, W. W. Dou, X. Y. Wang, and S. Akella, Tag-latent dirichlet allocation: Understanding hashtags and their relationships, presented at 2013 IEEE/WIC/ACM Int. Joint Conf. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 2013, pp. 260–267.

[26] P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski, The social media genome: Modeling individual topic-specific behavior in social media, in *Proc. 2013 IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining*, Niagara, Canada, 2013, pp. 236–242.

[27] G. Jang and S. H. Myaeng, Analysis of spatially oriented topic versatility over time on social media, in *Proc. 2015 IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining*, Paris, France, 2015, pp. 573–578.

[28] F. Z. Yao, K. C. C. Chang, and R. H. Campbell, Ushio: Analyzing news media and public trends in twitter, in *Proc. 2015 IEEE/ACM 8^{th} Int. Conf. Utility and Cloud Computing* (*UCC*), Limassol, Cyprus, 2015, pp. 424–429.

[29] S. S. Qian, T. Z. Zhang, C. S. Xu, and J. Shao, Multi-modal event topic model for social event analysis, *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 233–246, 2016.

[30] A. Musaev and Q. X. Hou, Gathering high quality information on landslides from twitter by relevance ranking of users and tweets, presented at 2016 IEEE 2nd Int. Conf. Collaboration and Internet Computing(CIC), Pittsburgh, PA, USA, 2016, pp. 276–284.

[31] V. A. Rohani, S. Shayaa, and G. Babanejaddehaki, Topic modeling for social media content: A practical approach, presented at 2016 3rd Int. Conf. Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 2016, pp. 397–402.

[32] A. Argyrou, S. Giannoulakis, and N. Tsapatsoulis, Topic modelling on Instagram hashtags: An alternative way to automatic image annotation? presented at 2018 13th Int. Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Zaragoza, Spain, 2018, pp. 61–67.

[33] A. Ejaz, S. K. Fatima, Q. N. Rajput, and S. A. Khoja, Analyzing news from electronic media and topics discussed on social media using ontology, presented at 2018 5th Int. Conf. Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, 2018, pp. 349–354.

[34] T. C. Fu, D. C. M. Sze, P. K. C. Leung, K. Y. Hung, and F. L. Chung, Analysis and visualization of time series data from consumer-generated media and news archives, presented at 2007 IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology-Workshops, Silicon Valley, CA, USA, 2007, pp. 259–262.

[35] M. Mathioudakis and N. Koudas, Twittermonitor: Trend detection over the twitter stream, in *Proc. 2010 ACM SIGMOD Int. Conf. Management of Data*, Indianapolis, IN, USA, 2010, pp. 1155–1158.

[36] J. Yang and J. Leskovec, Patterns of temporal variation in online media, in Proc. 4th ACM Int. Conf. Web Search and Data Mining, Hong Kong, China, 2011, pp. 177–186.

[37] W. X. Zhao, B. H. Shu, J. Jiang, Y. Song, H. F. Yan, and X. M. Li, Identifying event-related bursts via social media

activities, in *Proc. 2012 Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 2012, pp. 1466–1477.

[38] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl, Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition, presented at 2012 IEEE Conf. Visual Analytics Science and Technology (VAST), Seattle, WA, USA, 2012, pp. 143–152.

[39] H. I. Ahn and W. S. Spangler, Sales prediction with social media analysis, presented at 2014 Annu. SRII Global Conf., San Jose, CA, USA, 2014, pp. 213–222.

[40] K. Kucher, A. Kerren, C. Paradis, and M. Sahlgren, Visual analysis of stance markers in online social media, presented at 2014 IEEE Conf. Visual Analytics Science and Technology (VAST), Paris, France, 2014, pp. 259–260.

[41] P. Healy, G. Hunt, S. Kilroy, T. Lynn, J. P. Morrison, and S. Venkatagiri, Evaluation of peak detection algorithms for social media event detection, presented at 2015 10th Int. Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Trento, Italy, 2015, pp. 1–9.

[42] Y. Tsuboi, A. Jatowt, and K. Tanaka, Product purchase prediction based on time series data analysis in social media, presented at 2015 IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology (WI-IAT), Singapore, 2015, pp. 219–224.

[43] K. Nusratullah, S. A. Khan, A. Shah, and W. H. Butt, Detecting changes in context using time series analysis of social network, presented at 2015 SAI Intelligent Systems Conf. (IntelliSys), London, UK, 2015, pp. 996–1001.

[44] S. D. Johnson and K. Y. Ni, A pricing mechanism using social media and web data to infer dynamic consumer valuations, presented at 2015 IEEE Int. Conf. Big Data (Big Data), Santa Clara, CA, USA, 2015, pp. 2868–2870.

[45] K. Zhao, Y. Q. Zhang, B. G. Li, and C. F. Zhou, Repost number prediction of micro-blog on Sina Weibo using time series fitting and regression analysis, presented at 2015 Int. Conf. Identification, Information, and Knowledge in the Internet of Things (IIKI), Beijing, China, 2015, pp. 66–69.

[46] M. Ni, Q. He, and J. Gao, Forecasting the subway passenger flow under event occurrences with social media, *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1623–1632, 2017.

[47] E. A. Dahouei, A cloud-based dashboard for time series analysis on hot topics from social media, presented at 2017 Int. Conf. Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India, 2017, pp. 1–6.

[48] C. Comito, D. Falcone, and D. Talia, A peak detection method to uncover events from social media, presented at 2017 IEEE Int. Conf. Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 2017, pp. 459–467.

[49] H. X. Rui and A. Whinston, Designing a social-broadcasting-based business intelligence system, *ACM Transactions on Management Information Systems* (*TMIS*), vol. 2, no. 4, pp. 1–19, 2012.

[50] B. Yuan, Y. Liu, and H. Li, Sentiment classification in Chinese microblogs: Lexicon-based and learning-based approaches, *International Proceedings of Economics Development and Research*, vol. 68, pp. 1–5, 2013.

[51] Y. Yu, W. J. Duan, and Q. Cao, The impact of social and conventional media on firm equity value: A sentiment analysis approach, *Decision Support Systems*, vol. 55, no. 4, pp. 919–926, 2013.

[52] Z. X. Wang, V. J. C. Tong, and H. C. Chin, Enhancing machine-learning methods for sentiment classification of web data, in *Information Retrieval Technology*, A. Jaafar, N. M. Ali, S. A. M. Noah, A. F. Smeaton, P. Bruza, Z. A. Bakar, N. Jamil, and T. M. T. Sembok, eds. Cham, Switzerland: Springer, 2014, pp. 394–405.

[53] Z. X. Wang, C. S. Chong, L. Lan, Y. P. Yang, S. B. Ho, and J. C. Tong, Fine-grained sentiment analysis of social media with emotion sensing, presented at 2016 Future Technologies Conf. (FTC), San Francisco, CA, USA, 2016, pp. 1361–1364.

[54] A. Popescul and L. H. Ungar, Statistical relational learning for link prediction, in *Proc. Workshop on Learning Statistical Models from Relational Data at IJCAI-2003*, Acapulco, Mexico, 2003, pp. 109–115.

[55] D. L. Nowell and J. Kleinberg, The link-prediction problem for social networks, *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[56] G. Rossetti, M. Berlingerio, and F. Giannotti, Scalable link prediction on multidimensional networks, presented at 2011 IEEE 11th Int. Conf. Data Mining Workshops, Vancouver, Canada, 2011, pp. 979–986.

[57] R. Michalski, P. Kazienko, and D. Krol, Predicting social network measures using machine learning approach, in *Proc. 2012 Int. Conf. Advances in Social Networks Analysis and Mining*, Washington, DC, USA, 2012, pp. 1056–1059.

[58] M. A. Smith, NodeXL: Simple network analysis for social media, presented at 2013 Int. Conf. Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 2013, pp. 89–93.

[59] F. Erlandsson, P. Bródka, A. Borg, and H. Johnson, Finding influential users in social media using association rule learning, *Entropy*, vol. 18, no. 5, p. 164, 2016.

[60] J. Li, Z. P. Cai, M. Y. Yan, and Y. S. Li, Using crowdsourced data in location-based social networks to explore influence maximization, presented at IEEE INFOCOM 2016—The 35th Annu. IEEE Int. Conf. Computer Communications, San Francisco, CA, USA, 2016, pp. 1–9.

[61] X. H. Zhao, F. A. Liu, J. L. Wang, and T. L. Li, Evaluating influential nodes in social networks by local centrality with a coefficient, *ISPRS International Journal of Geo-Information*, vol. 6, no. 2, p. 35, 2017.

[62] W. Y. Tang, G. C. Luo, Y. B. Wu, L. Tian, X. Zheng, and Z. P. Cai, A second-order diffusion model for influence maximization in social networks, *IEEE Transactions on Computational Social Systems*, vol. 6, no. 4, pp. 702–714, 2019.

[63] T. O. Sprenger, Essays on the information content of microblogs and their use as an indicator of real-world events, PhD dissertation, Technische Universität München, München, Germany, 2011.

[64] G. K. Palshikar, Simple algorithms for peak detection in time-series, http://constans.pbworks.com/w/file/fetch/120908295/Simple_Algorithms_for_Peak_Detection_in_Time-Serie.pdf, 2009.

[65] P. Du, W. A. Kibbe, and S. M. Lin, Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[66] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, Dynamical classes of collective attention in twitter, in *Proc. 21st Int. Conf. World Wide Web*, Lyon, France, 2012, pp. 251–260.

[67] L. Y. Liu, J. L. Priestley, Y. Y. Zhou, H. E. Ray, and M. Han, A2Text-Net: A novel deep neural network for sarcasm detection, presented at 2019 IEEE 1st Int. Conf. Cognitive Machine Intelligence, Los Angeles, CA, USA, 2019, pp. 118–126.

[68] S. Desai and M. Han, Social media content analytics beyond the text: A case study of university branding in instagram, in *Proc. 2019 ACM Southeast Conf.*, Kennesaw, GA, USA, 2019, pp. 94–101.

[69] J. S. He, M. Han, S. L. Ji, T. Y. Du, and Z. Li, Spreading social influence with both positive and negative opinions in online networks, *Big Data Mining and Analytics*, vol. 2, no. 2, pp. 100–117, 2019.

[70] C. Banea, R. Mihalcea, and J. Wiebe, A bootstrapping method for building subjectivity lexicons for languages with scarce resources, in *Proc. Int. Conf. Language Resources and Evaluation*, Marrakech, Morocco, 2008, pp. 2764–2767.

[71] C. Strapparava and A. Valitutti, WordNet-affect: An affective extension of WordNet, in *Proc. 4th Int. Conf. Language Resources and Evaluation*, Lisbon, Portugal, 2004, pp. 1083–1086.

[72] S. Baccianella, A. Esuli, and F. Sebastiani, SentiWordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in *Proc. Int. Conf. Language Resources and Evaluation*, Valletta, Malta, 2010, pp. 2200–2204.

[73] I. Perikos and I. Hatzilygeroudis, A framework for analyzing big social data and modelling emotions in social media, presented at 2018 IEEE 4th Int. Conf. Big Data Computing Service and Applications (BigDataService), Bamberg, Germany, 2018, pp. 80–84.

[74] M. Han, M. Y. Yan, Z. P. Cai, Y. S. Li, X. Q. Cai, and J. G. Yu, Influence maximization by probing partial communities in dynamic online social networks, *Transactions on Emerging Telecommunications Technologies*, vol. 28, no. 4, p. e3054, 2017.

[75] M. Han, M. Y. Yan, J. B. Li, S. L. Ji, and Y. S. Li, Neighborhood-based uncertainty generation in social networks, *Journal of Combinatorial Optimization*, vol. 28, no. 3, pp. 561–576, 2014.

[76] M. Han, Z. J. Duan, C. Y. Ai, F. W. Lybarger, Y. S. Li, and A. G. Bourgeois, Time constraint influence maximization algorithm in the age of big data, *International Journal of Computational Science and Engineering*, vol. 15, nos. 3&4, pp. 165–175, 2017.

[77] E. Mykhalovskiy and L. Weir, The global public health intelligence network and early warning outbreak detection, *Canadian Journal of Public Health*, vol. 97, no. 1, pp. 42–44, 2006.

[78] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports, *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, 2008.

[79] N. Kshetri and J. Voas, The economics of "fake news", *IT Professional*, vol. 19, no. 6, pp. 8–12, 2017.

[80] Report on ECDC/JRC collaboration on development of online tool for epidemic intelligence, https://www.ecdc.europa.eu/en/news-events/report-ecdcjrc-collaboration-development-online-tool-epidemic-intelligence, 2011.

[81] Event-based surveillance, https://www.cdc.gov/globalhealth/healthprotection/gddopscenter/how.html, 2020.

[82] Severe weather 101–Floods, https://www.nssl.noaa.gov/education/svrwx101/floods/detection/, 2020.

[83] F. Cheong and C. Cheong, Social media data mining: A social network analysis of tweets during the Australian 2010–2011 floods, presented at PACIS 2011—15th Pacific Asia Conf. Information Systems: Quality Research in Pacific, Brisbane, Australia, 2011, pp. 1–16.

[84] D. Yates and S. Paquette, Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake, *International Journal of Information Management*, vol. 31, no. 1, pp. 6–13, 2011.

[85] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, Extracting information nuggets from disaster-related messages in social media, in *Proc. 10th Int. ISCRAM Conf.*, Baden-Baden, Germany, 2013, pp. 1–10.

[86] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, SensePlace2: GeoTwitter analytics support for situational awareness, presented at 2011 IEEE Conf. Visual Analytics Science and Technology (VAST), Providence, RI, USA, 2011, pp. 181–190.

[87] A. Musaev, Z. Jiang, S. Jones, P. Sheinidashtegol, and M. Dzhumaliev, Detection of damage and failure events of road infrastructure using social media, in *International Conference on Web Services*, H. Jin, Q. Wang, and L. J. Zhang, eds. Seattle, WA, USA: Springer, 2018, pp. 134–148.

[88] D. Murthy and A. J. Gross, Social media processes in disasters: Implications of emergent technology use, *Social Science Research*, vol. 63, pp. 356–370, 2017.

[89] A. Sheth, Citizen sensing, social signals, and enriching human experience, *IEEE Internet Computing*, vol. 13, no. 4, pp. 87–92, 2009.

[90] M. Laituri and K. Kodrich, On line disaster response

community: People as sensors of high magnitude disasters using internet GIS, *Sensors*, vol. 8, no. 5, pp. 3037–3055, 2008.

[91] J. Wichmann, Being B2B social: A conversation with Maersk Line's head of social media, https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/being-b2b-social-a-conversation-with-maersk-lines-head-of-social-media, 2013.

[92] S. Desai and M. Han, Social media content analytics beyond the text: A case study of university branding in instagram, in *Proc. 2019 ACM Southeast Conf.*, Kennesaw, GA, USA, 2019, pp. 94–101.

[93] D. Choudhery and C. K. Leung, Social media mining: prediction of box office revenue, in *Proc. 21st Int. Database Engineering & Applications Symp.*, Bristol, UK, 2017, pp. 20–29.

[94] W. G. Mangold and D. J. Faulds, Social media: The new hybrid element of the promotion mix, *Business Horizons*, vol. 52, no. 4, pp. 357–365, 2009.

[95] Attorney general Becerra publicly releases proposed regulations under the California consumer privacy act, https://oag.ca.gov/news/press-releases/attorney-general-becerra-publicly-releases-proposed-regulations-under-california, 2019.

[96] European Commission, General data protection regulation, Article 4(1), https://gdpr-info.eu/art-4-gdpr/, 2018.

[97] M. Han, Q. L. Han, L. J. Li, J. Li, and Y. S. Li, Maximising influence in sensed heterogeneous social network with privacy preservation, *International Journal of Sensor Networks*, vol. 28, no. 2, pp. 69–79, 2018.

[98] X. Zheng, Z. P. Cai, J. G. Yu, C. K. Wang, and Y. S. Li, Follow but no track: Privacy preserved profile publishing in cyber-physical social systems, *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1868–1878, 2017.

[99] X. Zheng, G. C. Luo, and Z. P. Cai, A fair mechanism for private data publication in online social networks, *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 880–891, 2020.

[100] M. Han, L. Li, Y. Xie, J. B. Wang, Z. J. Duan, J. Li, and M. Y. Yan, Cognitive approach for location privacy protection, *IEEE Access*, vol. 6, pp. 13 466–13 477, 2018.

[101] Z. P. Cai, Z. B. He, X. Guan, and Y. S. Li, Collective data-sanitization for preventing sensitive information inference attacks in social networks, *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2018.

[102] C. Castillo, M. Mendoza, and B. Poblete, Information credibility on twitter, in *Proc. 20th Int. Conf. World Wide Web*, Hyderabad, India, 2011, pp. 675–684.

[103] S. Y. Sun, H. Y. Liu, J. He, and X. Y. Du, Detecting event rumors on Sina Weibo automatically, in *Asia-Pacific Web Conference*, Y. Ishikawa, J. Li, W. Wang, R. Zhang, and W. Zhang, eds. Berlin, Germany: Springer, 2013, pp. 120–131.

[104] F. Yang, Y. Liu, X. H. Yu, and M. Yang, Automatic detection of rumor on Sina Weibo, in *Proc. ACM SIGKDD Workshop on Mining Data Semantics*, Beijing, China, 2012, pp. 1–7.

[105] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. J. Wang, Prominent features of rumor propagation in online social media, presented at 2013 IEEE 13th Int. Conf. Data Mining, Dallas, TX, USA, 2013, pp. 1103–1108.

[106] R. Ennals, D. Byler, J. M. Agosta, and B. Rosario, What is disputed on the web? in *Proc. 4th Workshop on Information Credibility*, Raleigh, NC, USA, 2010, pp. 67–74.

[107] Z. Zhao, P. Resnick, and Q. Z. Mei, Enquiring minds: Early detection of rumors in social media from enquiry posts, in *Proc. 24th Int. Conf. World Wide Web*, Florence, Italy, 2015, pp. 1395–1405.

[108] A. V. Sathanur, M. Sui, and V. Jandhyala, Assessing strategies for controlling viral rumor propagation on social media—A simulation approach, presented at 2015 IEEE Int. Symp. Technologies for Homeland Security (HST), Waltham, MA, USA, 2015, pp. 1–6.

[109] C. M. M. Kotteti, X. S. Dong, and L. J. Qian, Multiple time-series data analysis for rumor detection on social media, presented at 2018 IEEE Int. Conf. Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4413–4419.

[110] Y. G. Lin, Z. P. Cai, X. M. Wang, and F. Hao, Incentive mechanisms for crowdblocking rumors in mobile social networks, *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 9220–9232, 2019.

[111] Y. Q. Qin, M. M. Chi, X. Liu, Y. F. Zhang, Y. J. Zeng, and Z. M. Zhao, Classification of high resolution urban remote sensing images using deep networks by integration of social media photos, presented at IGARSS 2018-2018 IEEE Int. Geoscience and Remote Sensing Symp., Valencia, Spain, 2018, pp. 7243–7246.

[112] E. J. Hoffmann, M. Werner, and X. X. Zhu, Building instance classification using social media images, presented at 2019 Joint Urban Remote Sensing Event (JURSE), Vannes, France, 2019, pp. 1–4.

[113] M. Jing, B. W. Scotney, S. A. Coleman, and M. T. McGinnity, The application of social media image analysis to an emergency management system, presented at 2016 11th Int. Conf. Availability, Reliability and Security (ARES), Salzburg, Austria, 2016, pp. 805–810.

[114] Q. X. Hou, A. Musaev, Y. Yang, and C. Pu, A comparative study of increasing automation in the integration of multilingual social media information, presented at 2017 IEEE 3rd Int. Conf. Collaboration and Internet Computing (CIC), San Jose, CA, USA, 2017, pp. 319–327.

[115] L. S. Xia, N. Yamashita, and T. Ishida, Analysis on multilingual discussion for Wikipedia translation, presented at 2011 2nd Int. Conf. Culture and Computing, Kyoto, Japan, 2011, pp. 104–109.

[116] M. J. Fuadvy and R. Ibrahim, Multilingual sentiment analysis on social media disaster data, presented at 2019 Int. Conf. Electrical, Electronics and Information Engineering (ICEEIE), Denpasar, Indonesia, 2019, pp. 269–272.

[117] X. Zhou, X. Wan, and J. Xiao, Cross-lingual sentiment classification with bilingual document representation learning, in *Proc. 54th Annu. Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016, pp. 1403–1412.

[118] Y. S. Xie, Z. Z. Chen, K. P. Zhang, Y. Cheng, D. K. Honbo,

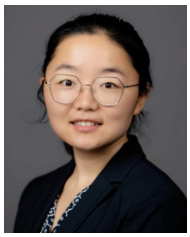
A. Agrawal, and A. N. Choudhary, MuSES: Multilingual sentiment elicitation system for social media data, *IEEE Intelligent Systems*, vol. 29, no. 4, pp. 34–42, 2014.

[119] L. Y. Liu and M. Han, Privacy and security issues in the 5g-enabled internet of things, in *5G-Enabled Internet of Things*, Y. L. Wu, H. J. Huang, C. X. Wang, and Y. Pan, eds. Boca Raton, FL, USA: CRC Press, 2019, pp. 241–268.

[120] Cisco annual internet report (2018–2023) white paper, https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html, 2020.

[121] S. Mattisson, Overview of 5g requirements and future wireless networks, presented at ESSCIRC 2017—43rd IEEE European Solid State Circuits Conf., Leuven, Belgium, 2017, pp. 1–6.

[122] IAB, Live video streaming: A global perspective, https://www.iab.com/insights/live-video-streaming-2018/, 2018.



Qixuan Hou is a master student in analytics at Georgia Institute of Technology. She received the BS degree from Georgia Institute of Technology in 2019. Her research interests are in the area of data science, with experience executing data-driven solutions.

Meng Han is an assistant professor at the College of Computing and Software Engineering, Kennesaw State University. He received the PhD degree in computer science from Georgia State University in 2017. He is currently an IEEE member and an IEEE COMSOC member. He has the unique research experiences built upon big data, cyber data, and social data with academic achievements of 6 book chapters/books, more than 20 first-authored and more than 20 co-authored publications in international journals and conferences, with 4 best paper awards and 2 best paper runner-up awards. His research interests include data-driven intelligence, AI security & privacy, and blockchain technologies.

Zhipeng Cai received the MS and PhD degrees from University of Alberta in 2004 and 2008, respectively. He is currently an associate professor at the Department of Computer Science, Georgia State University (GSU). Prior to joining GSU, he was a research faculty at the School of Electrical and Computer Engineering, Georgia Institute of Technology. His research areas focus on networking, big data, data security, and artificial intelligence. He is the recipient of an NSF CAREER Award.