Information Processing and Management 58 (2021) 102390

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches

Yuehua Zhao[a,b], Jingwei Da[a], Jiaqi Yan[a,*]

[a] School of Information Management, Nanjing University, No.163, Xianlin Road, Qixia District, Nanjing, China
[b] Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing, China

ARTICLE INFO

Keywords:
Health misinformation
Misinformation detection
Online health community

ABSTRACT

Curbing the diffusion of health misinformation on social media has long been a public concern since the spread of such misinformation can have adverse effects on public health. Previous studies mainly relied on linguistic features and textual features to detect online health-related misinformation. Based on the Elaboration Likelihood Model (ELM), this study proposed that the features of online health misinformation can be classified into two levels: central-level and peripheral-level. In this study, a novel health misinformation detection model was proposed which incorporated the central-level features (including topic features) and the peripheral-level features (including linguistic features, sentiment features, and user behavioral features). In addition, the following behavioral features were introduced to reflect the interaction characteristics of users: Discussion initiation, Interaction engagement, Influential scope, Relational mediation, and Informational independence. Due to the lack of a labeled dataset, we collected the dataset from a real online health community in order to provide a real scenario for data analysis. Four types of misinformation were identified through the coding analysis. The proposed model and its individual features were validated on the real-world dataset. The model correctly detected about 85% of the health misinformation. The results also suggested that behavioral features were more informative than linguistic features in detecting misinformation. The findings not only demonstrated the efficacy of behavioral features in health misinformation detection but also offered both methodological and theoretical contributions to misinformation detection from the perspective of integrating the features of messages as well as the features of message creators.

1. Introduction

Internet access drives information access. The number of users who utilize the Internet for health information seeking, ranging from healthy lifestyle advice to treatment and diseases, continues to grow (Chu et al., 2017). The number of users who utilize the Internet for health information seeking, ranging from healthy lifestyle advice to treatment and diseases, continues to grow (Chu et al., 2017). Pew Research Center's latest national survey reports that seven-in-ten (72%) adult internet users search online for a variety of health issues (Silver & Huang, 2020). In the Web 2.0 era, social media have flourished and increasingly influenced people's daily life and their health behaviors. With access to information shared on social media platforms, people find useful information more

* Corresponding author.
E-mail addresses: yuehua@nju.edu.cn (Y. Zhao), m18805156308@163.com (J. Da), jiaqiyan@nju.edu.cn (J. Yan).

https://doi.org/10.1016/j.ipm.2020.102390
Received 28 January 2020; Received in revised form 9 September 2020; Accepted 15 September 2020
Available online 06 October 2020

effectively and personally than traditional information retrieval through search engines (Zhao, Zhang & Wu, 2019).In recent years, the number of online health communities has increased rapidly as more patients seek to access alternate sources of health information as well as to connect with other patients with the same or similar disease. The large number of such communities is a testament to their popularity among health consumers (Jadad, Enkin, Glouberman, Groff & Stern, 2006). Different from traditional ways of information searching, social media offer health information seekers access not only to the information on the platforms, but also to other users (Zhao & Zhang, 2017). The basic idea behind so-called peer-to-peer healthcare is consulting about health issues with other peers. Social media connect patients with others who have the same concerns.

However, due to the increasing popularity of social media, the Internet has become an ideal breeding ground for the spread of fake news, misleading information, fake reviews, rumors., etc. (Zhang & Ghorbani, 2019). Social media enable unreliable sources to spread large amounts of unverified information among people speedily and wildly (Qazvinian, Rosengren, Radev & Mei, 2011). It is thus crucial to design models that can effectively detect the online misleading information.

Misinformation is defined as “the factually incorrect information that is not backed up with evidence” (Bode &Vraga, 2015). Misinformation on social media has become an urgent and vital issue, and even more so in the health-related fields (A. Ghenai & Mejova, 2018). Health information obtained from social media, including online health communities, may impact the health care outcomes of patients (Zhao & Zhang, 2017). Concerns about health misinformation have risen with the rise of health information seeking on social media platforms. The lack of gatekeepers in online communities discussions encourage the spread and reinforcement of health misinformation (Bode & Vraga, 2018). The extant literature has largely focused on the detection of fake reviews and fake news; however, there is a lack of comprehensive theory-driven framework in the literature that was designed for health misinformation detection, especially in the online health community context. Given the vast volume of health misinformation spreading in online health communities, there is a need to design an effective model to achieve the automatic detection of health misinformation in the health community context. Accordingly, the first research question leading this study was: How can we build an effective model to automatically detect the health misinformation in online health communities? To answer this questions, we identified different types of misinformation appearing in online health communities, and built a series of detection models based on different machine learning techniques and a real-world dataset to verify the performance of the models.

In order to build models and methods to identify the online misinformation, many researchers are dedicated to capturing the features of misinformation. Misinformation on social media can be seen as messages that are posted to persuade other users. To reveal the effective features to detect misinformation in online health communities, we drew on the Elaboration Likelihood Model (ELM) which can assist in comprehending how misinformation in online health communities persuade the users. ELM postulates that users build the attitude towards a message either the central or the peripheral route (Petty & Cacioppo, 1986). In the central route, users scrutinize the quality and strength of the information; whereas in the peripheral route users care more about superficial factors like source reputation, visual appeal, and presentation (Ebnali & Kian, 2020). In addition, well-supported results in social media studies have indicated that beyond the content of the message, some secondary information (such as the number of likes and stars) substantially increased the validity and trustworthiness of messages. This was observed especially in online health communities (Ebnali & Kian, 2020). Therefore, based on ELM, we proposed that the features of online health misinformation can be classified into two levels: central-level and peripheral-level. The central-level features of the messages persuade the users based on the content of the messages, while peripheral-level features convince the users through the impact of the message creators. Moreover, Castillo, Mendoza and Poblete (2011) and Qazvinian et al. (2011) have suggested that the best features to identify the misinformation on social media were those that looked into the user, message and topic features. According to these findings, building a health misinformation detection model that integrates central-level features and peripheral-level features needed further investigations. Therefore, this study's second research question was: What are the effective features to detect the health misinformation in online health communities? To answer this question, we proposed using central-level features (including topic features) and peripheral-level features (including linguistic features, sentiment features, and user behavioral features), and assessing their power in automatically distinguishing the health misinformation from the legitimate information in the online health community settings. We also investigated the discrimination power of different features using different machine learning techniques.

The rest of this paper is organized as follows. The Related Work section, describes the severity of the online health misinformation issues, the approaches that have been applied to track fake reviews and online rumors, and prior works detecting online health misinformation. Then, the Methodology section introduces the dataset and the methods dealing with the feature extraction and the classification model construction. The Experimental Results section reports the experimental results for the proposed model and the Discussion section discusses these findings. Finally, the Conclusion section provides the concluding remarks and directions for future work.

2. Related work

2.1. Health misinformation on social media

Misinformation has been widely studied in social science, especially in the fields of politics and mass communication. Health-related misinformation has long been a concern to the public (Li et al., 2019, July). Chou, Oh and Klein (2018) defined the health misinformation as "A health-related claim of fact that is currently false due to a lack of scientific evidence". In the context of online health communities, health misinformation might also include advertisements and other promotional messages. Recently, health misinformation has gained increasing attention because it is one of the most frequently transmitted information on social media and can have adverse effect on public health.

Social media's unrestricted access has made it an important and popular means by which individuals can access and discuss health information. A growing number of individuals, especially those with chronic diseases such as cancer, and their families, have turned to social media to find and share a variety of disease-related information. This includes finding information about disease prevention and treatment, sharing their experiences, and gaining social support to cope with the disease and manage emotions (Chen, Wang & Peng, 2018).

Nevertheless, individuals take great risks when using web-based resources because health information on social media is not always accurate (Chen et al., 2018). First, although medical professionals and traditional portals contribute to the health information provided on social media, people generate and disseminate more information based on their first-hand disease experience (Song et al., 2016). Moreover, compared to general online platforms, people are more eager to share their first-hand treatment experience (which may not be accurate) on the social media community to make a social contribution and improve their social status among friends (Fichman, Kohli & Krishnan, 2011). Second, compared to health information on websites, health information in social media posts is often oversimplified and may ignore small but important details. In addition, due to an overload of information on social media, people may not have the resources, knowledge and expertise to evaluate the accuracy of web-based disease-related information and identify informative and trustworthy information on social media (Vogel, 2017). Third, some stakeholders have an interest in creating fake comments to criticize conventional therapies in order to advertise their therapies or products in social media (Waszak, Kasprzycka-Waszak, & Kubanek,2018). There is also misinformation that is posted to promote health-related websites, blogs, or content on social media platforms (Zhang, Zhou, Kehoe & Kilic, 2016; Ott, Choi, Cardie, & Hancock, 2011). The promotional misinformation on social media might affect users' judgments or decisions about disease or treatments. Furthermore, the spread of health misinformation on social media is an echo chamber effect. Social media connects like-minded people to a closed network where people share similar content. This magnifies the risk of misinformation (Brady, Kelly & Stein, 2017; Chou et al., 2018).

Given the prevalence of social media and the serious consequences caused by health misinformation on social media, it is necessary to develop a model to automatically detect health misinformation.

2.2. Detection of fake news and fake reviews based on the ELM

Elaboration Likelihood Model (ELM) of persuasion is well-studied in communication and information processing. According to Elaboration Likelihood Model (ELM), there are two routes in which information may be conveyed to persuade people to accept information as true, namely central and peripheral routes (Petty & Cacioppo, 1986). The central route requires the recipients to make a considerable cognitive effort to systematically evaluate the information received, while the peripheral route relies on heuristic cues such as the source credibility rather than the actual content of the message (Singh, Ghosh & Sonagara, 2020). Based on the ELM, recipients can be convinced to consider a piece of information as true through either the central or peripheral route (Singh et al., 2020).

The ELM has been widely applied in information system research in different contexts. The ELM has been applied in detecting fake news and fake reviews by a few recent studies as summarized in Table 1 (Horne & Adali, 2017; Janze & Risius, 2017; Lee, Ham, Yang & Koo, 2018; Osatuyi & Hughes, 2018; Singh et al., 2020). Drawing on the ELM and previous works in the realm of user-generated content and social psychology, Janze and Risius (2017) developed a machine learning model incorporating cognitive, visual, affective, and behavioral cues to detect fake news shared on social media platforms. Horne & Adali (2017) considered the ELM as a theory to explain the spread of fake news and analyzed the difference of fake news and real news from the perspectives of title, complexity, language, and style of content. The authors concluded that real news persuades users through sound arguments while the persuasion of fake news is realized through heuristics. Lee, Ham, Yang & Koo (2018) adopted the ELM to explain the different patterns existing for authentic and fake reviews. Osatuyi and Hughes (2018) used the ELM as a theoretical framework to explain the strategies used by real and fake news platforms to present news. The authors hypothesized that fake news sites favor the peripheral

Table 1
Summary of research on fake news and fake reviews detection based on ELM.

Study	Subject	Features		Domain
		Central features	Peripheral features	
Janze and Risius (2017)	Fake news	—	Cognitive features, Affective features, Behavioral features, Visual features	Political news
Horne and Adali (2017)	Fake news	—	Stylistic features, Complexity features, Psychological features	Political news
Lee, Ham, Yang & Koo (2018)	Fake review	—	Self-disclosure, Number of friends, Linguistic features, Behavioral features, raw coverage, consistency, unique coverage	Restaurant reviews
Osatuyi and Hughes (2018)	Fake news platforms	—	Negative valences of full texts, Negative valences of vocabularies, Information amounts, News source variances	Online news posts
Singh et al. (2020)	Fake news	Content features	Organization features, Emotions features, Manipulation features	Online news stories

route by providing less information but more negative affective cues. Consistent with the peripheral route characteristics posited in the ELM, Osatuyi and Hughes (2018) revealed that fake news tends to provide a lower amount of information in order to facilitate the quicker processing of the information. Drawing on key theories of information processing and presentation, Singh et al. (2020) identified multiple text and visual features that are associated with fake or credible news and developed a multimodal approach combining these features to detect fake news online.

2.3. Health misinformation detection

Recently, many studies have focused on the characteristics of health misinformation on social media (as shown in Table 2). Content analysis has been widely used to identify the information and communication characteristics of health misinformation on social media (Li, 2019). As early as 2013, Syed-Abdul et al. (2013) studied misleading information about anorexia based on videos on YouTube. In 140 videos related to anorexia, they found 29.3% contain pre-anorexia information which promoted anorexia as a fashion and a source of beauty by sharing tips and methods for becoming and remaining anorexic. In 2016, Bessi et al. (2016) analyzed verified (science news) and unverified (conspiracy news) content in Facebook, and provided important insights into the systematic structure of the possible dissemination of erroneous information through comprehensive quantitative analysis. According to the judgment of rumor websites and health experts, Li, Zhang and Wang (2017) classified health information from WeChat as true or fake health information, and used Chi-square to identify significant features of fake health information in WeChat. They also developed a list of significant features of health misinformation. Waszak, Kasprzycka-Waszak and Kubanek (2018) analyzed the spread of fake news on social media and found 40% of news contained text classified as fake news. The most fallacious content was about vaccines. Chen et al. (2018) investigated the nature and spread of misinformation related to gynecological cancers on Weibo. The results showed about 30% of them contained misinformation, and the tweets about cancer treatment were found to contain a higher percentage of misinformation than those related to prevention. However, the false information related to prevention was much wider and deeper than the real information on social media.

In February 2016, the World Health Organization declared the Zika outbreak a Public Health Emergency of International Concern (Ghenai & Mejova, 2017). Since then, many studies have focused on spreading and analyzing of misinformation about Zika on social media. Sicilia, Giudice, Pei, Pechenizkiy & Soda (2017, 2018) developed a novel rumor detection system that focused on health-related rumors about Zika on Twitter and constructed a new features subset, which included influence potential and network characteristics features. The proposed method was tested on real datasets using Random Forest classifier, with an accuracy of 71.4%. Ghenai and Mejova (2017) constructed a tool to track health misinformation about Zika on Twitter. This tool incorporated health professionals, crowdsourcing, and machine learning. Sommariva, Vamos, Mantzarlis, Đào and Martinez Tyson (2018) studied the spread of health rumors and verified information on social networking sites (SNSs) using the Zika virus as a case study. They classified Zika-related news to three categories: verified news stories, rumors, and satire or parody. Rumors included three subcategories: misleading content, false connection and fabricated content.

Some studies applied features and techniques that achieved success in fake reviews or fake news identification to detect health misinformation. Purnomo, Sumpeno, Setiawan and Purwitasari (2017) used Random Forest classifier to classify fake news in medical archives of snopes.com with an accuracy of 95.95%. Kinsora, Barron, Mei, and Vydiswaran (2017) created a labeled dataset of misinformation and non-misinformation based on posted questions and comments from MedHelp, a health discussion forum. The dataset contains 14.2% misinformation comments. They identified the nine most descriptive features related to classification and obtained a classification accuracy of 90.1% using Random Forest classifier. A. Ghenai and Mejova (2018) aimed to automatically identify Twitter users who were prone to spread misinformation about cancer treatments. A logistic regression model was trained based on linguistic features extracted from LIWC (Linguistic Inquiry and Word Count) lexicon and text readability, with an accuracy of more than 90%. In addition, they found that readability, tentative language and avoidance of personal pronouns were important language markers associated with the propensity of posting cancer treatments misinformation.

Liu, Zhang, Susarla & Padman (2018) developed a logistic regression model to classify the level of medical knowledge (high-and low-medical knowledge videos) encoded in the collected videos from YouTube based on text-based and image-based medical knowledge. These were extracted from video descriptions and video frames using Bidirectional Long Short-Term Memory (BLSTM) and CNN. Li (2019) used topic analysis to reveal the frequent words and common topics based on a dataset collected from two healthcare-related websites, and built a linear regression models to detect false medical/healthcare information. Liu, Yu, Wu, Qing and Peng (2019) analyzed common characteristics of true and fake health-related information on Chinese online social media, and developed a health-related misinformation detection framework. It combined a feature-based method and a text-based method. Hou, Perez-Rosas, Loeb, and Mihalcea (2019) focused on the automatic detection of misinformation in YouTube videos. Using linguistic, acoustic and user-engagement features, they developed a classification model for health misinformation detection with accuracies of up to 74% (Table 2).

Table 2
Summary of research on health misinformation detection.

Study	Title	Method	Data Source	Disease
Syed-Abdul et al. (2013)	Misleading Health-Related Information Promoted Through Video-Based Social Media: Anorexia on YouTube	Content analysis	YouTube	Anorexia
Bessi et al. (2016)	Homophily and Polarization in the Age of Misinformation	Quantitative analysis	Facebook	Autism
Li et al. (2017)	Fake vs. Real Health Information in Social Media in China	Content analysis	We Chat	–
(Kinsora et al., 2017, August)	Creating a Labeled Dataset for Medical Misinformation in Health Forums	Design science	MedHelp	–
Purnomo et al. (2017)	Keynote speaker II: biomedical engineering research in the social network analysis era: stance classification for analysis of hoax medical news in social media	Design science	Snopes.com	–
(Sicilia et al., 2017, November)	Health-Related Rumour Detection On Twitter	Design science	Twitter	Zika
Ghenai and Mejova (2017)	Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter	Design science	Twitter	Zika
Chen et al. (2018)	Nature and Diffusion of Gynecologic Cancer–Related Misinformation on Social Media: Analysis of Tweets	Content analysis	Weibo	Gynecologic cancers
Sommariva et al. (2018)	Spreading the (Fake) News: Exploring Health Messages on Social Media and the Implications for Health Professionals Using a Case Study	Content analysis	Facebook, LinkedIn, Twitter, Pinterest, Google Plus	Zika
Waszak et al. (2018)	The Spread of Medical Fake News in Social Media–The Pilot Quantitative Study Health	Content analysis	Facebook, LinkedIn, Twitter, Pinterest	Cancer, neoplasm, heart attack, stroke, hypertension, diabetes, vaccinations, HIV, AIDS
(Sicilia et al. (2018))	Twitter Rumour Detection in the Health Domain	Design science	Twitter	Zika
A. Ghenai and Mejova (2018)	Fake Cures: User-centric Modeling of Health Misinformation in Social Media	Content analysis Design science	Twitter	Cancer
Liu, Zhang, Susarla & Padman (2018)	YouTube for Patient Education: A Deep Learning Approach for Understanding Medical Knowledge from User-Generated Videos	Design science	YouTube	–
(Li, 2019, July)	Detecting False Information in Medical and Healthcare Domains: A Text Mining Approach	Design science	Snopes.com PolitiFact.com	–
Liu et al. (2019)	Analysis and Detection of Health-Related Misinformation on Chinese Social Media	Design science	KepuChina, Weibo, WeChat, Guokr, Dingxiangyuan	–
Hou et al. (2019))	Towards Automatic Detection of Misinformation in Online Medical Videos	Design science	YouTube	Prostate cancer

In summary, there are two potential gaps in the existing literature that we attempted to address in this study. First, the extant literature has largely focused on the detection of fake reviews and fake news; however, there is a lack of detection framework guided by a comprehensive theory to identify health misinformation on social media, especially in the online health communities. Second, although scholarly attention has been drawn to the features of health misinformation on social media; previous studies mainly focused on linguistic features. Little is known about whether integrating user behavioral features with linguistic features, sentiment features, and topic features, could effectively distinguish the misinformation from the legitimate information in online health communities. Accordingly, to narrow the aforementioned gaps, this study combined these features to build the detection model targeting the misinformation spreading in the online health community context.

3. Methodology

Fig. 1 presents the data collection and data analysis process in this study and the pipeline of the proposed health misinformation detection model. The detailed procedures of the data analysis are described in the following sections.

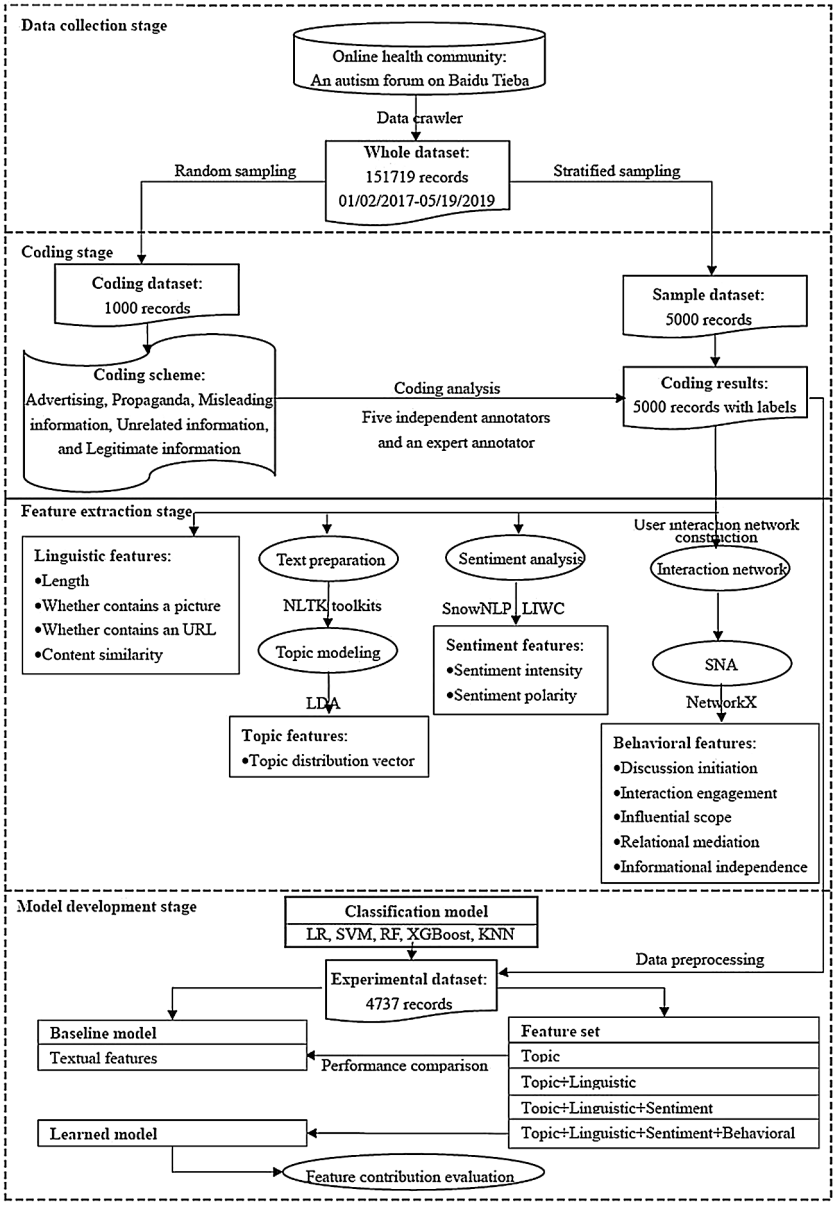

Fig. 1. Data collection and data analysis process.

Table 3
Descriptive statistics of the collected data.

	Count
Total number of threads	8097
Total number of replies	67,906
Total number of replies to reply	75,716
Average number of replies	8.4
Average number of replies to reply	1.1
Total number of involved users	13,193

3.1. Data collection

To label heath misinformation in online health communities for supervised learning, we selected Zibizheng Ba ("autism forum" in Chinese) on Baidu Tieba (www.tieba.baidu.com) as our data source. The Baidu Tieba claims to be one of the largest interest-based discussion platforms in China, where Internet users can create topic-based discussion forums on the platform, share information, and make friends with other users. It covers a wide range of topics, including society, education, health, entertainment, games, etc. Currently, there have been 1.5 billion register users and 22 million forums established on Baidu Tieba. Because the posts on Baidu Tieba are indexed by Baidu (www.baidu.com), China's most popular search engine, misinformation posted on Baidu Tieba can be easily found by the public when people search for related information through the search engine.

Zibizheng Ba, an autism forum, is under the health category on the Baidu Tieba, with 54,650 followed users and 465,323 posted threads. A python-based web crawler was developed to collect data from the forum. Any registered users could post threads and reply to existing threads. Each thread included the following information: the user who created the thread, the content of the thread, the specific time when it was posted, and the number of replies it received. According to the mechanism of the Baidu Tieba, users can not only respond to any original threads but also to replies to the threads (a "reply to a reply" is named as Louzhonglou in Chinese). Each reply to the original thread possesses the following components: the user who made the reply, the content of the reply, the specific time when the reply was made, and the number of replies it received. Content of the threads and the replies, may contain text(s), URL (s), photo(s), etc.

Data collection from the autism forum was performed on May 19, 2019. The data collection covered January 2, 2017 to May 19, 2019. All of the threads, replies, and replies to replies were captured by the web crawler. After the data preparation process, records with no content were eliminated from the dataset. A unique ID was assigned to identify each record. As a result, there were collected 151,719 records, including 8097 threads, 67,906 replies, and 75,716 replies to replies. Table 3 presents the basic descriptions of the collected data. For each record, the content of the thread/reply/reply to reply, the user who posted the record and the time stamp were extracted.

3.2. Data sampling

To train the health misinformation detection model, 5000 records were sampled from the whole dataset by stratification according to the three types of records (i.e. thread, reply, reply to reply) using stratified sampling methods. Therefore, the constituent types of the records (i.e. thread, reply, reply to reply) in the sample dataset were consistent with the composition of the whole dataset (as shown in Table 4).

Table 4
Descriptions of whole dataset and sample dataset.

	Total record	Sampled record	% in sample dataset
# Threads	8097	266	5%
# Replies	67,906	2238	45%
# Reply to replies	75,716	2496	50%
Total	151,719	5000	100%

Table 5
Summary of previous research on classification of online misinformation.

Study	Subject	Category
Waszak et al. (2018)	Fake medical news	Fabricated news Manipulated news Advertisement news
Sommariva et al. (2018)	Rumors	Misleading content False connection Fabricated content
Wardle (2017)	Fake news	Satire or parody Misleading content Imposter content Fabricated content False connection False context Manipulated content
Tandoc et al. (2018)	Fake news	News Satire News Parody News Fabrication Photo Manipulation Advertising and Public Relations Propaganda

3.3. Coding analysis

To train the misinformation detection model, we first needed to identify the types of misinformation appearing in online health communities. Previous studies had shed light into the detection of fake news and rumors and categorized fake news and rumors into different types. Table 5 summarizes previous research on the classification of fake news and rumors.

Coding analysis is one of the significant qualitative data analysis methods applying to organize and make sense of textual data (Basit, 2003). For the collaborative coding process, Richards and Hemphill (2017) described it in the following 6 progressive steps: (a) preliminary organization and planning, (b) open and axial coding, (c) development of a preliminary codebook, (d) pilot testing the codebook, (e) final coding process, (f) review the codebook and finalize the themes. In this study, we conducted the labeling process following the above steps. The coding analysis was performed on a coding dataset which contained 1000 records randomly selected out of the whole dataset using the random sampling algorithm. In the first stage of coding analysis, we processed the coding dataset by breaking down, examining, comparing, conceptualizing, and categorizing the records. Adapting from the previous studies on categorizing the fake news and the rumors, the coding scheme for classifying health misinformation in online health communities developed in this study is shown in Table 6.

In the second stage of the coding analysis, five coders were recruited to annotate the sample dataset of 5000 records. After being trained on the definitions and examples of the categories of health misinformation as shown in the coding scheme, five human annotators were asked to read each post and decide whether the post belonged to one of the categories of misinformation or ordinary

Table 6
Coding scheme for categorizing the health misinformation in online health communities.

Category	Definition	Example
Advertising	Advertising materials about treatments, therapies, training organizations, personal trainers, products, etc.	"Dongguan Smart Rabbit Rehabilitation Education Center was founded in Zhangmutou in 2005. Since its development, there are now seven campuses in Guangdong Province, which are located in Xintang, Guangzhou, Shenzhen, Dongguan, Zhangmutou, Dongcheng, Changping and Guancheng, with a total area of more than 10,000 m^2s. There are more than 150 faculty members.", "I am an autism special education teacher. I just want to bring the light to the autistic children. Contact me via phone or WeChat 159********"
Propaganda	Posts which are created by users to attract attentions to certain products, websites, or online content or groups on other social media platforms	"Please go to 'I want to finish it' to talk: www.***.com. It is better to communicate more. Please don't delete it.", "I found a good lesson about how to cure autistic babies. Scan the QR code I sent. You can listen to it."
Misleading information	Posts that are currently false or misleading due to the lack of scientific evidences	"Can you come to Shandong to take a photo? You will be able to see the wonders of the world! Severe autism was cured after acupuncture treatment. The patient was cured at 52 years old, and is now 61 years old. The world's first autism cured by a medical method!", "Copy the scriptures! Tested useful! Thank the Buddha with a sincere heart. Copy the scriptures with ease. Try it. I copy the scriptures every day. The condition of my son has been much better. Try it yourself. Namo Amitabha."
Unrelated information	Posts that are not clearly related to the community	"Watch movie here", "I envy you people who have stories. Unlike me, handsome runs through my life."

Table 7
Results from coding analysis.

Category	Count	Percentage
Advertising	259	5.18%
Propaganda	104	2.08%
Misleading information	208	4.16%
Unrelated information	189	3.78%
Legitimate information	4240	84.8%
Total	5000	100%

information sharing. Each coder annotated 3000 records out of the sample dataset. Each record was judged independently by three of the five coders. The assignments of the records to the coders were decided by the random sampling algorithm.

To assess the inter-coder reliability of the coding results generated by the five coders, Krippendorff's alpha interval statistic was adopted since it can be applied to any number of observers, categories, scale values or measures, and allows for missing data (Wu & Liu, 2019). The resultant alpha value was 0.80 according to Krippendorff's reliability formula (Krippendorff, 2012). This meant that the five coders achieved a substantial agreement in the category assignments. A majority-vote method consolidated the coding results. For each record, the majority vote (2 to 3) among the three coders was used to determine the category to which the record belonged. For the 32 records that the three coders annotated differently, all five coders discussed and reached the final decisions of the annotation together. Considering the professionalism of the misleading information, an expert annotator, an autism specialized doctor, was consulted to validate the records that were labeled as the misleading information. Table 7 lists the final outcome of the coding analysis.

3.4. Linguistic features

Previous studies had widely adopted linguistic features as a type of significant characteristics to track the fake reviews and rumors online (Ghenai & Mejova, 2018). The presence of suspicious tokens, such as URLs, hashtags, and photos in social communication data have been identified as good features for deceptive information detections (Zhang & Ghorbani, 2019). Therefore, in this study, we introduced four linguistic features to measure the characteristics of the content of the records: the number of words in the record, whether or not the record contains picture(s), whether or not the record contains URL(s), and content similarity.3.4.1 Text preparation

Text preparation process refers to a series of steps to prepare the raw text before more in-depth natural language processing, e.g. topic model training. Text preparation commonly consists of the selection, cleansing and preprocessing of text (Liddy, 2000). In this study, the text preparation process was operated by Python scripts using NLTK toolkits. Word segmentation is a significant process for the Chinese natural language processing since there is no clear word divider between words written Chinese. In this study, word segmentation was performed with one of the most widely used word segmentation tools for Chinese: Jieba (Sun, 2019). After removing all punctuation, stop words were filtered based on the stop words list produced by the Harbin Institute of Technology. Finally, all text were cleansed and checked manually to insure the accuracy of the subsequent analyses.

3.4.2. Content similarity feature

Content similarity determines how similar two pieces of text are. As crafting a new thread/post every time is time consuming, users who create misinformation are likely to disseminate the same or similar messages in the community. Therefore, content similarity and duplicity have been used to capture the online fake reviews (Lai, Xu, Lau, Li & Jing, 2010; Mukherjee et al., 2013).

Classical approaches, such as Jaccard similarity measurement, measure the content similarity based on the content overlap between documents. Cosine similarity measures the similarity between two vectors by comparing the cosine of the angle between the two vectors and determining whether the two vectors point in roughly the same direction. Cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance due to the size of the document, they may still be oriented close together (Han, Kamber, & Pei, 2012). The smaller the angle between the two vectors, the higher the cosine similarity between the two documents. We used the adjusted cosine similarity formula because the cosine similarity only focused on the similarity of directions and ignored the similarity of values (Sarwar, 2001).

3.4.3. Adjusted cosine similarity

To use adjusted cosine similarity, we needed to convert records into vectors. From each record we derived a vector. The set of records in our dataset then were viewed as a set of vectors in a vector space. Each term appearing in the records had its own axis. For this we represented records as bag-of-words so that each record was a sparse vector. In addition, we transferred the original bag-of-words matrix with TF (term frequency) to the matrix with TF-IDF (term frequency- inverse document frequency). The content similarity between any pairs of records (R_i and R_j) was defined as:

$$\text{adjusted\_cosine\_similarity}\ (R_i, R_j) = \frac{\sum_{c\in C}(R_{i,c} - \overline{R_i})(R_{j,c} - \overline{R_j})}{\sqrt{\sum_{c\in C}(R_{i,c} - \overline{R_i})^2}\sqrt{\sum_{c\in C}(R_{j,c} - \overline{R_j})^2}}$$

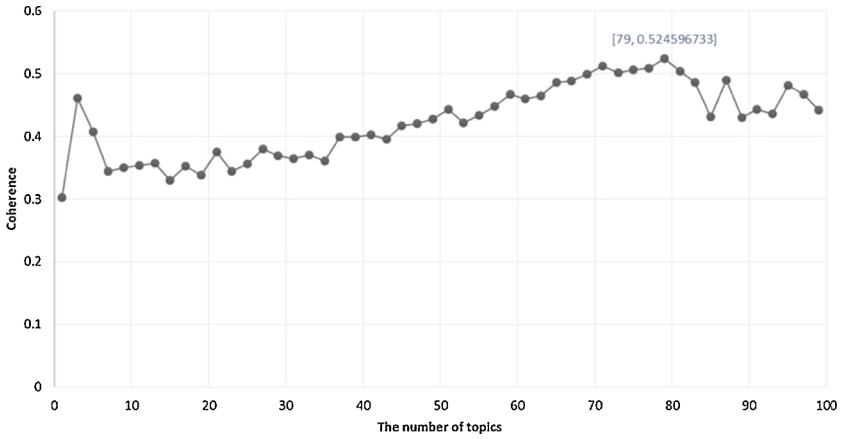

Fig. 2. The coherence for different settings of the number of topics.

Here C referred to the dimension of R_i and R_j. $R_{i,c}$ and $R_{j,c}$ referred to values of each dimension of R_i and R_j,respectively.

The content similarity feature of a given record (R_i) in the dataset, assuming the total number of records in the dataset is n, was defined as the average of the adjusted cosine similarity between the given record (R_i) and every other records in the dataset.

$$\text{content\_similarity}(R_i) = \frac{\sum_{j=1}^{n} |\text{adjusted\_cosine\_similarity}(R_i, R_j)|}{n}$$

3.5. Topic features

Topic modeling provides a powerful tool to identify latent content patterns from content. It views documents as mixtures of probabilistic topics and helps discover a set of topics that appear in a collection of documents (Griffiths & Steyvers, 2004). In addition to revealing the topics based on the text, applying topic modeling methods can achieve a vast dimensionality reduction of the textual feature extraction.

Griffiths and Steyvers (2004) proposed Latent Dirichlet Allocation (LDA) as a particular generative model for topic discovery. We implemented an LDA model in Python with the Sklearn package using the Gibbs Sampling inference method. The hyperparameters α and β control the amount of smoothing in the model estimation process. The hyperparameters α was set to 50/k (k is the number of topics) while β equals 0.01. The number of iterations was set as 500.

The pre-specified numbers of topics influence the performance of the topic model training. However, there exists no universally agreed formula to decide the optimal number of topics for a dataset (Savov, Jatowt & Nielek, 2020). Topic coherence has been proposed as a genuine assessment method for topic modeling techniques (Rosner, Hinneburg, Röder, Nettling, & Both, 2014). Topic coherence measures a single topic by the degree of semantic similarity between top terms of the topic (Rosner, Hinneburg, Röder, Nettling, & Both, 2014). To evaluate the performance of a topic model, the topic coherence score is calculated by the average of pairwise term similarities formed by the top terms of the generated topics. The number of topics (k) that marks the end of a growth of topic coherence usually presents interpretable topics and offers a meaningful topic model.

To find the optimal number of topics, a series of LDA models with different values of k were tested. We built 100 different models, each one with an incremental value of k (from 1 to 100), and for each of these we calculated the coherence score. Fig. 2 shows the distribution of the coherence scores given different values of k. As shown in Fig. 2, the coherence scores gradually increased with the number of topics, with a decline between 3 and 7. The coherence scores reached the highest (0.525) when the number of topics was set as 79. After producing the appropriate LDA models, the generated record-topic probability vectors were established as the topic features of the records.

3.6. Behavioral features

Previous efforts in search of features for online fake reviews and fake news detection have looked into the behavioral features of users. It has been widely recognized that deception, such as creating or disseminating fake news, is a type of strategic behavior (Wu & Liu, 2019). Thus, deceptive users probably behave differently in the communication network. To potentially enhance the performance of the detection model, in addition to the basic descriptive statistics of user behaviors (e.g. number of messages the user created), we introduced the use of node-level centrality measures as the features of interaction behavior in the user interaction network.

Social network analysis is paramount to understanding the social behavior of social network members. In the view of social network analysis, online health communities can be seen as social networks wherein users are nodes with the relationships between users represented as edges of the network. Implicit relationships between users are generated when users interact with each other

(such as replying to a thread, replying to a reply, and so on) in the online communities. The implicit relationships represent users' interaction behavior and movements within the communities.

3.6.1. Creation of the related matrices

After all data in the community were collected, multiple matrices were generated. These matrices defined relationships among the involved users (actors) in terms of making replies and replying to replies. The generation of matrices is vital and crucial for social network analysis.

Reply Node-Node Matrix (RNNM):

$$RNNM = \begin{pmatrix} r_{11} & \cdots & r_{1a} \\ \vdots & r_{ij} & \vdots \\ r_{a1} & \cdots & r_{aa} \end{pmatrix}$$

Here a was the number of all users who were involved in the replying interactions. r_{ij} was a cell in the matrix and referred to the number of replies that actor i made to the threads from actor j.

Reply to Reply Node-Node Matrix (RRNNM):

$$RRNNM = \begin{pmatrix} rr_{11} & \cdots & rr_{1b} \\ \vdots & rr_{ij} & \vdots \\ rr_{b1} & \cdots & rr_{bb} \end{pmatrix}$$

Here b was the number of all users who were involved in the reply to reply interactions. rr_{ij} was a cell in the matrix and referred to the number of likes that actor i made to the posts from actor j.

The original matrices had to be normalized before these matrices were combined. In the normalization, the size of a normalized original matrix should equal the size of the final node-node matrix (FNNM); the order of the actors in the normalized original matrix should be the same as the order of the actors in the FNNM. The normalized matrices were symmetric. The sizes of normalized matrices were equal to the number of actors who were involved in the replying behavior and/or replying to reply behavior.

Normalized Reply Node-Node Matrix (NRNNM) was defined as:

$$NRNNM = \begin{pmatrix} r'_{11} & \cdots & r'_{1q} \\ \vdots & r'_{ij} & \vdots \\ r'_{q1} & \cdots & r'_{qq} \end{pmatrix}$$

Here q was the number of actors who were involved in any of the two types of connections. If an actor in NRNNM did not appear in the RNNM, it meant that it was a newly added actor. Then the cells in its corresponding row and column were set to 0 in the NRNNM. This was an important procedure for the normalization process. All other cells in the normalized matrix were the same value as the original matrix. The NRNNM and FNNM shared the same matrix structure for the purpose of the normalization.

Normalized Reply to Reply Node-Node Matrix (NRRNNM) was defined as:

$$NRRNNM = \begin{pmatrix} rr'_{11} & \cdots & rr'_{1q} \\ \vdots & rr'_{ij} & \vdots \\ rr'_{q1} & \cdots & rr'_{qq} \end{pmatrix}$$

As described above, if an actor in NRRNNM did not appear in the RRNNM, it meant that it was a newly added actor. Then the cells in its corresponding row and column were set to 0 in the NRRNNM. All other cells in the normalized matrix were the same value as the original matrix.

Finally, the FNNM was created based on the above two matrices after the normalization process. Thereby the interaction network was constructed based on the FNNM in which each row/column represented a distinct node in the network, while each cell represented the strength of the connection between a pair of two nodes.

The final node-node matrix (FNNM) was defined as:

$$FNNM = NRNNM + NRRNNM = \begin{pmatrix} r'_{11} + rr'_{11} & \cdots & r'_{1q} + rr'_{1q} \\ \vdots & r'_{ij} + rr'_{ij} & \vdots \\ r'_{q1} + rr'_{q1} & \cdots & r'_{qq} + rr'_{qq} \end{pmatrix}$$

3.6.2. Centrality features

Over the past years, a number of centrality measures have been proposed by sociologists to detect the structural characteristics of actors in a network. The centrality indicators were designed to identify the "core actors" from different perspectives. Three node-level centrality measures (i.e. degree centrality, betweenness centrality, closeness centrality) were adopted to measure the interaction characteristics of users in the community.

Degree centrality refers to the number of connections incident upon a node. The degree centrality implies the potential communication ability of a certain actor. Actors with higher degree centrality have higher probability of receiving and transmitting the

Table 8
The descriptions of the behavioral features.

Behavioral feature	Measurement	Description
Discussion initiation	#Thread a user created	To reflect the activity of a user in terms of initiating new discussions
Interaction engagement	#Reply and #reply to reply a user created	To reflect the activity of a user in terms of interacting with other users
Influential scope	Degree centrality	To reflect the potential communication ability of a user
Relational mediation	Betweenness centrality	To assess the potential of a user for control of communication in the community
Informational independence	Closeness centrality	To assess the ability a user to instantly communicate with others without going through many intermediaries

information flows and thus can be considered to have influence over other actors in the network (Abraham, Hassanien, & Snášel, 2010). Freeman's (1978) betweenness centrality was based upon the frequency with which a point falls between pairs of other points on the shortest paths connecting them. Betweenness centrality can be used to assess the potential of an actor for control of communication in the knowledge flow network. Closeness centrality basically measures how close a node is located with respect to every other node in the network (Abraham, Hassanien, & Snášel, 2010). Actors with higher closeness are able to reach (or be reached by) more other nodes in the network through geodesic or shortest paths. An actor that is close to many others can instantly communicate and interact with others without going through many intermediaries (Makagon, McCowan & Mench, 2012).

3.6.3. Definitions of centrality features

Given a network N with n nodes, then the size of the network N is n. Here n_i, n_j, and n_k are three nodes in the network. The total connections of n_i is $d(n_i)$. This means the degree of n_i is $d(n_i)$. The number of the maximum connections for n_i in the network is n-1 when n_i is directly connected to all other nodes. $C_D(n_i)$ refers to the degree centrality for the n_i, $C_B(n_i)$ refers to the betweenness centrality for the n_i, and $C_C(n_i)$ refers to the closeness centrality for the n_i. Based on the above assumptions, the following equations describe the definitions of the network-level measurements and actor-level measurements applied in this study.

The degree centrality for n_i was defined as:

$$C_D(n_i) = d(n_i)$$

Here g_{ijk} was the number of paths from node i to node k that pass through node j; while g_{ik} was the number of all the paths from node i to node k in the network. The betweenness centrality for n_j was defined as:

$$C_B(n_j) = \sum_{i<k} \frac{g_{ijk}}{g_{ik}}$$

Here $g(n_i, n_j)$ was the geodesic distance from node i to node j. The closeness centrality for n_i was defined as:

$$C_C(n_i) = \left[\sum_{j=1}^{n} g(n_i, n_j)\right]^{-1}$$

According to the above definitions, the user behavior features of each user were calculated based on the constructed interaction network. The social network analysis was performed with the NetworkX Hagberg, 2019, a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. The centrality features were extracted to distinguish the misinformation from the legitimate information in terms of the information creators' interaction characteristics and importance in the interaction network. Table 8 summarizes the descriptions of the proposed behavioral features.

3.7. Sentiment features

Sentiment analysis centers on identifying the viewpoint underlying the documents. One particular and common type of sentiment analysis is to detect the sentiment polarity, which is the overall orientation of a certain text as positive or negative (Lau, Wang, Man, Yuen & King, 2014). Since early 2000, sentiment analysis has been applied to the analysis of online movie reviews, the discovery of public sentiment, the prediction of elections, etc. Moreover, sentiment features have been recognized as effective features to detect the online rumors and fake reviews (Rout et al., 2017). Therefore, we proposed to implement the sentiment features in the health misinformation detection model with sentiment analysis techniques.

To improve the accuracy of the extraction of the sentiment features of the records, two widely applied sentiment analysis

Table 9
The definitions of the sentiment features.

Sentiment feature	Method	Description
Sentiment intensity	SnowNLP	SIV: [0, 1]
Corpus-based sentiment polarity	LIWC	Positive: PE > NE Negative: PE < NE Neutral: PE = NE

Table 10
Summary of the proposed feature set.

Level	Subset	Feature	Description
Peripheral-level features	Linguistic features	Length	Numerical
		Whether contains a picture	1 (Yes), 0 (No)
		Whether contains a URL	1 (Yes), 0 (No)
		Content similarity	Numerical
Central-level features	Topic features	Distribution of the probabilities of the record associated with the generated topics	Vector
Peripheral-level features	Sentiment features	Sentiment intensity	Numerical
		Corpus-based sentiment polarity	1 (Positive) − 1 (Negative)0 (Neutral)
	Behavioral features	Discussion initiation	Numerical
		Interaction engagement	Numerical
		Influential scope	Numerical
		Relational mediation	Numerical
		Informational independence	Numerical

techniques were adopted in this study: SnowNLP (2015) and LIWC (Linguistic Inquiry and Word Count). The SnowNLP is a Python module for sentiment analysis for simplified Chinese using the Naïve Bayes classifier method (Tseng et al., 2018). It has been regarded as an effective machine learning based sentiment analysis tool for Chinese text processing (He, Li, Yao, & Ding, 2020). The sentiment feature of a given text is measured by the sentiment intensity value (SIV) ranging from 0 to 1 where 0 represents the most negative sentiment while 1 represents the most positive sentiment.

The LIWC offers a corpus-based sentiment analysis approach by counting words in different emotion categories. Empirical results using LIWC demonstrate its ability to detect meaning in emotionality (Tausczik & Pennebaker, 2010). In addition, it has been employed to extract the sentiment features for the detection of misinformation in online medical videos (Hou, Perez-Rosas, Loeb & Mihalcea, 2019). The LIWC measures the positive and negative emotions of a given text by counting the number of positive and negative words appearing in the text and outputs the value of positive emotions (PE) and negative emotions (NE) of the text. In this study, as shown in Table 9, we define the corpus-based sentiment polarity of the records as follows: positive (PE > NE), negative (PE < NE), and neutral (PE = NE).

3.8. Summary of the features

As mentioned above, Table 10 summarizes the features proposed in this study to detect the health misinformation in online health communities. In total, there were 12 features that were grouped into the following four subsets: linguistic features, topic feature, sentiment features, and behavioral features. Drawing on the ELM, these features can be classified into two levels: central-level features (including topic features) and peripheral-level features (including linguistic features, sentiment features, and user behavioral features).

3.9. Classification methods

As we mentioned in the Related Works section, detection of health misinformation can be regarded as a binary or multi-classification problem, which marks the tested samples as misinformation or non-misinformation and etc. In this paper, we chose five classification algorithms, which are common classification models used in health misinformation detection and belong to different learning paradigms (Ghenai & Mejova, 2017; Hou et al., 2019; Liu et al., 2019; Purnomo et al., 2017). We included Logistic Regression (LR) a discriminant model, Support Vector Machine (SVM) as a kernel machine, Random Forest (RF) as an ensemble of trees, k-Nearest Neighbor (KNN) as a statistical classifier, and eXtreme Gradient Boosting (XGBoost) as a gradient boosting algorithm. They are explained as follows:

LR is a widely used classification model. It assumes the input data follow the Bernoulli distribution, and solves the parameters by maximum-likelihood function and gradient descent method to achieve the goal of classification of the data (Jindal & Liu, 2008).

SVM constructs a hyperplane or set of hyperplanes in a high-dimensional space, mapping inputs into high-dimensional feature spaces through a non-linear transformation, and then obtains the optimal linear classification plane according to the kernel function (Cortes & Vapnik, 1995; Sicilia et al., 2017). It has shown many advantages in solving small samples of nonlinear and high-dimensional pattern recognition. In this paper, we used SVM with the radial basis function (RBF) as the kernel function.

RF is a modern classification and regression technology, and it is also a combined self-learning technology (Breiman, 2001). RF integrates several single classifiers for classification, and determines the final classification by combining the classification results of multiple classifiers in order to achieve better performance than a single classifier. Most importantly, Random Forest models are robust to oversampling biases in highly skewed datasets such as ours.

KNN classification algorithm is one of the simplest methods in data mining classification technology. The “K” in KNN means k nearest neighbors, which means that each sample can be represented by its nearest k neighbors (Liu et al., 2019). KNN has no explicit learning process. The dataset has classification and eigenvalues in advance. After receiving new samples, they are processed directly by measuring the distance between different eigenvalues.

XGBoost model is a learning model based on tree integration, which is integrated by multiple CART (Classification And Regression Trees). By combining multiple tree models with lower classification accuracy, it builds a model with higher accuracy. The model will be continuously iteratively improved, which is called gradient promotion (Chen, He, Benesty, Khotilovich & Tang, 2015).

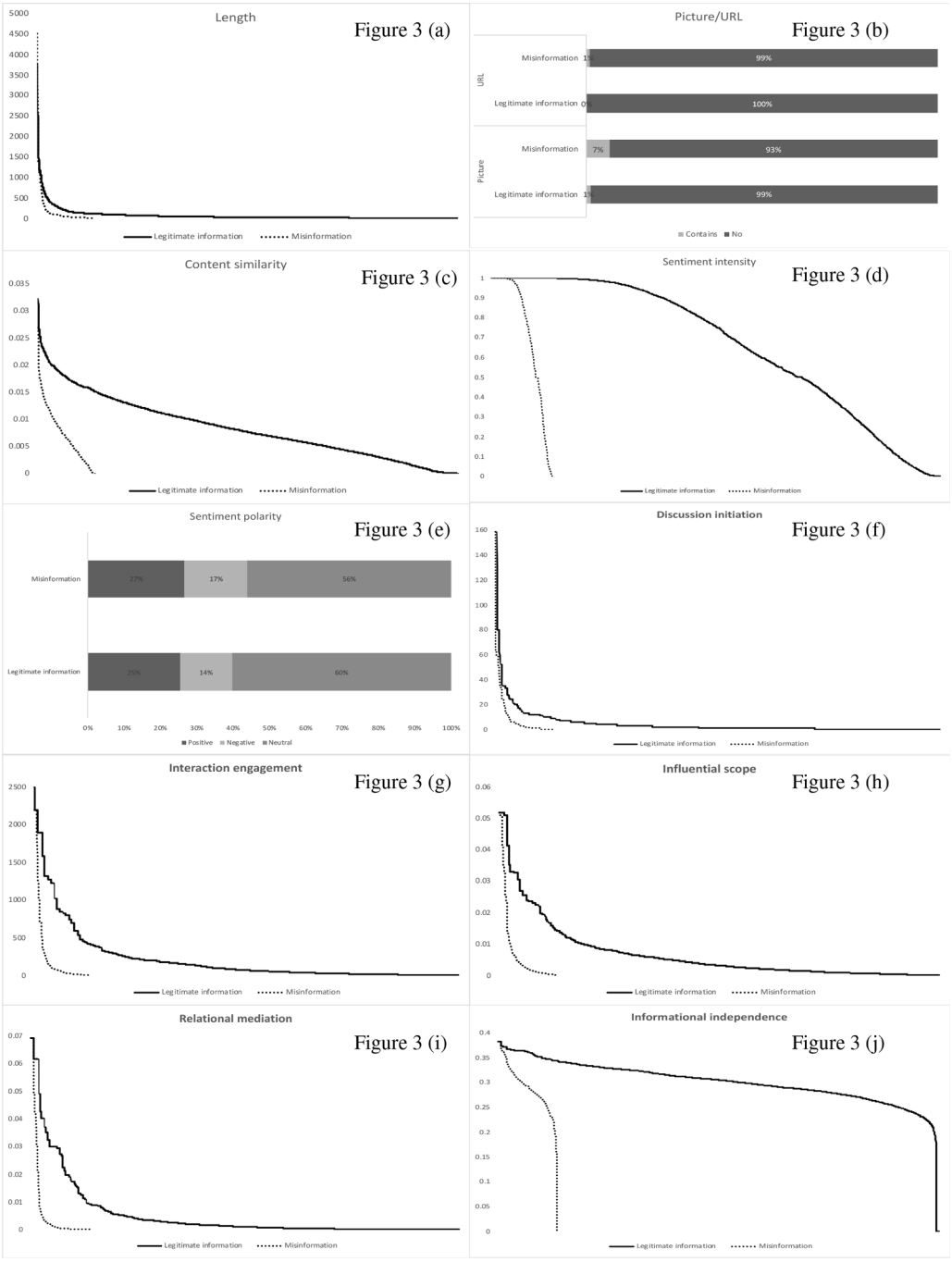

Fig. 3. The distributions of the legitimate information and the misinformation for the proposed features.

4. Experimental results

4.1. Feature extraction results

Fig. 3 shows the distribution of the different features for both legitimate information and misinformation in the sampled online health community. In the line charts, the solid line stands for the legitimate information while the dotted line stands for the misinformation. The plots in Fig. 3 suggest obvious differences between the two groups for the proposed features, thus potentially making them appropriate discriminators for the two types of information.

In terms of the extracted linguistic features, the misinformation, with 184 words on average in each record, tended to be much longer than the legitimate information that averaged 77 words. As shown in Fig. 3(b), 7% of the misinformation contained a picture (s) in the content while a mere 1% of the legitimate information contained a picture(s). In terms of content similarity, the legitimate information (average similarity equaled to 0.008) held higher similarity in comparison to the misinformation (average similarity equaled to 0.007). This might have been caused by selecting the mean value as the content similarity value of each record after calculating the adjusted cosine formula in the content similarity calculation.

Through the sentiment analyses, on average, the emotions expressed by the misinformation (0.72) were more positive than the legitimate information (0.66). Fig. 3(e) shows that 27% of the misinformation contained more positive words and 17% of the misinformation contained more negative words, while 25% of the legitimate information contained more positive words and 14% of the legitimate information contained more negative words.

Fig. 3(f-j) display the distributions of the behavioral features extracted from the misinformation and the legitimate information. The misinformation's authors created significantly both more threads (9.9 vs. 4.9) and more replies (255.6 vs. 214.4) than the legitimate information's authors. It implies that the users who created the misinformation tended to be more engaged in the discussions in the community. With regards to the centrality features, in comparison to the legitimate information's authors, the misinformation's authors achieved higher degree centralities (0.0076 vs. 0.0068) and higher betweenness centralities (0.0076 vs. 0.0054) whereas they attained lower closeness centralities (0.2848 vs. 0.3011). It suggests that the misinformation creators tended to be more influential compared to the users who posted the legitimate information in the community.

4.2. Model

4.2.1. Dataset

As mentioned in the Methodology section, we collected a dataset from a real online health community, and randomly selected 5000 records to be labeled as five categories: legitimate information, advertising, propaganda, misleading information, and unrelated information. Because of the skewed frequency distribution, our experiments were conducted as binary classification. Records with the label of “Legitimate information” were categorized as legitimate information and those with labels of “Advertising”, “Propaganda”, and “Misleading information” were categorized as misinformation. Because of text processing in the process of topic extraction, some records became empty after removing the stop words and we delete these records. The final experimental dataset contained 4737 records. This included 569 records which were categorized as misinformation and 4168 records which were labeled as legitimate information.

4.2.2. Classification methods

To validate the classification performance of our proposed feature set, we chose the following machine learning models to detect health misinformation: logistic regression, SVM, KNN, Random Forest and XGBoost. We used the implementation available in scikit-learn (Pedregosa et al., 2011). We did not adjust many model parameters and left most parameters at their default values because it was reasonable that the classifier which won on average on all the experiments would also win if a better setting was performed (FernáNdez et al., 2013). Furthermore, in a framework where the classifiers were not tuned, the winning learning model tended to correspond to the most robust one, which was also a desirable characteristic (Sicilia et al., 2018). Specifically, we chose: for logistic regression, solver of liblinear; for SVM, the penalty coefficient was set to 1.0 and RBF kernel was chosen for classification; for KNN, the leaf size was set to 30; and for Random Forest and XGBoost, the number of trees was set to 10. Because our dataset was imbalanced and had less records of heath misinformation, we oversampled the dataset using the RandomOverSampler function of scikit-learn package in Python 3.7 in the training of the classification models, and used 10-fold cross validation for evaluation. The average performance of each model was reported.

4.2.3. Feature set

Because content similarity, sentiment features and topic features were all based on the textual content, we have tested the correlations between these features using the feature_selector packages in Python (Koehrsen, 2019). The results showed that the absolute value of the correlations magnitudes between these features were all less than 0.4, indicating that there was no strong or high correlation between these features (Taylor, 1990); therefore, there was no need to control the correlation between these features.

During our experiments, we added each feature subset in turn, and tested the predictive ability of the feature subsets we proposed, and finally built a model that integrated all features. In order to improve the classification accuracy of models, we used the StandardScaler function of scikit-learn package in Python 3.7 to standardize some numerical features. The data was transformed into a normal distribution with the mean value of 0 and the variance of 1.

Table 11
Health misinformation classification results.

Category	#Features	Model	Accuracy	F1-score	Precision	Recall	Time/s	Space/MiB
Topic features	79	LR	0.566	0.642	0.804	0.566	0.91	12.2266
		SVM	0.719	0.752	0.801	0.719	62.22	55.9961
		RF	0.804	0.802	0.801	0.804	2.12	12.7070
		XGBoost	0.609	0.676	0.820	0.609	11.67	20.5078
		KNN	0.691	0.734	0.800	0.691	2.91	11.1836
Topic+ Linguistic features	83	LR	0.639	0.700	0.818	0.639	1.16	10.1875
		SVM	0.834	0.817	0.823	0.834	61.33	18.7656
		RF	0.851	0.828	0.814	0.851	1.97	12.7344
		XGBoost	0.620	0.685	0.824	0.660	11.84	21.5117
		KNN	0.696	0.739	0.803	0.696	2.56	11.5117
Topic+ Linguistic +Sentiment features	85	LR	0.650	0.708	0.828	0.620	1.43	11.2812
		SVM	0.798	0.799	0.819	0.798	62.54	73.1055
		RF	0.859	0.831	0.816	0.859	1.97	13.0820
		XGBoost	0.634	0.697	0.823	0.634	12.11	22.2695
		KNN	0.695	0.739	0.804	0.695	2.13	12.0234
Topic+ Linguistic+ +Sentiment+ Behavior features	90	LR	0.685	0.734	0.83	0.685	1.7	11.9219
		SVM	0.825	0.827	0.833	0.825	61.18	79.0117
		RF	0.876	0.848	0.845	0.876	1.83	13.1406
		XGBoost	0.737	0.773	0.846	0.737	13.18	22.4023
		KNN	0.708	0.749	0.813	0.708	2.47	12.6367
Textual features	1372	LR	0.822	0.839	0.87	0.822	1.16	139.9219
		SVM	0.828	0.837	0.858	0.828	893.99	275.3984
		RF	0.830	0.830	0.835	0.830	13.05	139.2891
		XGBoost	0.833	0.844	0.860	0.833	136.92	230.5195
		KNN	0.824	0.826	0.829	0.824	67.87	84.9141

In addition, we chose the model only based on textual features as our baseline model to better evaluate the effectiveness of our proposed feature set. Textual features refer to the basic components of natural language. Bag of words, n-gram, term frequency (TF), term frequency inverted document frequency (TF-IDF) are the most commonly used textual features in natural language processing (Zhang & Ghorbani, 2019). In this paper we chose TF-IDF values to account for the importance of each word. We used Python package Jieba to achieve Chinese word segmentation, and then discarded stop words and special characters based on the common stop words list.

4.2.4. Evaluation measures

To evaluate the performance of these models, we chose accuracy, precision, recall and F1-score as our metrics. These are commonly used in classification tasks. They are defined in Eqs. (1)-(4). Accuracy measures the proportion of correctly predicted samples to the total number of samples. Precision measures the proportion of truly positive samples in identified positive samples. Recall measures the proportion of correctly identified samples in real positive samples. The F1-score combines precision and recall as an overall assessment of the performance. Additionally, we measured the time and space complexity of the models by calculating the time and memory consumption of running these models. This was calculated using memory-profiler (version 0.57.0) in Python (Pedregosa, 2020).

$$\text{Accuracy} = \frac{TP + TN}{TP+TN + FP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2^{*}\frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2^{*}\text{Precision}^{*}\text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4.2.5. Classification results

Classification results obtained with each feature subset are shown in Table 11, along with the number of features per set. As can be seen from Table 11, the models based only on topic features had acceptable levels of precision. The models based on all features had a F1-score more than 0.75 except the LR and KNN model. This indicated that they made a more accurate classification of health misinformation.

Among the five models based on different feature sets, the Random Forest model using all features achieved the best performance, with an accuracy of 0.876 and a F1-score of 0.848. Compared with the SVM model, which has a good performance, RF model has lower time and space consumption. Therefore, RF model is also suitable considering that in actual use where the amount of data will

Table 12
Summary of research on health misinformation detection using design science.

Study	Dataset	Dataset classification	Features	Classification model	Reported accuracy
(Kinsora et al., 2017, August)	2225 comments from MedHelp labeled with misinformative comments and non-misinformative comments	Binary (14.2% misinformative comments, 85.8% non-misinformative comments)	Linguistic features [Word count features; references to the self and the initiator of the thread (the question), corresponding to the "I", "Ppron", and "You" LIWC features; usage of words suggesting authenticity and authority ("Authentic" and "Clout"); health-related words ("Health" and "Bio")]	Random Forest	90.1%
Purnomo et al. (2017)	78 medical articles of snopes.com categorized into true, false, and unverified facts	Multiple (19 true claims, 42 false claims, 17 unverified texts)	Linguistic features [Headlines features, claim-headline features]	Logistic Regression,Random Forest	95.95%
(Sicilia et al., 2017, November)	800 tweets manually labeled with rumor, non-rumor and unknown	Multiple classification (rumour, non-rumour, and unknown)	Influence potential features, network characteristics features, personal interest features	Random Forest	71.4%
R. Sicilia et al. (2018)	709 tweets manually labeled with rumor, non-rumor and unknown	Multiple (54% rumor, 30% non-rumor, and 16% unknown)	Influence potential features, network characteristics features, personal interest features	SVM, MLP, Nearest Neighbor,multiclass Adaboost, Random Tree, Random Forest	86.13%
Ghenai and Mejova (2017)	26,728 tweets labeled with rumor, clarification or other	Binary (32% rumor and 68% non-rumor)	Linguistic features, Sentiment features, Twitter features, Readability features, Medical/ Domain features	Random Forest Naive Bayes Random Decision Tree	92.9%
A. Ghenai and Mejova (2018)	4000 control tweets and 4152 rumor tweets	Binary (Whether a user posts about a rumor)	Linguistic features, Sentiment features, Twitter features, Readability features, Medical features	Logistic regression with LASSO regularization	0.906(R^2)
Liu, Zhang, Susarla & Padman (2018)	600 videos from YouTube labelled with high medical knowledge or low medical knowledge	Binary (62.8% high medical knowledge and 37.2% low medical knowledge)	Text-based medical knowledge Image-based medical knowledge	Logistic Regression	85% (precision)
(Li, 2019, July)	416 claims from Snopes.com and 1692 claims from Politifact.com	Linear regression	Linguistic features, Source features	Linear Regression	0.5475(R^2)
Liu et al. (2019)	4381 health-related articles from WeChat subscription articles, unreliable news websites, Weibo and BBS.	Binary (52.4% reliable articles and 47.6% unreliable articles)	75 features based on specific word frequencies, the frequency of some punctuation; the length of title, text and paragraphs; the number of paragraphs, whether using numbers in title	FastText Decision Tree, SVM, KNN, AdaBoost, GBDT, Random Forest	84.1%
Hou et al. (2019))	250 videos from YouTube related to prostate cancer, labelled with 5-point Likert scale, 1 indicating the video contains no misinformation and 2–5 indicating an increasing level of misinformation.	Binary (47.2% misinformative and 52.8% trustworthy)	Viewer engagement features; Linguistic features (LIWC, Ngrams, Lexical richness, Syntax (CFG), Readability) Raw acoustic features	SVM (Linear kernel)	74.41%

be much larger than the data in the experiment. Additionally, when using other feature sets the RF models had better performance than other models. This might be because it is robust to oversampling biases in highly imbalanced datasets such as ours.

Furthermore, as each feature subset was added, the performance of the models was improved. This indicated that the features from different modalities contributed complementary information. This was especially true for Behavioral features which after combining the Behavioral features, the accuracy and F1-score of the optimal model improved by 2% and 2.1%, respectively.

The RF model applying all features also outperformed the baseline models relying on only textual features with the highest accuracy of 0.833 and the highest F1-score of 0.844. Compared to the model based on textual features, the feature set we constructed achieved a higher performance in the detection of health misinformation. In addition, the model based on our feature set only contains 90 features and has lower time and space consumption.

5. Discussion

5.1. Classification and annotation for the misinformation in online health communities

One of the major challenges in the area of fake review and online rumor detection research is the availability of gold standard datasets (Rout, Singh, Jena & Bakshi, 2017). Ott, Choi, Cardie and Hancock (2011) created a publicly available fake review dataset that consisted of 800 truthful reviews on hotels as well as 800 deceptive reviews obtained from Amazon Mechanical Turk (AMT). A number of studies have adopted this dataset as the experimental settings to test with supervised learning techniques for fake review detection (Stanton & Irissappane, 2019). Although this dataset was claimed as gold standard, researchers argued that the fake reviews generated by paid Turkers do not reflect actual behavioral and psychological state of mind of fake reviewers in a real scenario (Rout et al., 2017). In addition, the Turkers may not have sufficient domain knowledge or experience to write convincing fake reviews (Mukherjee et al., 2013). Mukherjee et al. (2013) testified that the fake reviews produced via crowd sourcing methods are not valid training data because the models do not generalize well on real life test data.

Yet, to the best of our knowledge, there is no existing open dataset for the health misinformation detection in online health communities. Due to the lack of a labeled dataset, in this study we employed the dataset from a real online health community as it provided us with a real time scenario for data analysis. Drawing inspiration from previous work in the area of online rumor detection, we created a coding scheme to classify the misinformation in online health communities into 3 types including advertising, propaganda, and misleading information. During the coding process, five independent coders as well as an expert coder, together annotated the dataset and achieved a satisfied level of agreement. The annotated misinformation accounted for 11.42% of all records in the sample dataset. Although the actual percentage of misinformation is unknown, the identified percentage of misinformation was consistent with previous deception prevalence studies (Ott et al., 2012, April); Mukherjee et al., 2013), which have reported 8–15% spam rate in online review sites.

5.2. Feature selection

Studies in detection of fake reviews or fake news commonly have adopted two kinds of features. One is reviews' linguistic features like words frequency and semantic features and the other is reviewers' behavioral features including network features and graph-based features. In a word, detecting fake reviews combines the features derived from reviews and reviewers.

However, as summarized in Table 12, previous studies on health misinformation detection have been focused on linguistic features, including LIWC features, N-grams features, sentiment features, features on specific word frequency, the frequency of some punctuation or special words, the length of review and so on (Ghenai & Mejova, 2017; A. 2018; Hou et al., 2019; Kinsora, Barron, Mei & Vydiswaran, 2017; Liu et al., 2019; Purnomo et al., 2017). Only a few studies have adopted behavioral features (Sicilia, Giudice, Pei, Pechenizkiy, & Soda, 2017, 2018); Hou et al., 2019). Sicilia et al. (Sicilia, Giudice, Pei, Pechenizkiy, & Soda, 2017, 2018) considered measures of centrality (including the closeness centrality and the betweenness centrality) and popularity originally conceived for the graph theory. Hou et al. (2019)) incorporated features that were found related to health misinformation such as viewer engagement features including the average number of video views per day, the number of comments posted below the video, the number of thumbs up or thumbs down, the video duration in seconds and the video category. Besides these studies in videos, researchers have extracted some features specific for videos. Liu, Zhang, Susarla & Padman (2018) combined linguistic features with image-based features extracted from video frames to detect the videos which have low medical knowledge. Hou et al. (2019)) adopted raw acoustic features of videos for detection of misinformation in videos.

Building upon previous studies and the ELM, we built a feature set integrating both the features of the message (thread/reply/reply to reply) and the behavioral features of the user who posted the message. The final feature set contained two levels of features: central-level features (topic features) and peripheral level features (including linguistic features, sentiment features, and behavioral features). In the context of OHCs, the perceived expertise of users with more engagement might be higher and thus the messages they posted could convince the recipients more easily. Therefore, health misinformation creators might intentionally establish their credibility and expertise within the communities by engaging more in user interactions. Moreover, in the process of coding analysis, it was observed that some users intentionally promoted certain products by sharing suspicious personal experiences in the community. Users who posted propaganda in the OHCs tended to make comments to a large number of posts to increase their influence in the community. Therefore, we proposed a series of behavioral features measuring the interaction involvement of the information creators in order to help detect the health misinformation.

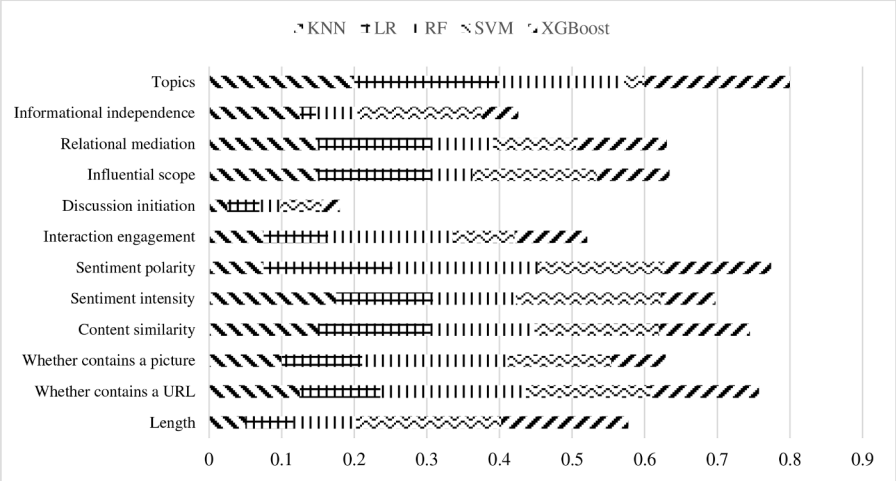

Fig. 4. Stacked histogram of the rank analysis.

As shown in the experiment results, after combining the behavioral features with the three aforementioned subsets of features, the accuracy and F1-score of the best model improved 2% and 2.1%, respectively. Therefore, it suggests that the proposed feature set, incorporating behavioral features with linguistic features, topic features, and sentiment features, was feasible and well-performed in the detection of health misinformation.

5.3. Classification methods for health misinformation detection

As we can see from Table 12, previous studies generally regarded the detection of health misinformation as a binary classification problem and used machine learning methods to detect health misinformation. Random Forest and Support Vector Machine models are most used. In this paper, we chose five classification models that were commonly used in previous studies to build classification models for health misinformation detection. The results showed that the Random Forest model applying all features had higher performance than other models, including the baseline models which only used the textual features, and has a lower time and space consumption. Considering the feasibility of the practical application, we chose the Random Forest model as the optimal model in our experiment.

5.4. Feature evaluation

The evaluation of features is an important step in analyzing which features are the most informative and whether the features are meaningful or not for the classification purpose. (R. Sicilia et al., 2018) Among the different techniques for feature evaluation, we applied the wrapper method which used the performance of a classification algorithm to compare different candidate feature sets (Huang, 2015). We analyzed each feature by evaluating the F1-score of each classifier after it was removed from the feature set. We chose the F1-score because it was more measurable for imbalanced data classification by combining precision and recall as an overall assessment of the performance. We measured the discrimination of features by the following score:

$$S(f) = F1 - score_{complete\ set} - F1 - score_{leave\ f\ out\ set}$$

Table 13
The important score of each feature in the best model.

Scope	Feature	Importance score
Linguistic	Length	0.0455
	Whether contains a picture	0.0046
	Whether contains a URL	0.0013
	Content similarity	0.0703
Topic	Topics	0.0064
Sentiment	Sentiment intensity	0.0540
	Sentiment polarity	0.0106
Behavioral	Discussion initiation	0.0517
	Interaction engagement	0.0538
	Influential scope	0.0477
	Relational mediation	0.0474
	Informational independence	0.0792

Where f was considered the descriptor in the feature set. The larger the value of S(f) was, then the more discriminant the feature f was.

Given the differences of the classification algorithms, we used multiple classifiers for the evaluation of the features in order to avoid any deviation. The relative performance of one feature relative to others was calculated by rank analysis. First, for each classifier we used, we sorted the values of S(f), and then we ranked each variable f relative to other features. The highest level was 12, which was assigned to the smallest distinguishing feature, and the lowest level was 1, which was assigned to the most informative feature. Finally, the rank of each feature f on each classifier was summarized and then normalized with respect to the highest possible value (which was calculated as highest level * number of classifiers). Fig. 4 shows the rank of each feature f among the five classifiers. Each bar in the plot is the normalized rank provided by the five classification algorithms. Each algorithm is symbolized by a different style. The bars show the contribution of each feature on each classifier; the shorter the bar the more informative the feature was.

As can be seen from Fig. 4, the behavioral features were considerably more informative compared to the linguistic features, topic features, and sentiment features. In particular, Discussion initiation, Informational independence and Interaction engagement were the three most informative features. Among them, Discussion initiation was the most informative feature in the feature set we built. It represented the number of threads the user created. Interaction engagement measured the number of replies and reply to replies the user created. As shown in Fig. 3(f, g), the misinformation's authors created more threads (9.9 vs. 4.9) and more replies (255.6 vs. 214.4) than the legitimate information's authors. Misinformation's authors often create as many threads as possible to attract attention and then they reply in the threads they have created, so that they can upload more health misinformation and spread it in a larger scope. Informational independence measured how close a node was located with respect to every other node in the network (Abraham, Hassanien, & Snášel, 2010). As Fig. 3(j) shows the misinformation's authors achieved lower closeness centralities (0.2848 vs. 0.3011) compared with the legitimate information's authors. Although some of misinformation creators achieved high closeness centrality in the network, most misinformation's authors had lower informational independence compared to legitimate users in the community. Thus, the above behavioral features represented valuable information for the health misinformation detection task.

Among all features, Topics showed the highest rank. It suggests that topic features were not as informative as other features. The topic features represented the probability distribution of records associated with generated topics. Although the experimental results showed that the models based only on topic features achieved acceptable levels of precision, the contribution of topic features in the whole feature set was relatively small.

Sentiment features have been seen as effective features in detecting online rumors and fake reviews (Rout et al., 2017). It is also widely used in health misinformation detection (Ghenai & Mejova, 2017; 2018). However, it was noticed that the sentiment features were less informative than other features in our feature set. In the dataset, on average, the emotions expressed by the misinformation (0.72) were not substantially different from the legitimate information (0.66). Therefore, the sentiment features did not discriminate well the samples in the feature space.

In addition, based on the best model we built in the Experimental Results section, we estimated the power of each feature. The importance score of each feature is listed in Table 13. It indicates the contribution of each feature in detection of health misinformation. The most influential feature was Informational independence. In addition, it was found that the most discriminative ones were related to user's behavior such as the number of threads and replies they created, and the centrality features of the user in the network. The linguistic and sentiment features (with the exception of Content similarity) had lower importance scores indicating that behavioral features considerably contribute to the detection of health misinformation.

6. Conclusion

With the development Web 2.0 technology, the credibility concerns governing online health information are renewed by the prevalence of user-generated-content. The user-generated content is not peer-reviewed, which means they usually have no source citations. It appears that unreliable health information is presented and shared by users (Ma & Atkin, 2017). User-generated content encourages participation in online health communities but the information created by lay people might contain misleading information that may lead to serious health problems.

Due to the exacerbation of the spread of health misinformation on social media, this research developed a machine learning techniques-based model incorporating linguistic features, topic features, sentiment features, and behavioral features in detecting the misinformation appearing in online health communities. The contributions of this study are three-fold. First, even though misinformation detection has been widely studied, there is a lack of detection framework guided by a comprehensive theory to identify health misinformation on social media, especially in the online health community context. By identifying different types of misinformation appearing in online health communities, this work provided the first step to attempting how to curb the spread of health misinformation through the peer-to-peer communications. Second, based on the Elaboration Likelihood Model (ELM), this study proposed that the features of online health misinformation can be classified into two levels: central-level and peripheral-level. By doing so, this study developed a feature set that integrated central-level features (including topic features) and peripheral-level features (including linguistic features, sentiment features, and user behavioral features). It verified the power of user behavioral features in distinguishing the health misinformation from the legitimate information in the online health community settings. Third, by validating the proposed model as well as the features on a real-world dataset, our findings not only extended the research on misinformation detection in the context of online health community settings, but also provided resolutions of how to detect health misinformation automatically with promising accuracy.

The theoretical contributions of this study lie in uncovering the features of the health misinformation in online health communities from a perspective of the Elaboration Likelihood Model. Lately, some efforts have explored the utilization of ELM in understanding and detecting online fake news and fake reviews. However, in theoretical terms, to the best of our knowledge, no research has yet attempted to revealing features of health misinformation in online health communities based on the ELM. By doing so, this study proposed a detection framework guided by a comprehensive theory to identify health misinformation in the online health community context. In addition, previous studies on health misinformation detection have been focused on linguistic features, including LIWC features, N-grams features, sentiment features, features on specific word frequency, the frequency of some punctuation or special words, the length of review and so on (Hou, Perez-Rosas, Loeb, & Mihalcea, 2019; Appendix 2. ODD design concepts, initialisation, input data, and submodels; Kinsora, Barron, Mei, & Vydiswaran, 2017). Only a few studies have adopted behavioral features (Sicilia, Giudice, Pei, Pechenizkiy, & Soda, 2017, 2018; Hou et al., 2019). This study proposed four types of features: linguistic features, topic features, sentiment features, and behavioral features. Although all four types of features have been applied in detecting fake news and fake reviews, to the best of our knowledge, no research has yet attempted to detect health misinformation, especially in the online health community context, combining all four types of features proposed in this study. Therefore, we believe the theoretical contributions of this study lie in the exploration of building a feature set integrating both the features of the message (including linguistic features, topic features, sentiment features) and the behavioral features of the user who posted the message to detect health misinformation in the online health community context. As shown in the experiment results in this study, after combining the behavioral features with the other three subsets of features, the accuracy and F1-score of the best model improved 2% and 2.1%, respectively. Therefore, compared to the existing literature, this study identified that behavioral features were feasible and well-performed in the detection of health misinformation in the online health community context. The proposed features and detection model can be employed to explore the misinformation appearing in online communities focusing on topics other than health domains. Specifically, our findings suggest the importance of considering the linguistic features (e.g., length, content similarity), topic feature, sentiment features, and behavioral features for the detection of misinformation, rather than relying on the high dimensional textual features (e.g., N-grams). In practice, this study verified the performance of the proposed models in distinguishing misinformation in online health communities. The proposed model could be applied by social media platforms to develop and implement screening tools for misinformation in online communities.

Despite the contributions and implications of this study, it also has several limitations. First, this study adopted the dataset from an autism-related online health community and extracted the different features of the identified misinformation. Second, the label of the misinformation relied on human annotations. In future studies, we would consider developing domain knowledge-based methods to assist the annotation of health misinformation. Third, in this study, we tried to explore the features of misinformation based on previous literature. Future works could explore other features and adopt deep learning methods to capture more features in order to distinguish misinformation from legitimate information.

Acknowledgment

This study is supported in part by the Social Science Foundation of Jiangsu Province (No. 19TQC005), the National Natural Science Foundation of China (No. 72004091, No. 71701091), the Humanities and Social Sciences Youth Foundation, Ministry of Education of the People's Republic of China (No. 20YJC870014, No. 17YJC870020), and the Key Projects of Philosophy and Social Sciences Research of Chinese Ministry of Education under Grant 19JZD021. The authors hereby declare that there are no other conflicts of interest. The authors wish to thank the anonymous reviewers for their highly constructive and helpful comments. The authors would like to thank Xu Han, Sicheng Zhu, Yuan Liu, and Yanxu Gao for their assistance.

CRediT authorship contribution statement

Yuehua Zhao: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Jingwei Da:** Data curation, Investigation, Writing - original draft. **Jiaqi Yan:** Supervision, Writing - review & editing.

References

Abraham, A., Hassanien, A.-. E., & Snášel, V. (2010). *Computational social network analysis: Trends, tools and research advances.* London: Springer.

Basit, T. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational Research, 45*(2), 143–154. https://doi.org/10.1080/0013188032000133548.

Bessi, A., Petroni, F., Del Vicario, M., Zollo, F., Anagnostopoulos, A., Scala, A., et al. (2016). Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics,, 225*(10), 2047–2059. https://doi.org/10.1140/epjst/e2015-50319-0.

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication, 65*(4), 619–638.

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication, 33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312.

Brady, J. T., Kelly, M. E., & Stein, S. L. (2017). The trump effect: With no peer review, how do we know what to really believe on social media? *Clinics in colon and rectal surgery, 30*(04), 270–276. https://doi.org/10.1055/s-0037-1604256.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32 insights.ovid.com.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684). . https://doi.org/10.1145/1963405.1963500.

Chen, L., Wang, X., & Peng, T. Q. (2018). Nature and diffusion of gynecologic cancer–related misinformation on social media: Analysis of tweets. *Journal of medical Internet research, 20*(10), https://doi.org/10.2196/11515 e11515.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: Extreme gradient boosting. *R package version 0.4-2,* 1–4.
Chou, W. Y. S., Oh, A., & Klein, W. M. (2018). Addressing health-related misinformation on social media. *JAMA, 320*(23), 2417–2418.
Chu, J. T., Wang, M. P., Shen, C., Viswanath, K., Lam, T. H., & Chan, S. S. C. (2017). How, when and why people seek health information online: qualitative study in Hong Kong. *Interactive Journal of Medical Research, 6*(2), https://doi.org/10.2196/ijmr.7000 e24.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
Ebnali, M., & Kian, C. (2020). Nudge users to healthier decisions: A design approach to encounter misinformation in health forums. In A. G. Ho (Ed.). *Advances in human factors in communication of design* (pp. 3–12). (Ed.). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-20500-3_1.
FernáNdez, A., LóPez, V., Galar, M., Del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems, 42*, 97–110.
Fichman, R. G., Kohli, R., & Krishnan, R. (2011). Editorial overview—The role of information systems in healthcare: Current research and future trends. *Information Systems Research, 22*(3), 419–428. https://doi.org/10.1287/isre.1110.0382.
Ghenai, A., & Mejova, Y. (2017). Catching Zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on Twitter. *In 2017 IEEE International Conference on Healthcare Informatics (ICHI)*IEEEhttps://doi.org/10.1109/ICHI.2017.58.
Ghenai, A., & Mejova, Y. (2018). Fake cures: User-centric modeling of health misinformation in social media. In: *Proc. ACM Hum.-Comput. Interact. 2*(CSCW) 58:1–58:20.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America. 101. Proceedings of the National Academy of Sciences of the United States of America* (pp. 5228–5235). . https://doi.org/10.1073/pnas.0307752101.
Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Amsterdam: Elsevier.
He, C., Li, S., Yao, Y., & Ding, Y. (2020). Research on the method of identifying opinion leaders based on online word-of-mouth. In H. Yang, R. Qiu, & W. Chen (Eds.). *Smart service systems, operations management, and analytics* (pp. 209–222). (Eds.). Cham, Switzerland: Springer International Publishing.
Hou, R., Perez-Rosas, V., Loeb, S., & Mihalcea, R. (2019, October). Towards automatic detection of misinformation in online medical videos. *In 2019 International Conference on Multimodal Interaction* (pp. 235–243). ACM. https://doi.org/10.1145/1122445.3353763.
Huang, S. H. (2015). Supervised feature selection: A tutorial. *Artif. Intell. Research, 4*(2), 22–37.
Isnowfy. (2015). snownlp (Version 0.12.3) [Software]. Available from https://pypi.org/project/snownlp/.
Jadad, A. R., Enkin, M. W., Glouberman, S., Groff, P., & Stern, A. (2006). Are virtual communities good for our health. *BMJ (Clinical Research Ed.), 332*(7547), 925–926. https://doi.org/10.1136/bmj.332.7547.925.
Janze, C., & Risius, M. (2017). Automatic detection of fake news on social media platforms. *PACIS 2017 Proceedings*https://aisel.aisnet.org/pacis2017/261.
Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. *Proceedings of the 2008 international conference on web search and data mining* (pp. 219–230). ACM.
Kinsora, A., Barron, K., Mei, Q., & Vydiswaran, V. V. (2017). Creating a labeled dataset for medical misinformation in health forums. *In. 2017 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 456–461). IEEE.
Koehrsen, W. (2019, October 14). Feature-selector. Retrieved April 19, 2020, from https://github.com/WillKoehrsen/feature-selector.
Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. London: SAGE.
Lai, C. L., Xu, K. Q., Lau, R. Y. K., Li, Y., & Jing, L. (2010). Toward a language modeling approach for consumer review spam detection. *2010 IEEE 7th International Conference on E-Business Engineering* (pp. 1–8). . https://doi.org/10.1109/ICEBE.2010.47.
Lau, T. P., Wang, S., Man, Y., Yuen, C. F., & King, I. (2014). Language technologies for enhancement of teaching and learning in writing. *Proceedings of the 23rd International Conference on World Wide Web* (pp. 1097–1102). New York, NY: ACM. https://doi.org/10.1145/2567948.2580058.
Lee, K., Ham, J., Yang, S.-. B., & Koo, C. (2018). Can you identify fake or authentic reviews? An fsQCA approach. In B. Stangl, & J. Pesonen (Eds.). *Information and communication technologies in tourism 2018* (pp. 214–227). (Eds.). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-72923-7_17.
Li, J. (2019). Detecting false information in medical and healthcare domains: A text mining approach. *International Conference on Smart Health* (pp. 236–246). Springer, Cham. https://doi.org/10.1007/978-3-030-34482-5_21.
Li, Y. J., Cheung, C. M. K., Shen, X. L., & Lee, M. K. O. (2019). Health misinformation on social media: A literature review. *23rd Pacific Asia Conference on Information Systems (PACIS 2019)*https://scholars.cityu.edu.hk/en/publications/health-misinformation-on-social-media(991aef31-8d00-43b9-b8fc-7321e35c2f86).html.
Li, Y., Zhang, X., & Wang, S. (2017). Fake vs. real health information in social media in China. *Proceedings of the Association for Information Science and Technology. 54. Proceedings of the Association for Information Science and Technology* (pp. 742–743). . https://doi.org/10.1002/pra2.2017.14505401139.
Liddy, E. D. (2000). Text mining. *Bulletin of the American Society for Information Science and Technology, 27*(1), 13–14. https://doi.org/10.1002/bult.184.
Stanton, G., & Irissappane, A. A. (2019). *GANs for Semi-Supervised Opinion Spam Detection.* Retrieved from https://www.ijcai.org/proceedings/2019/723.
Horne, B., & Adali, S. (2017). *This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News.* Retrieved from https://arxiv.org/abs/1703.09398.
Liu, Y., Yu, K., Wu, X., Qing, L., & Peng, Y. (2019). Analysis and detection of health-related misinformation on Chinese social media. *IEEE Access, 7*, 154480–154489.
Ma, T. J., & Atkin, D. (2017). User generated content and credibility evaluation of online health information: A meta analytic study. *Telematics and Informatics, 34*(5), 472–486. https://doi.org/10.1016/j.tele.2016.09.009.
Makagon, M. M., McCowan, B., & Mench, J. A. (2012). How can social network analysis contribute to social behavior research in applied ethology. *Applied Animal Behaviour Science, 138*(3–4), 152–161. https://doi.org/10.1016/j.applanim.2012.02.003.
Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., et al. (2013). Spotting opinion spammers using behavioral footprints. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 632–640). .
Osatuyi, B., & Hughes, J. (2018). A tale of two internet news platforms-real vs. fake: An elaboration likelihood model perspective. *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 3986–3994). . https://doi.org/10.24251/hicss.2018.500.
Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. *Proceedings of the 21st international conference on World Wide Web* (pp. 201–210). ACM. https://doi.org/10.1145/2187836.2187864.
Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume. 1. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume* (pp. 309–319). Association for Computational Linguistics.
Hagberg, A. (2019) networkx (Version 2.4) [Software]. Available from https://pypi.org/project/networkx/2.4/.
Liu, X., Zhang, B., Susarla, A., & Padman, R. (2018). *YouTube for Patient Education: A Deep Learning Approach for Understanding Medical Knowledge from User-Generated Videos.* Retrieved from https://arxiv.org/abs/1807.03179.
Pedregosa, F. (2020). memory-profiler (Version 0.57.0) [Software]. Available from https://pypi.org/project/memory-profiler/.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*(Oct), 2825–2830.
Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. Eds. In R. E. Petty, & J. T. Cacioppo (Eds.). *Communication and persuasion: Central and peripheral routes to attitude change* (pp. 1–24). New York: Springer. https://doi.org/10.1007/978-1-4612-4964-1_1.
Purnomo, M. H., Sumpeno, S., Setiawan, E. I., & Purwitasari, D. (2017). Keynote speaker II: Biomedical engineering research in the social network analysis era: Stance classification for analysis of hoax medical news in social media. *Procedia Comput. Sci. 116. Procedia Comput. Sci* (pp. 3–9).
Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1589–1599). .
Richards, K. A. R., & Hemphill, M. A. (2017). A practical guide to collaborative qualitative data analysis. *Journal of Teaching in Physical Education, 37*(2), 225–231. https://doi.org/10.1123/jtpe.2017-0084.
Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014).*Evaluating topic coherence measures.* Retrieved from https://arxiv.org/abs/1403.6397.
Rout, J. K., Singh, S., Jena, S. K., & Bakshi, S. (2017). Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications, 76*(3),

3187–3211.
Sarwar, B. M. (2001, February 19). Adjusted Cosine Similarity. Retrieved April 22, 2020, from http://www.www10.org/cdrom/papers/519/node14.html.
Savov, P., Jatowt, A., & Nielek, R. (2020). Identifying breakthrough scientific papers. *Information Processing & Management, 57*(2), Article 102168. https://doi.org/10.1016/j.ipm.2019.102168.
Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2017). Health-related rumour detection on Twitter. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1599–1606). IEEE. https://doi.org/10.1109/BIBM.2017.8217899.
Sicilia, R., Giudice, S. L., Pei, Y., Pechenizkiy, M., & Soda, P. (2018). Twitter rumour detection in the health domain. *Expert Systems with Applications, 110,* 33–40.
Silver, L., & Huang, C. (2020, September 11). Smartphone, Social Media Users Have Broader Social Networks In Emerging Economies. Retrieved September 28, 2020, from https://www.pewresearch.org/internet/2019/08/22/in-emerging-economies-smartphone-and-social-media-users-have-broader-social-networks/.
Singh, V. K., Ghosh, I., & Sonagara, D. (2020). Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology,* 1–15. https://doi.org/10.1002/asi.24359.
Sommariva, S., Vamos, C., Mantzarlis, A., Đào, L. U. L., & Martinez Tyson, D. (2018). Spreading the (fake) news: Exploring health messages on social media and the implications for health professionals using a case study. *American Journal of Health Education, 49*(4), 246–255. https://doi.org/10.1080/19325037.2018.1473178.
Song, H., Omori, K., Kim, J., Tenzek, K. E., Hawkins, J. M., & Lin, W. Y. (2016). Trusting social media as a source of health information: Online surveys comparing the United States, Korea, and Hong Kong. *Journal of medical Internet research, 18*(3), 1–12 e25.
Sun, J. (2019). jieba (Version 0.40) [Software]. Available from https://pypi.org/project/jieba/.
Syed-Abdul, S., Fernandez-Luque, L., Jian, W. S., Li, Y. C., Crain, S., Hsu, M. H., et al. (2013). Misleading health-related information promoted through video-based social media: Anorexia on YouTube. *Journal of medical Internet research, 15*(2) e30.
Tandoc, E. C., Jr, Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism, 6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143.
Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54.
Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of diagnostic medical sonography, 6*(1), 35–39.
Tseng, C.-. W., Chou, J.-. J., & Tsai, Y.-. C. (2018). Text mining analysis of teaching evaluation questionnaires for the selection of outstanding teaching faculty members. *IEEE Access, 6*, 72870–72879. https://doi.org/10.1109/ACCESS.2018.2878478.
Vogel, L. (2017). Viral misinformation threatens public health. *CMAJ: Canadian Medical Association Journal, 189*(50), https://doi.org/10.1503/cmaj.109-5536 E1567–E1567.
Wardle, C. (2017, May 15). Fake news. It's complicated. Retrieved September 29, 2020, from https://firstdraftnews.org/latest/fake-news-complicated/.
Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media–the pilot quantitative study. *Health Policy and Technology, 7*(2), 115–118. https://doi.org/10.1016/j.hlpt.2018.03.002.
Wu, J., & Liu, Y. (2019). Deception detection methods incorporating discourse network metrics in synchronous computer-mediated communication. *Journal of Information Science*Article 0165551518823176. https://doi.org/10.1177/0165551518823176.
Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems, 33*(2), 456–481.
Zhang, X., & Ghorbani, A. A. (2019). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*Article 102025.
Zhao, Y., & Zhang, J. (2017). Consumer health information seeking in social media: A literature review. *Health Information & Libraries Journal, 34*(4), 268–283. https://doi.org/10.1111/hir.12192.
Zhao, Y., Zhang, J., & Wu, M. (2019). Finding users' voice on social media: An investigation of online support groups for Autism-affected users on Facebook. *International Journal of Environmental Research and Public Health, 16*(23), https://doi.org/10.3390/ijerph16234804.