Social Network Analysis and Mining (2022) 12:94
https://doi.org/10.1007/s13278-022-00921-9

REVIEW PAPER

Applications of machine learning for COVID-19 misinformation: a systematic review

A. R. Sanaullah[1] · Anupam Das[1] · Anik Das[2] · Muhammad Ashad Kabir[3] · Kai Shu[4]

Received: 28 January 2022 / Revised: 2 July 2022 / Accepted: 4 July 2022 / Published online: 29 July 2022

Abstract

The inflammable growth of misinformation on social media and other platforms during pandemic situations like COVID-19 can cause significant damage to the physical and mental stability of the people. To detect such misinformation, researchers have been applying various machine learning (ML) and deep learning (DL) techniques. The objective of this study is to systematically review, assess, and synthesize state-of-the-art research articles that have used different ML and DL techniques to detect COVID-19 misinformation. A structured literature search was conducted in the relevant bibliographic databases to ensure that the survey was solely centered on reproducible and high-quality research. We reviewed 43 papers that fulfilled our inclusion criteria out of 260 articles found from our keyword search. We have surveyed a complete pipeline of COVID-19 misinformation detection. In particular, we have identified various COVID-19 misinformation datasets and reviewed different data processing, feature extraction, and classification techniques to detect COVID-19 misinformation. In the end, the challenges and limitations in detecting COVID-19 misinformation using ML techniques and the future research directions are discussed.

Keywords COVID-19 · Misinformation · Classification · Machine learning · Deep learning

1 Introduction

Misinformation is a piece of false information or inaccurate information that is intentionally created to get more attention from people (Fernandez and Alani 2018). There are many terms related to misinformation such as fake news, rumor, false information, misleading information, and disinformation (Wu et al. 2019). Despite their similarities, they differ slightly in terms of usage contexts, degrees of incorrectness as well as the functions of serving in various propagation scenarios (Su et al. 2020; Wang et al. 2019).

During this COVID-19 pandemic situation, there has been an expeditious growth in the usage of social media platforms and blogging websites which has passed 3.8 billion marks of active users (Huang and Carley 2020). People are now getting more involved in these platforms, especially on Facebook, Twitter, Instagram, etc., and expressing their thoughts, and opinions as well as sharing the news and information related to COVID-19. Every now and then, they seek information about COVID-19, e.g., symptoms, medicines, vaccines, mask usage, post complications, and dangers (UNICEF 2021). They gather information about COVID-19 from any news media or social media platforms

✉ Anik Das
x2021gmg@stfx.ca

A. R. Sanaullah
u1604050@student.cuet.ac.bd

Anupam Das
u1604054@student.cuet.ac.bd

Muhammad Ashad Kabir
akabir@csu.edu.au

Kai Shu
kshu@iit.edu

[1] Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh

[2] Department of Computer Science, St. Francis Xavier University, Antigonish, NS B2G 2W5, Canada

[3] Data Science Research Unit, School of Computing, Mathematics and Engineering, Charles Sturt University, Bathurst, NSW 2795, Australia

[4] Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA

and share it with others without fact-checking the information. Along with factual information, it is observed that a large amount of misinformation related to COVID-19 is circulating through these platforms, which is causing panic, and affecting people's mental health, daily lives, and behaviors (Su et al. 2021). For instance, the health officials in Nigeria found a number of cases overdosed on Chloroquine (a drug formerly used for the treatment of Malaria) after the news to treat coronavirus with the drug through the news media (Busari and Adebayo 2020). World Health Organization (WHO) called this situation an 'infodemic'—an overabundance of both inaccurate and accurate information to explain the misinformation about the virus and makes it harder for people to find trustworthy and reliable sources for any claim made on any online platforms during the pandemic (WHO 2020; Zarocostas 2020).

It is now a global concern to combat the spread of COVID-19 misinformation on online platforms. It has already gained a great deal of attention from researchers all around the world. A significant number of research works (Elhadad et al. 2021; Chen 2020; Kar et al. 2020) have applied various ML techniques for detecting COVID-19 misinformation in online platforms. As there are still many challenging issues in the existing studies that need further investigations, it is important to explore potential research directions that can improve the efficiency of the systems to combat the spread of misinformation in this pandemic. Hence, it is necessary to review the existing research on COVID-19 misinformation detection to understand the state-of-the-art research, their limitations and explore potential future research directions that can improve the effectiveness and efficacy of the approaches to combat the spread of misinformation in this pandemic.

In this study, we have conducted a survey of state-of-the-art research on COVID-19 misinformation detection. We systematically search and select 43 research articles based on our inclusion criteria. We include papers that aim to detect COVID-19 misinformation using either traditional ML or DL techniques. We have outlined and grouped various COVID-19 misinformation datasets including their sources, number of instances, classes, and links to download. We have analyzed the pre-processing and feature extraction methods and the performance of various classification techniques used in COVID-19 misinformation detection. Finally, we have discussed the research gaps and future research directions on COVID-19 misinformation detection.

The rest of the paper is organized as follows. Section 2 provides an overview of COVID-19 misinformation and its impact. Section 3 presents our methodology to search databases along with the selection criteria of the articles. Section 4 outlines different datasets for COVID-19 misinformation and presents an analysis of various pre-processing, feature extraction, and classification methods used in the state-of-the-art research. Section 5 discusses open issues and future research directions. Finally, Sect. 6 concludes the paper.

2 COVID-19 misinformation

2.1 Misinformation types

According to Fetzer (2004), misinformation is 'false, mistaken, or misleading information.' Others define misinformation as inaccurate information, which is created to misguide the readers (Fernandez and Alani 2018; Zhang et al. 2018) or 'any claim of fact that is currently false due to lack of scientific evidence' (Chou et al. 2018). Many terms are related to misinformation such as fake news, rumor, false information, misleading information, and disinformation. Despite the similarities, there exist some differences between them which are easily distinguishable. Figure 1 depicts the categorization of COVID-19 misinformation within the scope of this survey.

Fake news is a modified version of original news which is used to misguide the people or manipulate public opinion using traditional mass media and online social media (Cui and Lee 2020). It is also known as fabricated information which differs in organizational procedure or purpose but looks similar to news media content (Lazer et al. 2018). It can be misleading or dangerous when it is out of context and original sources. It is used to describe phony press releases, hoaxes, and spam since there is no official definition (Su et al. 2020). These kinds of news are unreliable and create misconceptions among the people.

Conspiracy theory is created by the secret or powerful groups rather than as natural disasters or caused by clear action to identify the reason behind varied events as plots (Bale 2007; Swami et al. 2011; Douglas et al. 2016). These are created for doing harm to the people with the help of internet access (van Prooijen and Douglas 2018; Douglas et al. 2016). People believe in conspiracy theories during societal crises, such as natural disasters, financial crises or diseases, wars, and terrorist attacks (Fritsche et al. 2017; Van Prooijen and Douglas 2017). For example, many conspiracy theories are created during the COVID-19 crisis, such as '5G cellular network is the root cause of the virus,' and 'Bill Gates is using the virus as a cover for his desire to create a worldwide surveillance state through the enforcement of a global vaccination program' (Shahsavari et al. 2020).

Rumor is basically a story of uncertain or doubtful truth. It spreads online very quickly (Lin et al. 2019). Sometimes, it is called 'false rumor' or 'fake news' when a rumor's veracity value is false (Li et al. 2019). Many kinds of rumors are circulating during this COVID-19 pandemic. For example,

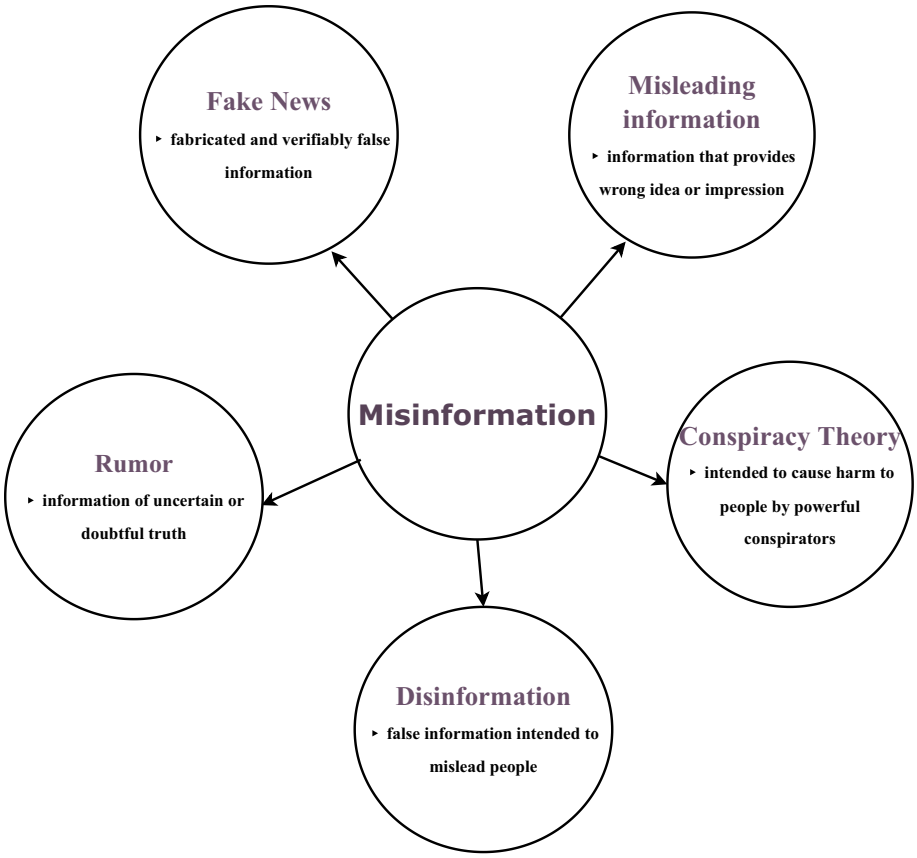

Fig. 1 Types of COVID-19 misinformation

among the rumors spread at the beginning of coronavirus infection in Bangladesh are: 'Coronavirus would not come in Bangladesh as its temperature is more than 30 degrees,' and 'Drinking 3 cups of tea daily can get rid of coronavirus' (Akon and Bhuiyan 2020).

Misleading information is defined as incorrect information which is given to an eyewitness following an event (tutor2u 2020). It may be planned to upset the economy of nations, diminish individuals' trust in their governments or elevate a particular item to accomplish huge benefits, which have already happened with COVID-19 (Elhadad et al. 2020).

Disinformation is treated as a part of misinformation (Losee 1997; Zhou et al. 2004). Inaccurate information is referred to as 'Misinformation,' whereas deceptive information is referred to as 'Disinformation' (Karlova and Fisher 2013). It creates misconceptions among the people. One recent disinformation related to COVID-19 is that drinking pure alcohol can kill the coronavirus (Bernard 2020), which is truly misguiding and injurious to health.

2.2 Impact of COVID-19 misinformation

Since the beginning of the COVID-19 pandemic, misinformation has become a major issue worldwide. The main reason behind this is the substantial increase in internet use during this pandemic for different purposes, e.g., communication (Nguyen et al. 2020), business (Papadopoulos et al. 2020; Petratos 2021), health-related information (Li et al. 2020a), etc. Due to the anxiety, worry, and panic over local transmission and multiple infections among the population, which can trigger xenophobia on the continent, a group of people is currently circulating various types of misinformation on social media platforms (Ahinkorah et al. 2020; Mejova and Kalimeri 2020; Shimizu 2020). Facebook, a popular social networking site, has reported that approximately 90 million pieces of content during the March and April of 2020 are related to COVID-19 misinformation (Spring 2021). A study Li et al. (2020a) also reported that approximately 23% to 26% of YouTube videos related to COVID-19 were misleading information. It hampers the practice of healthy behaviors and promotes unsound practices, which negatively affect both the physical and mental health (Tasnim et al. 2020). Furthermore, some misinformation might create a serious threat by misleading the general population (Ahinkorah et al. 2020). The unwillingness of taking the COVID-19 vaccine among people is an example in this regard (Loomba et al. 2021).

Table 1 Database search string

Database name	Query string/Keywords
Scopus	TITLE-ABS-KEY ((COVID-19 OR coronavirus) AND ("fake news" OR misinformation OR rumors OR misleading) AND (detection OR classification OR clustering))
Web of Science	TS=((COVID-19 OR coronavirus) AND (fake news OR misinformation OR misleading OR rumors) AND (detection OR classification OR clustering))
Google Scholar	COVID-19 fake news detection, COVID-19 fake news classification, COVID-19 misinformation detection, COVID-19 misleading news detection, COVID-19 rumor detection

3 Methodology

In this section, we present our search scope and database search methods for collecting articles related to our study. We outline three prominent databases and the queries used for searching relevant articles and present the selection criteria process based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al. 2009) method where we illustrate step by step systematic approach for selecting the articles.

3.1 Search scope

In this survey, we have searched three prominent databases such as Scopus, Web of Science, and Google Scholar. Scopus and Web of Science are the popular authentic databases that maintain the published paper from IEEE, ACM, Elsevier, Springer, etc. Google Scholar also provides a simple way to broadly search for scholarly literature.

3.2 Database search method

We have used the query string/keyword-based searching method in our study. Our query string/keyword includes COVID-19-related misinformation, fake news, rumors, and misleading information-related studies that have used detection, classification, and clustering techniques using ML algorithms. The search keywords and query strings are listed in Table 1. We have searched the different formatted query strings on these databases between July 18, 2021 and July 24, 2021.

3.3 Selection criteria

For the selection of the papers for our systematic review, we have defined five inclusion criteria: (i) the article must be focused on the detection of COVID-19 misinformation, (ii) The subject matter of this study exists anywhere in the title, abstract, or keywords of the article, (iii) the article should either employ any traditional ML and/or DL model(s) to classify misinformation or present a dataset related to COVID-19 misinformation, (iv) article employing classification model(s) must have presented performance evaluation of the adopted model(s), and (v) article must be written in English.

Figure 2 shows the systematic selection process of the articles using PRISMA (Moher et al. 2009). A total of 260 papers were found in the 'identification' phase of our study by searching the databases. After removing 38 duplicate articles, the remaining 222 articles were screened by their titles and abstracts in the 'screening' phase. In this phase, the articles are further filtered out with the inclusion criteria and 134 articles were excluded accordingly. In the 'eligibility' phase, full texts of the remaining 88 articles were studied for final selection. A total of 45 articles were eliminated during this phase for not relating to COVID-19 misinformation classification or not employing any traditional ML or DL techniques. Finally, in the 'included' phase, we have found 43 papers that were included and analyzed in this survey.

4 Analysis

In this section, we reviewed the datasets, different preprocessing and feature extraction techniques, and the classification methods used for COVID-19 misinformation detection along with their evaluation results.

4.1 Dataset description

Relevant and sufficient training data are considered the basis to achieve precise results from any ML-based misinformation detection system. To perform the misinformation classification task, data from various platforms such as social media, news websites, fact-checking sites, and government or well-recognized authentic websites are being used frequently. But manually determining the authenticity of news is a very challenging task because it usually requires annotators with domain expertise. Therefore, to facilitate future research work related to the COVID-19 misinformation task, some recent and existing datasets are presented in Table 2 which are described in the next subsections.

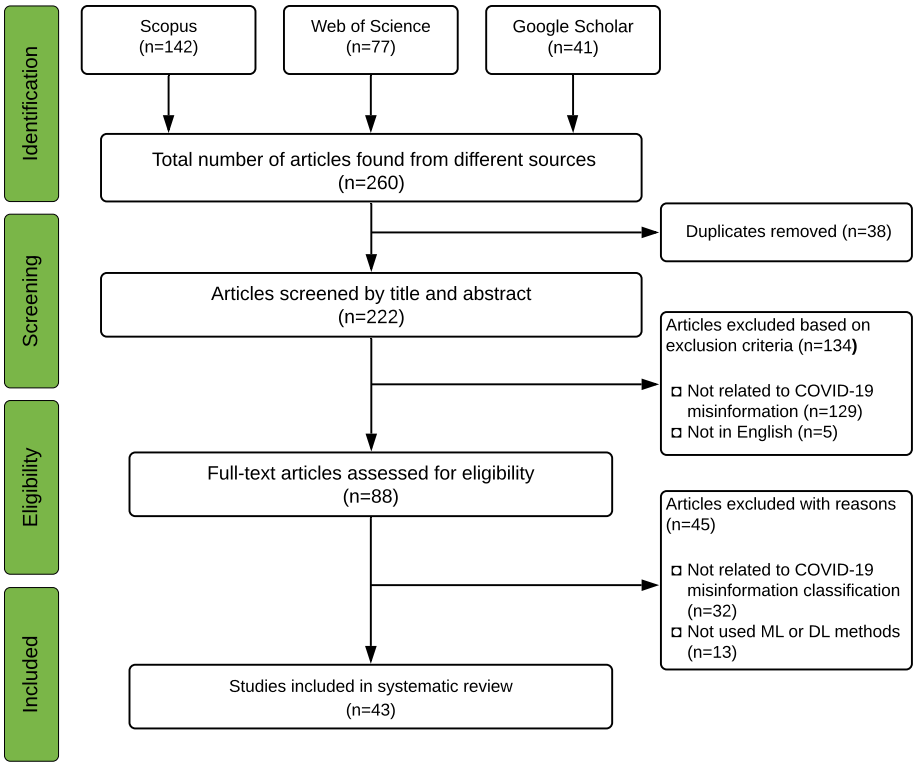

Fig. 2 Prisma flow diagram for the systematic selection and evaluation of the articles

4.1.1 Data sources

Studies included in this review paper cover data from multiple sources. The articles Elhadad et al. (2021, 2020) utilized the data which is collected from official websites and official Twitter accounts of the UNICEF, WHO, and UN as well as from different fact-checking websites (i.e., Snopes, PolitiFact). A large number of studies (Kar et al. 2020; Madani et al. 2021; Alkhalifa et al. 2020; Hossain et al. 2020; Al-Rakhami and Al-Amri 2020; Bandyopadhyay and Dutta 2020; Kumar et al. 2021; Alsudias and Rayson 2020; Alam et al. 2021; Dimitrov et al. 2020; Lamsal 2020; Qazi et al. 2020; Banda et al. 2021; Chen et al. 2020; Lopez and Gallemore 2020; Shahi et al. 2021; Preda 2020; Mahlous and Al-Laith 2021 Boukouvalas et al. 2020; Dharawat et al. 2020; Alqurashi et al. 2020; Micallef et al. 2020) used the Twitter platform as a data source. Several Twitter APIs such as streaming API and Tweepy API are generally used to collect the tweets from this platform. Medina Serrano et al. (2020) used video data that were collected from YouTube using YouTube's Data API. In Zhou et al. (2020a), the dataset includes news articles as well as tweets related to the news articles. These articles are crawled from a set of reliable news sites referenced by news fact-checking websites: NewsGuard, MBFC, and the tweets are collected by using Twitter Premium Search API. Haouari et al. (2020) created a dataset containing COVID-19-related claims and their relevant tweets. They were collected from Arabic fact-checking platforms (Fatabyyano and Misbar), English fact-checking websites (e.g., PolitiFact, Snopes), and the Twitter accounts of WHO, UNICEF, etc. Another study Cui and Lee (2020) released a dataset containing news articles, claims, and social media posts. News articles were collected from various reliable news outlets, e.g., Healthline, Medical News Today, etc., claims were collected referring to the WHO official website, WHO official Twitter account, etc., and finally, the social media posts were collected from Facebook, Twitter, Instagram, YouTube, and TikTok. Gao et al. (2020) introduced a multilingual dataset containing microblogs related to COVID-19 from Twitter and Chinese social media platform Weibo.

On the other hand, Chen (2020) utilized the data fetched from various Chinese rumor-refuting platforms such as Sina News, Baidu, and 360 rumor-refuting platforms. The study Shahi and Nandini (2020) used data that were collected from different fact-checking websites by getting references from Poynter and Snopes. Ng and Carley (2021) collected fact-checked stories regarding coronavirus to make a dataset. The stories were curated from popular fact-checking websites such as Poynter, Snopes, and PolitiFact. Song et al. (2020) gathered their data from the IFCN Poynter website. WANG et al. (2021) collected their rumor data from Snopes. The study Koirala (2020) released a dataset by scraping the data from various news and blog sites using Webhose.io API. In the studies Yang et al. (2021); Shi et al. (2020), the authors used a dataset containing microblogs related to COVID-19

Table 2 Summary of the datasets

Type	Paper	Dataset name	Data source				Dataset link	Language	Instances	Labels
			SP	FCW	NW	O				
Misleading	Elhadad et al. (2020)	COVID-19-FAKES	✓	✓	–	✓	i	En,Ar	7486	2
Fake News	Kar et al. (2020)	Indic-covidemic tweet dataset	✓	–	–	–	ii	En,B,H	1438	2
	Shahi and Nandini (2020)	FakeCovid	–	✓	–	–	iii	En, H, G and Other 37	5182	2
	Koirala (2020)	Abhishek Koirala	–	–	✓	–	–	En	3119	2
	Paka et al. (2021)	COVID-19 Twitter Fake News (CTF)	✓	✓	–	✓	xxxi	En	45,261	2
	Madani et al. (2021)	Madani et al.	✓	–	–	–	–	En	2000	2
	Hossain et al. (2020)	COVID19-Lies	✓	–	–	–	vi	En	6761	3
	Boukouvalas et al. (2020)	COVID-19 Twitter Data	✓	–	–	–	vii	En	560	2
	Li et al. (2020b)	MM-COVID	✓	✓	✓	✓	xxvi	En, S, P, H, It, F	11,173	2
	Al-Rakhami and Al-Amri (2020)	Al-Rakhami et al.	✓	–	–	–	–	En	409,484	2
	Bandyopadhyay and Dutta (2020)	Fake news dataset	✓	–	–	✓	–	En	19,873	2
	Kumar et al. (2021)	Kumar et al.	✓	–	–	–	–	En	1970	4
	Patwa et al. (2021)	COVID-19 Fake News Dataset	✓	✓	–	–	viii	En	10,700	2
	Micallef et al. (2020)	Counter-covid19-misinformation	✓	–	–	–	ix	En	155,468	3
	Zhou et al. (2020a)	ReCOVery	✓	–	✓	–	x	En	News 2029 tweets 140,820	2
	Dharawat et al. (2020)	Covid-HeRA	✓	–	–	–	xi	En	61,286	5
	Shahi et al. (2021)	Misinformation COVID-19	✓	✓	–	–	xii	En	1500	2
	Cui and Lee (2020)	CoAID	✓	✓	✓	–	xiv	En	298,778	2
	Haouari et al. (2020)	ArCOV-19	✓	✓	–	–	xvi	Ar	9414	3
	Kaliyar et al. (2021)	FN-COV	–	–	✓	–	–	En	69,976	2
	Ayoub et al. (2021)	Ayoub et al.	–	✓	✓	✓	–	En	984	2
	Ng and Carley (2021)	Ng et al.	–	✓	–	–	–	En	6731	5
	Mahlous and Al-Laith (2021)	Arabic Fake News corpora	✓	–	–	–	xxv	Ar	36,066	2
	Yang et al. (2021)	CHECKED	✓	–	–	–	xix	C	2104	2
Rumor	Chen (2020)	Shuaipu Chen	–	✓	–	–	–	En	3737	3
	WANG et al. (2021)	Wang et al.	–	✓	–	–	xxvii	En	7179	3
	Shi et al. (2020)	Shi et al.	✓	–	–	–	–	En	1537	2
	Alkhalifa et al. (2020)	CLEF dataset	✓	–	–	–	iv	En	962	2
	Alsudias and Rayson (2020)	COVID-19 Arabic tweets	✓	–	–	–	xvii	Ar	2000	3
	Cheng et al. (2021)	COVID-19-rumor-dataset	✓	✓	✓	–	xxviii	En	6834	3
Conspiracy	Medina Serrano et al. (2020)	*YouTube_misinfo*	✓	–	–	–	v	En	Videos 180 comments 151,567	2
Disinformation	Alam et al. (2021)	COVID-19 Infodemic Twitter Dataset	✓	–	–	–	xviii	En, Ar	722	2
	Song et al. (2020)	COVID-19 Disinformation corpus	–	✓	–	–	–	En	1293	10

Table 2 (continued)

Type	Paper	Dataset name	Data source				Dataset link	Language	Instances	Labels
			SP	FCW	NW	O				
Unlabeled	Dimitrov et al. (2020)	TweetsCOV19	✓	–	–	–	xiii	En	8,151,524	–
	Lamsal (2020)	COV19Tweets Dataset	✓	–	–	–	xx	En	Over 310 million	–
	Paka et al. (2021)	CTF	✓	✓	–	✓	xxix	En	21.85 million	–
	Qazi et al. (2020)	GeoCoV19	✓	–	–	–	xxi	En, S and other 60	524,353,432	–
	Banda et al. (2021)	COVID-19 Twitter Chatter Dataset	✓	–	–	–	xxii	En, F, S, G and others	Over 1.12 billion	–
	Alqurashi et al. (2020)	COVID-19-Arabic-Tweets-Dataset	✓	–	–	–	xxiii	Ar	3,934,610	–
	Chen et al. (2020)	COVID-19 Twitter dataset	✓	–	–	–	xxiv	En, S, I, F, P and other 62	123,113,914	–
	Lopez and Gallemore (2020)	*COVID19_Tweets_Dataset*	✓	–	–	–	xv	En, S, P and other 63	785,118,723	–
	Preda (2020)	COVID19 Tweets	✓	–	–	–	xxx	En	179,108	–
	Gao et al. (2020)	NAIST COVID	✓	–	–	–	xxxi	En, Ja, C	25,925,773	–

SP= social platform; *FCW* = fact checking website; *NW*= news website; *O*= others; *En*= English; *B* = Bengali; *C* = Chinese; *Ja*= Japanese; *H* = Hindi; *Ar* = Arabic; *G* = German; *S* = Spanish; *P* = Portuguese; *I* = Indonesian; *F* = French; *It* = Italian

⋆ Dataset links are provided in the appendix section

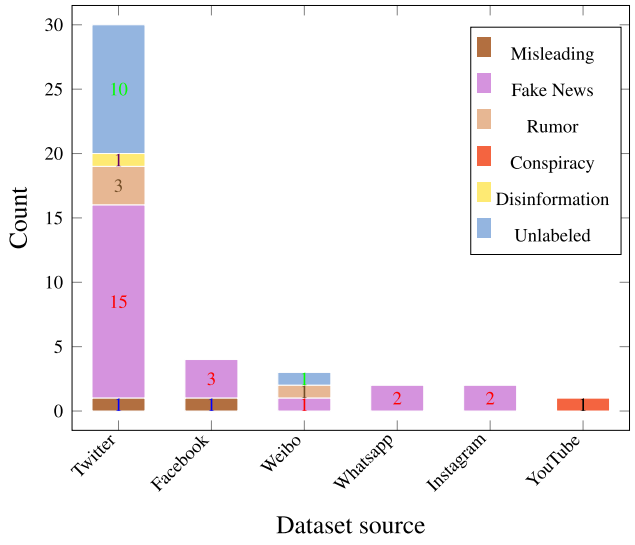

Fig. 3 Datasets from various social media platforms

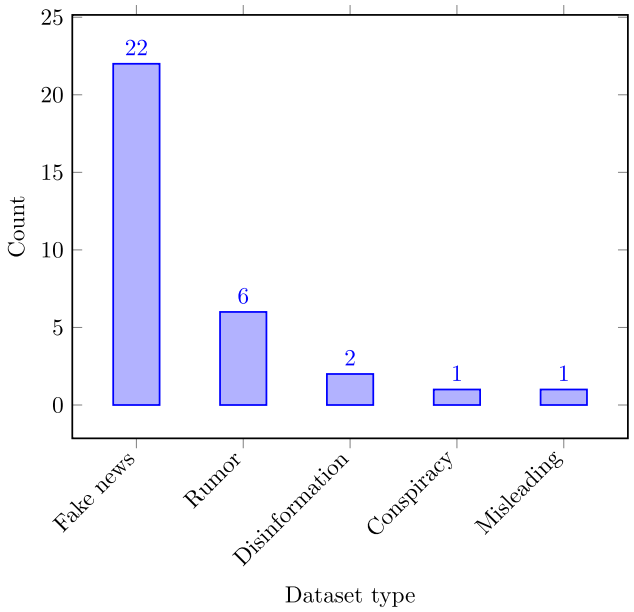

Fig. 4 Count of different dataset types

which were crawled from the popular Chinese social media platform Weibo. Patwa et al. (2021) used the data collected from public fact-verification websites and other sources, e.g., World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), etc. Cheng et al. (2021) introduced a COVID-19 rumor dataset that contains rumors regarding COVID-19 from a wide range of sources. These rumors were collected from various news sites (e.g., CNN, BBC News), fact-checking websites (e.g., Poynter, FactCheck), and Twitter platforms. This dataset also includes some metadata of the rumors which are source website, date of publication, reposts or retweets, etc. In Kaliyar et al. (2021), a dataset has been used that contains news articles regarding COVID-19 published worldwide. Ayoub et al. (2021) introduced a dataset that contains data collected from new sites (e.g., Aljazeera, CNN), fact-checking sites (e.g., Snopes, Poynter), and other reliable sources like WHO, CDC, etc. Li et al. (2020b) proposed a news dataset named MM-COVID which contains multilingual and multidimensional COVID-19 fake news data. They used fact-checking websites like Snopes and Poynter to collect fake content and several health-related websites to collect COVID-19-related real information. Social media (Facebook, Twitter, Instagram, etc.) posts and news articles posted on blog sites and traditional news agencies were considered to collect both fake and real news.

The variation of data collection from various social platforms is shown in Fig. 3. In this figure, the number of the datasets that cover data from various social platforms (e.g., Facebook, Twitter, YouTube, Weibo, WhatsApp, Instagram) are shown using different colors. On the other hand, Fig. 4 represents the number of datasets against their application purposes, such as fake news, rumor, disinformation, conspiracy theory, and misleading information.

4.1.2 Dataset class labels

In this survey paper, we have classified the existing studies into five major misinformation categories which include misleading information, fake news, rumor, conspiracy theory, and disinformation. Table 3 represents the datasets along with their corresponding class labels as well as the studies that introduced or used them.

In the misleading category, the studies Elhadad et al. (2020, 2021) used a dataset containing two class labels—Real and Misleading, where 'Real' indicates accurate information relating to COVID-19 and 'Misleading' indicates inaccurate information.

Several studies fall under the fake news category. The datasets used in these studies contain class labels varying from 2 to 5. The studies (Kar et al. (2020); Cui and Lee (2020); Shahi and Nandini 2020; Koirala 2020; Bandyopadhyay and Dutta 2020; Patwa et al. 2021; Yang et al. 2021; Shahi et al. 2021; Kaliyar et al. 2021; Ayoub et al. 2021; Mahlous and Al-Laith 2021; Paka et al. 2021; Li et al. 2020b; Madani et al. 2021) used datasets having two class labels: Fake and Real, where 'Fake' represents false news regarding COVID-19 and 'Real' represents true news pieces related to COVID-19, but in some datasets, the class 'Real', 'genuine' are represented as 'True' or 'Not Fake'. Some studies (Boukouvalas et al. 2020; Al-Rakhami and Al-Amri 2020; Zhou et al. 2020a) used different names to represent class labels. 'Unreliable' or 'Non-credible' is

Table 3 Datasets and their class labels (continued)

Dataset name	Class labels	Used in
COVID-19-FAKES	Real, Misleading	Elhadad et al. (2020, 2021)
Indic-covidemic tweet dataset	Fake, Non-Fake (Real)	Kar et al. (2020)
FakeCovid	False (Fake), Others	Shahi and Nandini (2020)
Abhishek Koirala	Fake, True (Real)	Koirala (2020)
Madani et al.	Fake, Real	Madani et al. (2021)
CTF	Fake, Genuine (Real)	Paka et al. (2021)
COVID-19 Twitter Data	Reliable, Unreliable	Boukouvalas et al. (2020)
MM-COVID	Fake, Real	Li et al. (2020b)
Al-Rakhami et al.	Credible, Non-credible	Al-Rakhami and Al-Amri (2020)
Fake news dataset	Fake, Real	Bandyopadhyay and Dutta (2020)
Kumar et al.	Irrelevant, Conspiracy, True (Real), False (Fake)	Kumar et al. (2021)
COVID-19 Fake News Dataset	Fake, Real	Patwa et al. (2021); Wani et al. (2021); Bang et al. (2021); Ayoub et al. (2021); Chen et al. (2021)
Counter-covid19-misinformation	Misinformation, Counter-misinformation, Irrelevant	Micallef et al. (2020)
ReCOVery	Reliable, Unreliable	Zhou et al. (2020a)
Covid-HeRA	Real, Refutes/Rebuts, Highly severe, Possibly severe, Not severe	Dharawat et al. (2020)
COVID19-Lies	Agree, Disagree, No Stance	Hossain et al. (2020)
Misinformation COVID-19	False (Fake) and Partially False (Partially fake)	Shahi et al. (2021)
CoAID	Fake, True (Real)	Cui and Lee (2020); Kaliyar et al. (2020)
ArCOV-19	False (Fake), True (Real), Other	Haouari et al. (2020)
FN-COV	Fake, Real	Kaliyar et al. (2020, 2021)
Ayoub et al.	Fake, True (Real)	Ayoub et al. (2021)
Ng et al.	True (Real), Partially true (Partially real), Partially false (Partially fake), False (Fake), Unknown	Ng and Carley (2021)
Arabic Fake News corpora	Fake, Not Fake (Real)	Mahlous and Al-Laith (2021)
CHECKED	Fake, Real	Yang et al. (2021)
Shuaipu Chen	Health rumor, Science rumor, Society rumor	Chen (2020)
Wang et al.	Fake (Rumor), Real, Unverified	WANG et al. (2021)
Shi et al.	Rumor, Real	Shi et al. (2020)
CLEF dataset	Rumor, Non-rumor	Alkhalifa et al. (2020)
COVID-19 Arabic tweets	True (Real), False (Rumor), Unrelated	Alsudias and Rayson (2020)
COVID-19-rumor-dataset	True (Real), False (Rumor), Unverified	Cheng et al. (2021)
YouTube_misinfo	Conspiracy, Agreement	Medina Serrano et al. (2020)
COVID-19 Infodemic Twitter Dataset	Yes (Not trustworthy), No (Trustworthy)	Alam et al. (2021); Bang et al. (2021)
COVID-19 Disinformation corpus	PubAuthAction, CommSpread, GenMedAdv, PromActs, Consp, VirTrans, VirOrgn, PubRec, Vacc, None	Song et al. (2020)

used to represent fake news, and 'Reliable' or 'Credible' is used to represent true news pieces about COVID-19. In Hossain et al. (2020), the dataset includes three misconception classes—Agree, Disagree, and No Stance. These labels are defined by determining the expression of a tweet to the misconception. If the tweet is a positive expression of the misconception then it is labeled as Agree, if the tweet disagrees with the misconception then it is labeled as Disagree and finally, if the tweet is neutral or not relevant to the misconception then it is labeled as No Stance. Another study Haouari et al. (2020) also used a dataset of three classes labeled False, True, and Other. If a tweet expresses a veracious claim then it is labeled as True, if not then the tweet is labeled as False, if the tweet cannot be labeled as one of the two earlier cases then it is labeled as Other. In Micallef et al. (2020), the authors labeled their data with three class labels named Misinformation, Counter-misinformation, and Irrelevant. 'Misinformation' is used to label a tweet

if the tweet includes decontextualized truths, falsehoods, inaccuracies, etc., if the tweet refutes false claims then it is labeled as Counter-misinformation and the tweet is labeled as Irrelevant if a tweet cannot be categorized in the prior two classes. Ng and Carley (2021) used five classes to label their collected stories regarding COVID-19. These classes include true, partially true, partially false, false, and Unknown. A story is labeled as 'True' if it is verifiable by trusted sources (e.g., CDC), 'Partially True' if it contains verifiable true facts, and facts that cannot be verified, 'Partially False' if it has verifiable false facts, and the facts that cannot be verified, 'False' if it is proved false by trusted sources, and finally 'Unknown' if it cannot be verified at all. The other two studies (Kumar et al. 2021; Dharawat et al. 2020) under this category used four and five class labels, respectively, to organize their data.

In the rumor category, the studies Shi et al. (2020); Alkhalifa et al. (2020) used data that are labeled as Rumor and Non-rumor or Real. If a tweet needs check-worthiness for the topic, it is labeled as a Rumor, otherwise, it is labeled as Non-rumor or real. In Chen (2020), the author labeled the COVID-19 rumor data into three categories where health-related rumors are labeled as health rumors, scientific rumors are labeled as science rumors, and the rumors about the society are labeled as society rumors. Alsudias and Rayson (2020) organized their data using three-class labels which are True, False, and Unrelated. The tweets which represent correct information are labeled as 'True', whereas the tweets containing rumors or false information are labeled as 'False', and irrelevant tweets are labeled as 'Unrelated'. In Cheng et al. (2021), the authors also used three classes to label their collected data. An instance is labeled as 'True' if it contains logical and authentic facts related to COVID-19, 'False' if it contains any false information or rumor, and 'Unverified' if the authenticity cannot be verified.

In the category named conspiracy theory, Medina Serrano et al. (2020) used YouTube videos along with their corresponding comments and labeled them as Conspiracy and Agreement. If the comments express any agreement then they are labeled as Agreement, oppositely comments amplifying misinformation with a conspiracy theory are labeled as Conspiracy. In the disinformation category, Song et al. (2020) developed a dataset containing 10 class labels. These are used to label the debunks of COVID-19 disinformation. Labels include PubAuthAction (Public authority), CommSpread (Community spread and impact), GenMedAdv (Medical advice, self-treatments, and virus effects), PromActs (Prominent actors), Consp (Conspiracies), VirTrans (Virus transmission), VirOrgn (Virus origins and properties), PubRec (Public Reaction), Vacc (Vaccines, medical treatments, and tests), and None (Cannot determine). Another study Alam et al. (2021) labeled their collected tweet data into two major classes named Yes and No for their binary classification task. Tweets are labeled based on the answers to some questions, e.g., 'Is the tweet contain any factual claim?', 'To what extent does the tweet contain false information?', etc.

4.1.3 Dataset language and availability

Among the labeled datasets, most of them contain data only in the English language. The datasets (Haouari et al. 2020; Mahlous and Al-Laith 2021; Alsudias and Rayson 2020) contain the tweets only in the Arabic language while the dataset used in this studies Elhadad et al. (2021, 2020) contains data in two different languages—English and Arabic. The dataset Yang et al. (2021) contains data in the Chinese language. Some studies (Kar et al. 2020; Shahi and Nandini 2020; Li et al. 2020b; Alam et al. 2021) also introduced multilingual datasets containing data in multiple languages. Datasets used in the studies (Chen 2020; Koirala 2020; Madani et al. 2021; Song et al. 2020; Al-Rakhami and Al-Amri 2020; Bandyopadhyay and Dutta 2020; Kumar et al. 2021; Kaliyar et al. 2021; Ayoub et al. 2021; Ng and Carley 2021; WANG et al. 2021) have not been made publicly available. Additionally, we have collected some unlabeled datasets that are vast in size. All of these datasets (Dimitrov et al. 2020; Lamsal 2020; Qazi et al. 2020; Banda et al. 2021; Alqurashi et al. 2020; Chen et al. 2020; Lopez and Gallemore 2020; Preda 2020; Li et al. 2020b; Gao et al. 2020) are publicly available to use. The datasets (Qazi et al. 2020; Banda et al. 2021; Chen et al. 2020; Lopez and Gallemore 2020; Li et al. 2020b; Gao et al. 2020) are multilingual while others (Lamsal 2020; Preda 2020; Madani et al. 2021; Alqurashi et al.2020) are monolingual containing data in English (first three) and Arabic, respectively. The dataset proposed by Paka et al. (2021) contains both labeled and unlabeled data in the English language and it is publicly available. After making some modifications and proper annotations, future research works may be conducted in this domain by utilizing these datasets.

4.2 Data pre-processing

Data pre-processing is one of the significant parts before feeding the data into any ML algorithm. It includes data cleaning, transformation, normalization, feature extraction, and selection. This step aims to facilitate data manipulation, reduce the required memory, and speed up the processing of large quantities of data. The pre-processing techniques used in COVID-19 misinformation studies are reported in Table 4 and discussed below.

Tokenization and stop-word removal both are the most common methods performed during the data

Table 4 Data pre-processing techniques used in existing research

Techniques	Explanation	Papers
Tokenization	Splitting the text into smaller units, known as 'Token'	Boukouvalas et al. (2020); Song et al. (2020); Kumar et al. (2021); Elhadad et al. (2021); Patwa et al. (2021); Alkhalifa et al. (2020); Shahi and Nandini (2020); Koirala (2020); Medina Serrano et al. (2020); Hossain et al. (2020); Dharawat et al. (2020); Alam et al. (2021); Alsudias and Rayson (2020); Chen et al. (2021); Ayoub et al. (2021)
Stop-words Removal	Removing the words which do not provide much context and hold less useful information	Boukouvalas et al. (2020); Song et al. (2020); Kumar et al. (2021); Elhadad et al. (2021); Patwa et al. (2021); Alkhalifa et al. (2020); Shahi and Nandini (2020); Koirala (2020); Medina Serrano et al. (2020); Elhadad et al. (2020); Bandyopadhyay and Dutta (2020); Alam et al. (2021); Alsudias and Rayson (2020); Chen et al. (2021); Wani et al. (2021); Ayoub et al. (2021)
Case-folding	Converting the characters of a sentence into lower case	Kaliyar et al. (2021); Wani et al. (2021)
Stemming	Converting a word to its grammatical roots so that they can be presented in one term only	Elhadad et al. (2021, 2020); Wani et al. (2021); Ng and Carley (2021); Mahlous and Al-Laith (2021)
Lemmatization	Transforming a word to its root which is also known as 'lemma' depending on the context	Kumar et al. (2021); Ayoub et al. (2021); Ng and Carley (2021); Mahlous and Al-Laith (2021)
POS tagging	Assigning one of the part-of-speech to a given word	Elhadad et al. (2020, 2021)
Data Augmentation	Increasing the data by modifying existing data	Kar et al. (2020); Ayoub et al. (2021)
Others	Removing HTML tags, URLs and other special characters from texts	Kaliyar et al. (2021); Wani et al. (2021); Ng and Carley (2021); Mahlous and Al-Laith (2021); Chen et al. (2021)

pre-processing step. In the tokenization process, the entire text or paragraphs are split into small units, called tokens, whereas the removal of the stop-word is the process of eliminating the words which do not provide much context. These steps are performed effectively in a number of studies (Song et al.2020; Medina Serrano et al. 2020; Boukouvalas et al. 2020). Patwa et al. (2021) represented their data by performing tokenization and removing stop-words with non-alphanumeric characters and unnecessary links. Hossain et al. (2020) conducted this tokenization process using NLTK Library. Bandyopadhyay and Dutta (2020) deleted incomplete news and communal news in their pre-processing steps as it had no need for misinformation detection. They also removed data that had no relation to COVID-19 because of their specific research on COVID-19 misinformation analysis.

Kumar et al. (2021) used the NLTK Library for text processing and removed stop-words. They also removed unnecessary tweets, usernames, etc., from their data and performed a lemmatization technique for converting a word to its roots which helps to extract features in the next step. In Alsudias and Rayson (2020), the authors removed hashtags, URLs, emojis, numbers, stop-words, repetitive tweets, and characters as they had no significance in their study. They also performed normalization and tokenization techniques in the tweets data for better representation of the data.

Elhadad et al. (2020) used some steps such as text parsing, data cleaning, and stop-word removal, POS tagging, stemming for data pre-processing. In the data cleaning process, they applied the regular expression to get the combination of English alphabets and numbers and eliminate others. They also transformed the digit into the text. In Elhadad et al. (2021), the authors removed links, symbols, stop-words, HTML encoding, and repeated words and performed POS tagging and stemming.

Alkhalifa et al. (2020) presented different kinds of pre-processing techniques such as Segment2Token, Segment-2Root, Word2id, and padding. Shahi and Nandini (2020) used a python-based library named 'langdetect' to identify different kinds of languages to assign respective languages to the articles. They also used NLTK, TEXTblob, and regular expression for data cleaning and the pre-processing steps like tokenization and spell correction. In another study Dharawat et al. (2020), reserved tokens such as URLs, mentions, and retweets are filtered out from the tweet data. Alam et al. (2021) performed case folding and removed non-ASCII characters and hash symbols and replaced the URLs and usernames by using URL tag and user tag. Data augmentation, a popular technique for increasing the data volume, has been used in the studies (Kar et al. 2020; Ayoub et al. 2021). Ng and Carley (2021) conducted stemming and lemmatization on the words to find their grammatical roots, as

Table 5 Feature extraction methods used in the literature

Methods	Papers
Pre-trained BERT	Kar et al. (2020); Alkhalifa et al. (2020); Hossain et al. (2020); Song et al. (2020); Dharawat et al. (2020); Ng and Carley (2021); WANG et al. (2021); Cheng et al. (2021)
mBERT	Kar et al. (2020)
COVID-Twitter-BERT	Alkhalifa et al. (2020); Hossain et al. (2020); Dharawat et al. (2020)
RoBERTa	Chen et al. (2021)
GloVe	Cui and Lee (2020); Elhadad et al. (2021); Koirala (2020); Hossain et al. (2020); Kumar et al. (2021); Dharawat et al. (2020); Wani et al. (2021); WANG et al. (2021)
ELMo	Alkhalifa et al. (2020)
Word2Vec	Alsudias and Rayson (2020); WANG et al. (2021)
FastText	Alsudias and Rayson (2020); Wani et al. (2021); WANG et al. (2021)
BoW	Cui and Lee (2020); Medina Serrano et al. (2020); Ng and Carley (2021)
Count Vector	Alsudias and Rayson (2020); Koirala (2020); Mahlous and Al-Laith (2021)
TF	Elhadad et al. (2020)
TF-IDF	Elhadad et al. (2020); Alkhalifa et al. (2020); Koirala (2020); Medina Serrano et al. (2020); Hossain et al. (2020); Patwa et al. (2021); Alsudias and Rayson (2020); Dharawat et al. (2020); Ayoub et al. (2021); Mahlous and Al-Laith (2021); Bang et al. (2021)
PCA	Boukouvalas et al. (2020)
ICA	Boukouvalas et al.(2020)
LIWC	Medina Serrano et al. (2020); Zhou et al. (2020a)
RST	Zhou et al. (2020a)
VAE	Cheng et al. (2021)

well as removed special characters from the textual contents of the stories.

Chen et al. (2021) performed tokenization to split the texts into a set of tokens and removed all the URLs, stop-words, and nor-alphanumeric characters from the texts in the pre-processing step. In Kaliyar et al. (2021), the authors conducted several pre-processing tasks which include the removal of HTML tags, special characters, and numbers, and the conversion of text characters into lowercase and number words into numeric forms. Wani et al. (2021) performed stemming to convert the words into their respective roots and normalization to transform the characters into lowercase. They also removed HTML tags, stop-words, special characters, white spaces, etc., to pre-process the data. In Ayoub et al. (2021), the authors developed a data augmentation technique called back-translation to increase the size of the dataset. In the back-translation technique, a text is translated back to its original language after translating it into an intermediate one. The authors also performed other commonly used pre-processing tasks such as tokenization, lemmatization, and stop-words removal. Mahlous and Al-Laith (2021) carried out some pre-processing tasks which include removal of mentions, hashtags, hyperlinks, punctuations, repeated characters, non-Arabic words, etc. They also used ISRIStemmer[1] for stemming and Tashaphyne[2] to generate the roots of the words.

[1] https://www.kite.com/python/docs/nltk.ISRIStemmer.

[2] https://pypi.org/project/Tashaphyne/.

4.3 Feature extraction

Feature extraction is a process of dimensionality reduction without losing important information. In the text categorization, a document generally consists of a large number of words, and phrases which creates a high computational burden in the learning process. Also, it is difficult to learn from high-dimensional data. Besides, the classifier's accuracy can decrease by taking irrelevant features. Taking relevant and important features can help to speed up the learning process. We have found different feature extraction methods in our study. The methods used in the papers (see Table 5) are summarized below.

PCA (F.R.S. 1901) is a method that is used for dimensionality reduction. By using this process, it produces lower-dimensional feature sets. It is very important to determine the number of principal components in PCA. If p is the number of principal components to be chosen among all of the components, the values of p should represent the data at their very best. In Boukouvalas et al. (2020), the authors applied PCA in their training dataset after removing the mean value from the initial vectors for centering all the features. This operation projected the training dataset onto the N-dimensional sub-space and reduced the dimension.

ICA (Hérault and Jutten 1987) is a linear transformation method in which the desired representation is the one that minimizes the statistical dependence of the components of the representation. It does not focus on the mutual orthogonality of the components and the issue of the variance among the data points. In Boukouvalas et al. (2020), the authors performed the ICA after performing the PCA technique in their dataset to reduce the statistical dependence.

Bag-of-Words (BoW) model is a text representation used in NLP. In this method, a text is represented as a bag (multiset) of its words and does not regard any grammar or word order but it maintains multiplicity. Each word's occurrence is considered a feature in this representation. This method has been adopted for vector representation of texts in several studies (Cui and Lee 2020; Medina Serrano et al. 2020). Ng and Carley (2021) used BoW to generate vector representation of word occurrences in each sentence of the stories regarding COVID-19. They also used the Term Frequency-Inverse Document Frequency (TF-IDF) as a weighting scheme with the BoW model to represent the relative importance of a word in the sentences.

TF-IDF (Robertson 2004) is a feature extraction method that comes from language modeling theory. It states that words in a text can be divided into two groups: words with eliteness and words without eliteness. It is calculated by combining two metrics, one of which represents the number of times a word occurs in a document and the other representing the inverse document frequency of a word over a set of documents. In Elhadad et al. (2020), the authors extracted different kinds of TF and TF-IDF features (unigram, bigram, trigram, character level, and N-gram word size) on their collected ground-truth data and those features showed different outstanding results for different models. Another study Medina Serrano et al. (2020) used these standard TF-IDF features to pre-process the comments for getting better performance from their classification model. Ayoub et al. (2021) applied the TF-IDF method to represent the texts into a vector space and extract relevant features. The authors used these features as input to the ML classification models. In Mahlous and Al-Laith (2021), the authors used three different TF-IDF representations to convert texts of the Arabic tweets into a vectorized form. In word-level TF-IDF, they represented each word in the TF-IDF matrix; in n-gram-level TF-IDF, they used unigram, bigram, and trigram sequence in the TF-IDF matrix, and in character-level TF-DF, the matrix was formed representing the TF-IDF character scores. Hossain et al. (2020) used this method for the extraction of both unigram and bigram TF-IDF vectors and utilized the extracted features to perform the classification task. In Alkhalifa et al. (2020), the authors trained word embeddings on the training data with the classification model and merged them with the TF-IDF representation of the tweets. The combined features led to improved performance over the n-gram baseline. Other studies (Koirala 2020; Alsudias and Rayson 2020; Patwa et al. 2021; Dharawat et al. 2020) also applied this feature extraction method to convert the data into a matrix of TF-IDF features and extract valuable features for the classification purposes.

Count Vector converts the text into a vector-based on the frequency (count) of each word found in the text. By using CountVectorizer, a matrix is created in which each specific word is represented by a column and each text sample from the document is represented by a row. The count of the word in that specific text sample is the value of each cell. In the studies Koirala (2020); Mahlous and Al-Laith (2021), the authors used this technique to represent the textual contents into a vector space containing the counts of the terms present in the texts. Alsudias and Rayson (2020) followed this approach for the extraction of linguistic features from COVID-19 tweets. They utilized the extracted features in the classification phase and obtained a good performance.

LIWC (Pennebaker et al. 2001) stands for Linguistic Inquiry and Word Count. It is a psycholinguistic lexicon and can count the words in the article. It counts the words based on one or more of 93 linguistic, psychological, and topical categories. Zhou et al. (2020a) used this lexicon for their study to extract valuable features from the news articles. Medina Serrano et al. (2020) used the Logistic Regression model using LIWC's lexicon-derived frequencies as features during the training step.

RST (MANN and Thompson 1988) stands for Rhetorical Structure Theory which points out the relationship between the parts of the text and represents them as a content of a tree. In the study Zhou et al. (2020a), the authors used a pre-trained RST parser (Ji and Eisenstein 2014) and got a representational view of the tree for each news article which enumerated the rhetorical relation within a tree. By performing this action, 45 features are extracted and classified in a traditional statistical learning framework.

Word2vec (Mikolov et al. 2013) is a word embedding technique developed by Google based on a shallow neural network. There are two types of Word2vec. One is Skip-gram and another is Continuous Bag of Words (CBOW). The CBOW method takes the context of each word as the input and tries to predict the word related to the context. It has better representations for more frequent words. On the other hand, the distributed representation of the input word is used to predict the context in the Skip-gram model which works well with a small amount of data and is found to represent rare words well. WANG et al. (2021) used GoogleNews-vectors-negative300[3] which constitutes 100 billion words trained on the Google News dataset. In Alsudias and

[3] https://code.google.com/archive/p/word2vec/.

Rayson (2020), the authors used Word2vec to create word embeddings from the tweets. They utilized the word embeddings produced by Word2vec as input features to the classification models. This word embedding method can capture the importance of the relevant information from the texts, hence capable of showing good performance.

FastText (Joulin et al. 2017) is an extension of Word2vec model developed by Facebook AI research lab. As a technique of extracting the n-gram feature, the generated vector for a word includes the sum of this character's n-grams. It can derive word vectors for unknown words by taking morphological characteristics of words even if a word wasn't seen during training. So it works well with rare words and can provide any vector representation. In Alsudias and Rayson (2020), the authors applied FastText to generate word embedding-based features from the tweet texts. Later, they used these features for the classification task. WANG et al. (2021) had chosen crawl-300d-2M4 embedding[4], a 2 million word vectors which are trained with subword information on Common Crawl (600B tokens). In Wani et al. (2021), the authors used 300-dimensional pre-trained Fast-text embeddings to convert the input texts into a sequence of word vectors. These word vectors preserved important features of the texts and fed them to the classification models as input.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019), a technique for NLP pre-training is developed by Google. It is deeply bidirectional because it learns text representation for both directions for a better understanding of the context and relationship. There are mainly two kinds of models. One is BERT Base and another is BERT Large, as well as there are some models based on languages such as English, Chinese, and a multi-lingual model (mBERT) which covers 102 languages and is trained on Wikipedia. In Kar et al. (2020), the authors performed feature extraction using pre-trained BERT embeddings with (or without) different extracted features (e.g., text features, Twitter user features, fact verification score, source tweet embedding, etc.). This study also used the multilingual BERT (mBERT) model and fine-tuned it to learn the textual features from tweets. Dharawat et al. (2020) used pre-trained BERT embeddings (Devlin et al. 2019) to determine how close each tweet was by using the centroid of its respective category based on their document vectors. The studies (Alkhalifa et al. 2020; Song et al. 2020; WANG et al. 2021; Bang et al. 2021) also used pre-trained BERT embedding and different pre-trained BERT-based models to transform their features into a word-level vector representation. Hossain et al. (2020) executed a contextualized word embedding using the pre-trained BERT model to find the semantic similarity between the tweets and misconceptions. In Ng and Carley (2021), the authors used the pre-trained BERT to generate contextualized word embeddings of the stories related to COVID-19. Cheng et al. (2021) used BERT to convert the rumor texts into a contextualized vector form. They also used an LSTM-based variational autoencoder (VAE) (Cheng et al. 2020) after the BERT to extract the important features from the vectors generated by BERT. The generative nature of the VAE model makes it more robust in the extraction of relevant features. The authors used these features extracted by VAE with a classification model and achieved a good performance score. In the studies (Alkhalifa et al. 2020; Hossain et al. 2020; Dharawat et al. 2020), the authors used COVID-Twitter-BERT (Müller et al. 2020) model to learn the features more effectively because the COVID-Twitter-BERT (CT-BERT) model has domain adaptive pre-training on COVID-19 Twitter data.

GloVe (Pennington et al. 2014) stands for Global Vectors for Word Representation which is developed at Stanford University as an open-source project. It is an unsupervised learning algorithm that is used for generating word embeddings. Here, all the words are mapped into a meaningful space where the distance between words is related to semantic similarity. An aggregated global word co-occurrence matrix from a corpus is used for training. Therefore, the resulting representations indicate interesting linear substructures of the word in vector space. In the studies Elhadad et al. (2021); Kumar et al. (2021), the authors applied an embedding layer of dimension 300 using the GloVe pre-trained word embedding model. This embedding layer can transform the tweet texts into a vector representation to capture the relevant features. Dharawat et al. (2020) used 100-dimensional pre-trained GloVe embeddings with different classifiers as text representation. Other studies Cui and Lee (2020); Wani et al. (2021) also employed the same dimensional pre-trained GloVe embeddings for their feature extraction process. WANG et al. (2021) had chosen the GloVe.840b.300d.3[5] which is trained on Common Crawl consisting of 2 million words. Hossain et al. (2020) used GloVe to generate non-contextualized word embedding. The authors utilized GloVe vectors of 300 dimensions to extract non-contextualized word embeddings from the texts and later used them as the features. The study Koirala (2020) used 300-dimensional GloVe vectors for word embedding purposes. The author applied GloVe to create an embedding matrix of words with the indices of tokenized words.

ELMo (Peters et al. 2018) stands for Embeddings from Language Models developed in 2018 by AllenNLP. It is a deep contextualized word representation that does not

[4] https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M.vec.zip.

[5] http://nlp.stanford.edu/data/glove.840B.300d.zip.

Table 6 Classification strategies used in the literature

Strategy	Papers
Binary class	Cui and Lee (2020); Elhadad et al. (2021); Kar et al. (2020); Shahi and Nandini (2020); Alkhalifa et al. (2020); Koirala (2020); Medina Serrano et al. (2020); Boukouvalas et al. (2020); Elhadad et al. (2020); Al-Rakhami and Al-Amri (2020); Patwa et al. (2021); Bandyopadhyay and Dutta (2020); Alam et al. (2021); Zhou et al. (2020a); Shi et al. (2020); Paka et al. (2021); Bang et al. (2021); Kaliyar et al. (2020); Yang et al. (2021); Mahlous and Al-Laith (2021); Kaliyar et al. (2021); Ayoub et al. (2021); Chen et al. (2021); Wani et al. (2021)
Multi-class	Chen (2020); Hossain et al. (2020); Alsudias and Rayson (2020); Song et al. (2020); Kumar et al. (2021); Alam et al. (2021); Dharawat et al. (2020); WANG et al. (2021); Cheng et al. (2021); Ng and Carley (2021)

use fixed embedding for each word but for creating word representations, it employs a deep, bidirectional LSTM model. Unlike other traditional word embeddings such as Word2vec and GloVe, it analyzes words within the context that they are used rather than a dictionary of words or their corresponding vectors. As a result, the same word can generate different word vectors under different contexts. In Alkhalifa et al. (2020), the authors used ELMo to create a word-level representation of the tweets. The word embeddings produced by the pre-trained ELMo model were fed as input features into a classification model. But, ELMo embedding did not show good performance in the classification task.

4.4 Classification methods

To perform a classification task, two types of classification strategies are commonly used—binary classification and multi-class classification. As can be seen in Table 6, binary classification is the most used classification strategy for classifying COVID-19 misinformation compared to multi-class classification.

4.4.1 Traditional ML methods

Traditional ML methods perform very well in the detection of misinformation on COVID-19. Several traditional ML algorithms have been used to perform the classification

Table 7 Traditional ML methods used in the literature

Methods	Papers
SVM	Cui and Lee (2020); Kar et al. (2020); Medina Serrano et al. (2020); Boukouvalas et al. (2020); Elhadad et al. (2020); Al-Rakhami and Al-Amri (2020); Patwa et al. (2021); Alsudias and Rayson (2020); Dharawat et al. (2020); Zhou et al. (2020a); Ng and Carley (2021); Mahlous and Al-Laith (2021); Bang et al. (2021)
LR	Cui and Lee (2020); Medina Serrano et al. (2020); Elhadad et al. (2020); Patwa et al. (2021); Alsudias and Rayson (2020); Dharawat et al. (2020); Zhou et al. (2020a); Ayoub et al. (2021); Ng and Carley (2021); Mahlous and Al-Laith (2021)
RF	Cui and Lee (2020); Kar et al. (2020); Medina Serrano et al. (2020); Al-Rakhami and Al-Amri (2020); Dharawat et al. (2020); Zhou et al. (2020a); Ayoub et al. (2021); Mahlous and Al-Laith (2021); Elhadad et al. (2020)
DT	Elhadad et al. (2020); Patwa et al. (2021); Zhou et al. (2020a); Ayoub et al. (2021)
NB	Al-Rakhami and Al-Amri (2020); Alsudias and Rayson (2020); Zhou et al. (2020a); Ng and Carley (2021); Mahlous and Al-Laith (2021)
MNB	Medina Serrano et al. (2020); Elhadad et al. (2020)
BNB	Elhadad et al. (2020)
kNN	Elhadad et al. (2020); Al-Rakhami and Al-Amri (2020); Bandyopadhyay and Dutta (2020); Zhou et al. (2020a)
XGBoost	Elhadad et al. (2020); Mahlous and Al-Laith (2021); Shi et al. (2020)
GDBT	Patwa et al. (2021)
C4.5	Al-Rakhami and Al-Amri (2020)
Perceptron	Elhadad et al. (2020)
BN	Al-Rakhami and Al-Amri (2020)
Linear Classifier	Hossain et al. (2020)

of COVID-19 misinformation. The studies that have used traditional ML methods are shown in Table 7.

Based on the ICA model (Hérault and Jutten 1987), Boukouvalas et al. (2020) proposed a data-driven solution where knowledge discovery and detection of misinformation are achieved jointly. Their proposed method helps to generate low-dimensional representations of tweets with respect to their spatial context and deployed SVM (Cortes and Vapnik 1995) by using different kinds of popular kernel methods, e.g., Gaussian, RBF, Polynomial. Using the SVM model with the Gaussian kernel method, accuracy of 81.2% was reported in their study. Bang et al. (2021) used the SVM model for setting up the baseline of their experiment which is trained on the TF-IDF feature and cross-entropy (CE) as a loss function. The authors were able to acquire a 93.32% accuracy and F1-score for both.

Elhadad et al. (2020) proposed a voting ensemble ML classifier based on ten classification algorithms (DT, MNB, BNB, LR, kNN, Perceptron, SVM, RF, and XGBoost). They used TF and TF-IDF with character level, unigram, bigram, trigram, and N-gram word size and word embeddings as feature extraction techniques to extract effective features. The study reports the performance of the models using different evaluation metrics such as 99.63% accuracy using the SVM model with character level features, 99.36% accuracy using the LR model with TF features, 99.20% accuracy using the DT model with character level features, etc. Overall, LR and DT classifiers showed the best outcomes. The authors used reliable ground-truth data and appropriate feature extraction methods which helped them to obtain good results for the models. Kar et al. (2020) used SVM and RF classifiers with pre-trained BERT embeddings for the classification of COVID-19 fake tweets in English. The authors used 80% of the dataset in the training phase, whereas the remaining 20% was used in the testing phase. In the classification task, SVM showed an F1-score of 75%, which is slightly higher (by 1%) than the RF model.

Medina Serrano et al. (2020) presented a model that uses user comments to detect COVID-19 misinformation videos on YouTube. For classifying user comments, they trained some models where two models are trained as baselines. One is the LR model based on LIWC's lexicon-derived frequencies (Tausczik and Pennebaker 2010) as features and another is the MNB model using BOW as features. They used the percentage of conspiracy comments on each video as a feature to classify the videos and extracted content features from the video's titles and the first hundred comments per video. They set six features such as title, conspiracy, comments, and their combination, and used LR, SVM, and RF where the SVM model is trained using different kernel methods such as linear, sigmoid, and RBF kernel. Among all the six features, comments with conspiracy features got slightly better accuracy. Alsudias and Rayson (2020) employed three ML algorithms named SVM, LR, and NB for the classification of COVID-19-related rumors in the Arabic language. The authors used different types of features such as Count Vector, TF-IDF, Word2vec, and FastText with the classification models. In the classification of rumors, the highest accuracy of 84.03% was obtained from both the SVM classifier (with TF-IDF features) and the LR classifier (with Count Vector features). On the other hand, the NB classifier showed slightly lower performance by achieving an accuracy of 81.04% using Count Vector features.

Bandyopadhyay and Dutta (2020) used the kNN classifier to find the truthfulness of the news shared on social media using their own collected dataset during four months of lockdown. Before fitting into the classifier, they pre-processed the dataset based on the similarity news on social media. They got a decent accuracy using this classifier. In Al-Rakhami and Al-Amri (2020), kNN classifier was used as candidate weak-learners during the experimental phase of ensemble learning where this algorithm obtained an accuracy of 94.39% for tenfold cross-validation.

Cui and Lee (2020) used different classification methods on their own created dataset as baselines for the comparative analysis of the misinformation detection task. They used BOW features and fed the representations to a linear kernel SVM and RF classifier. For feeding into the LR model, they concatenated all the word embeddings together. Although these models did not achieve a good score in this dataset, the comparative analysis helped to find the overall model performance. In Zhou et al. (2020a), extensive experiments are conducted using the ReCOVery dataset which included the baseline performances using either single-modal or multi-modal information of news articles for predicting news credibility and allowing future methods to be compared. Different kinds of methods such as LR, NB, kNN, RF, DT, and SVM are adopted in their experiment using LIWC and RST features.

Dharawat et al. (2020) performed experiments with several multiclass classification models on their own created new benchmark dataset- 'Covid-HeRA.' They used RF, SVM, and LR models with BOW and 100-dimensional pre-trained GloVe embeddings and achieved a very good accuracy above 95%. In Patwa et al. (2021), the authors performed an experiment with their annotated benchmark dataset using four ML baselines (DT, LR, GB, and SVM) and obtained the best performance of 93.46% F1-score with SVM using the TF-IDF feature. Hossain et al. (2020) trained linear classifiers on three datasets, i.e., SNLI, MultiNLI, and MedNLI using TF-IDF vectors and average GloVe embeddings as features separately. These classifiers did not show good performance in terms of macro F1-scores. The authors stated that the bad performance is due to the NLI datasets which lack the texts related to COVID-19 and also these texts are linguistically different from the tweets.

Ayoub et al. (2021) experimented with three ML algorithms (e.g., LR, RF, and DT) using TD-IDF features. The

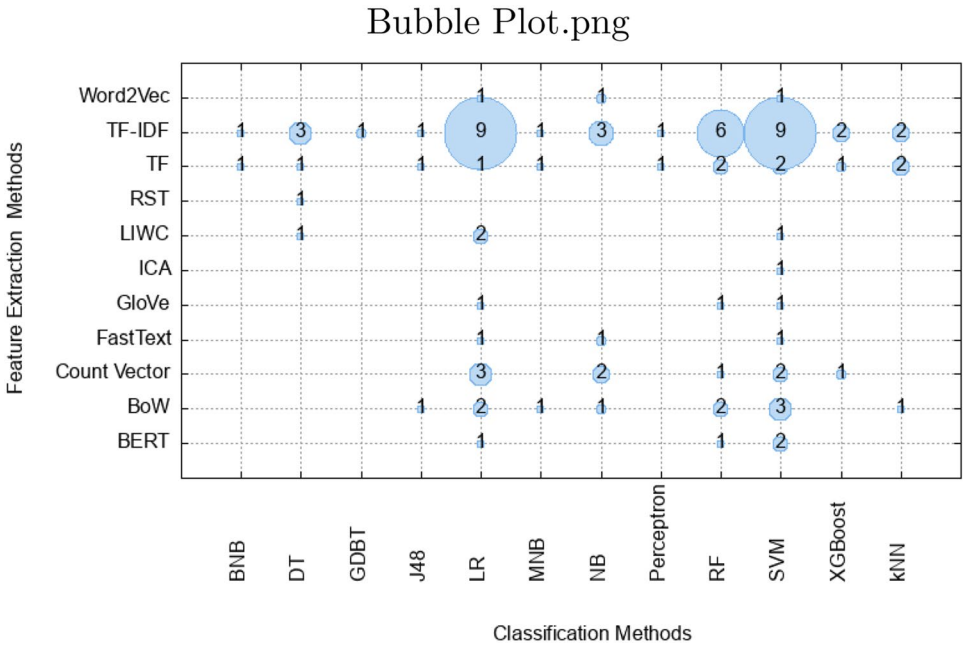

Fig. 5 Relationships between feature extraction and traditional ML techniques

authors trained these models with both the original and augmented datasets. They achieved relatively higher test accuracy from the models trained on augmented data. Among these models, the augmented LR model gained the highest accuracy score of 95.4% in the classification of COVID-19 claims. In Ng and Carley (2021), the authors experimented on the validity classification of stories regarding COVID-19. For this task, the authors used three machine-learning algorithms, i.e., LR, SVM, and NB, with two types of word embeddings. They trained the classifiers with the enhanced BoW representation that includes TF-IDF as a weighting scheme. They also used BERT word embeddings with two of the above classifiers (SVM and LR). In the classification step, the LR model (with enhanced BoW representation) showed relatively higher performance among the classifiers with an average F1-score of 89%.

Mahlous and Al-Laith (2021) employed some ML algorithms with different feature representations for the classification of Arabic tweets regarding COVID-19. They used count vector, word-level TF-IDF, n-gram-level TF-IDF, and character-level TF-IDF features with the classification models such as NB, LR, SVM, and XGB. Besides, the authors trained the classifiers on the corpus without pre-processing (i.e., raw text) and with pre-processing (i.e., stemming and rooting) steps. In the classification task, the LR model using the count vector feature showed the highest performance among all the models. This study reports an F1-score of 93.3% from the LR model which was trained on raw data.

In Shi et al. (2020), the authors introduced a model using the XGBoost ensemble learning algorithm, where 16 basic features of four types such as text characteristic, user-related, interaction-based, and emotion-based features are used in their collected rumor data from microblog. They showed that the accuracy of the model is not satisfactory when these features are used individually. Among the four types of features, the model using user-related features achieved the highest accuracy, reaching 87% and the model of interaction-based features achieved the highest precision, reaching 94%. However, by combining all four types of features, a model with 91% accuracy can be achieved, which is higher than the accuracy of each feature separately.

The relationship between the feature extraction and traditional ML methods is shown in Fig. 5. Here, the bubbles contain the number of articles that employed the classification method (expressed on the X-axis) along with the feature extraction method (expressed on the Y-axis). From this figure, it can be seen that a maximum number of nine studies used the TF-IDF as a feature extraction method before applying both the logistic regression (LR) and support vector machine (SVM) models. However, six studies employed the Random Forest (RF) model using the TF-IDF features. The total count of the studies that use other combinations of feature extraction and classification methods is also illustrated accordingly in the figure.

4.4.2 DL methods

Over the last few years, DL is playing a vital role in misinformation detection tasks. Various DL methods have already been used to conduct the classification task of misinformation in the pre-COVID situation. During this COVID-19 situation, DL has emerged as one of the significant

Table 8 DL methods used in the literature

Methods	Papers
NN	Elhadad et al. (2020); Kar et al. (2020)
DNN	Cheng et al. (2021)
MLP	Kar et al. (2020); Mahlous and Al-Laith (2021)
Transformer	Yang et al. (2021)
BERT-base	Chen (2020); Shahi and Nandini (2020); Koirala (2020); Medina Serrano et al. (2020); Song et al. (2020); Boukouvalas et al. (2020); Kumar et al. (2021); Alam et al. (2021); Dharawat et al. (2020); Chen et al. (2021); Wani et al. (2021); Ayoub et al. (2021); Bang et al. (2021); WANG et al. (2021)
BERT-large	Kumar et al. (2021); Chen et al. (2021); Bang et al. (2021)
Distil-BERT	Kumar et al. (2021); Wani et al. (2021); Ayoub et al. (2021)
mBERT	Alam et al. (2021)
AraBERT	Alam et al. (2021)
RoBERTa-base	Medina Serrano et al. (2020); Kumar et al. (2021); Alam et al. (2021); Chen et al. (2021); Bang et al. (2021)
RoBERTa-large	Kumar et al. (2021); Chen et al. (2021); Bang et al. (2021)
Distil-RoBERTa	Kumar et al. (2021)
ALBERT-base	Alam et al. (2021); Kumar et al. (2021); Chen et al. (2021); Bang et al. (2021)
ALBERT-large	Kumar et al. (2021); Chen et al. (2021)
ALBERT-xlarge	Chen et al. (2021)
CT-BERT	Chen et al. (2021); Wani et al. (2021)
Covid-bert-base	Wani et al. (2021)
Ro-CT-BERT	Chen et al. (2021)
XLNet	Medina Serrano et al. (2020)
CNN	Cui and Lee (2020); Koirala (2020); Elhadad et al. (2021); Alkhalifa et al. (2020); Dharawat et al. (2020); Wani et al. (2021); WANG et al. (2021)
RCNN	Elhadad et al. (2021)
MCNNet	Kaliyar et al. (2020)
TextCNN	Chen (2020); Kumar et al. (2021); Zhou et al. (2020a); Yang et al. (2021)
TextRNN	Chen (2020); Yang et al. (2021)
Att-TextRNN	Yang et al. (2021)
LSTM	Koirala (2020); Boukouvalas et al. (2020); Kumar et al. (2021); Wani et al. (2021); WANG et al. (2021)
BiLSTM	Hossain et al. (2020); Boukouvalas et al. (2020); Kumar et al. (2021); Dharawat et al. (2020); WANG et al. (2021)
BiLSTM-Attention	Wani et al. (2021)
BiGRU	Cui and Lee (2020)
Sequential Model	Elhadad et al. (2021)
SBERT	Hossain et al. (2020)
SBERT (DA)	Hossain et al. (2020)
XLM-r	Alam et al. (2021)
FastText	Alam et al. (2021); Yang et al. (2021)
SCHOLAR	Song et al. (2020)
SAFE	Zhou et al. (2020a)
SAME	Cui and Lee (2020)
HAN	Cui and Lee (2020); Dharawat et al. (2020); Wani et al. (2021)
Cross-SEAN	Paka et al. (2021)
dEFEND	Cui and Lee (2020); Dharawat et al. (2020)
CSI	Cui and Lee (2020)
CANTM	Song et al. (2020)

technologies to make efficient systems that can detect and classify the misinformation related to COVID-19. Several DL methods have been employed in the existing studies of COVID-19 misinformation detection and classification task (see Table 8). These methods are thoroughly reviewed here in the next.

Neural networks (NNs) are the most basic architectures among the DL methods. Few studies on COVID-19 misinformation detection employed NN as a classification model. For instance, Elhadad et al. (2020) implemented a NN model with different feature extraction techniques such as TF, TF-IDF, and Word Embedding to construct a voting ensemble system. In this study, the authors proposed an ensemble system that takes the output of the NN model and uses it to classify the misleading information on COVID-19. They achieved an accuracy of 99.68% and an F1-score of 99.80% from the NN classification model. The authors admitted that the selection of appropriate feature extraction methods along with the use of reliable ground-truth data results in such good outcomes. In Kar et al. (2020), the authors employed a multi-layer perceptron (MLP) model using pre-trained BERT embeddings and a NN model using multilingual BERT (mBERT) embedding for the classification of COVID-19 fake tweets in Indic-Languages (e.g., Hindi, Bengali) along with the English language. The MLP model did not show good performance due to the smaller size of the dataset. But, the NN model was able to deal with the smaller data size problem and achieved more than 80% F1-scores in both monolingual (for English) and multilingual (for English, Hindi, and Bengali) settings. Mahlous and Al-Laith (2021) experimented with an MLP model using different feature representations (e.g., count vector, TF-IDF) for the classification of Arabic tweets regarding COVID-19. This study reports a maximum F1-score of 88.6% from the MLP model using count vector features. Another study Elhadad et al. (2021) used a sequential model with a GloVe embedding vector to detect misleading information related to COVID-19. Cheng et al. (2021) proposed a system for COVID-19 rumor veracity classification based on deep neural networks (DNN). The authors used an LSTM-based variational autoencoder (VAE) (Cheng et al. 2020) followed by the pre-trained BERT model to extract significant features from the vectors of textual contents. A DNN classifier takes these features as input and gives the classification result. This study reports an average F1-score of 85.98% from the DNN classifier in the veracity classification of rumors.

CNN is one of the most popular and widely used models in NLP tasks. Similarly, some of the existing studies on COVID-19 misinformation classification also adopted CNN and its other variants for classification purposes. For example, Cui and Lee (2020) implemented a CNN model for detecting COVID-19 healthcare misinformation. They used word embedding initialized by GloVe and fed it into the CNN model. In Elhadad et al. (2021), the authors deployed a CNN model using pre-trained GloVe embedding to build up a system for detecting misleading information related to COVID-19. They utilized the word-level representation of features to preserve their order which enabled them to obtain high accuracy in results. WANG et al. (2021) also used a CNN model with FastText, Word2vec, and GloVe but the models could not achieve good results in their rumor-related dataset. Alkhalifa et al. (2020) introduced a CNN-based classification system with different pre-processing approaches and embedding methods to classify the COVID-19 rumors. In this work, the best performing model comprises a CNN model with COVID-Twitter-BERT (CT-BERT) (Müller et al. 2020) embedding which is pre-trained on COVID-19 Twitter data. Another study Koirala (2020) applied a CNN model with an embedding layer in front of it for the classification of fake news related to COVID-19. This study reported that lower weights of minority classes cause overfitting problems. By increasing the weights of the minority class, the author was able to reduce the overfitting problem significantly and increased the test accuracy as well. Wani et al. (2021) experimented with a CNN model using two types of word embeddings (e.g., GloVe and FastText) for the classification of COVID-19 fake news. They achieved an accuracy of 93.50% using GloVe embeddings and an accuracy of 94% using FastText embeddings in the classification task from the CNN model. Kaliyar et al. (2020) proposed a generalized fake news detection system called MCNNet using a multichannel CNN architecture. This architecture uses different sized kernels and filters in different parallel CNN networks. It concatenates different channel features into a single vector and uses some dropout layers to provide generalization capability in the classification of fake news. The authors experimented with this model on two different COVID-19 fake news datasets named FN-COV and CoAID. Although MCNNet has the ability to generalize any fake news detection task, it showed relatively higher accuracy in the FN-COV dataset. This study reports an accuracy of 98.2% and an F1-score of 98.1% with MCNNet from these datasets. Moreover, the authors used an attention-based CNN (AttCNN) model with a fake news dataset (not related to COVID-19) for their experimental purpose.

Dharawat et al. (2020) introduced a dataset for health risk assessment of COVID-19 misinformation. The authors also experimented with CNN to classify the misinformation categories using both binary and multi-class classification methods. They implemented CNN with multiple kernels and used pre-trained GloVe embedding as an initialization of word embedding. Among all the studies that experimented with CNN, the study Elhadad et al. (2021) achieved the highest performance with CNN by reporting the accuracy and F1-score of 99.999 % and 99.966 %, respectively. In some studies, the authors used TextCNN, a CNN architecture for text classification, to classify COVID-19 rumors (Chen

2020), fake news (Zhou et al. 2020a; Yang et al. 2021) and misinformation (Kumar et al. 2021) in COVID-19 tweets. The TextCNN model uses a one-dimensional convolution layer and max-over-time pooling layer to capture the associations between the neighboring words in texts. The study Chen (2020) obtained the highest performance with accuracy and F1-score of 98.40% and 97.24%, respectively, among the studies that adopted the TextCNN model. Elhadad et al. (2021) used an RCNN model which combines the properties of RNN and CNN to detect COVID-19 misleading information. In the RCNN architecture, a recurrent structure is responsible to capture the contextual information and the max-pooling layer can easily determine the words which are playing the key roles in the texts (Lai et al. 2015). In this study, RCNN performed very well with an accuracy of 99.997%. The authors were able to attain such accuracy by integrating the properties of RNN and CNN into one and fine-tuning the hyperparameters to their optimum levels.

RNN has the ability to capture better contextual information from the texts. Therefore, various studies utilized RNN and its other variants for the classification of COVID-19 misinformation. In particular, the studies Chen (2020); Yang et al. (2021) used the TextRNN (Liu et al. 2016) model to classify COVID-19 rumors and fake news, respectively. TextRNN model uses different LSTM layers inside its architecture. In Chen (2020), higher accuracy (98.40%) was obtained in the classification results as TextRNN was able to strongly capture the relationship between the semantics and the contexts of the texts. As LSTM (Hochreiter and Schmidhuber 1997) has the advantage of learning long-term dependencies over RNNs, some studies implemented the LSTM model for the better classification of misinformation related to COVID-19 (Koirala 2020; Boukouvalas et al. 2020; Kumar et al. 2021; Wani et al. 2021). Among these studies, Wani et al. (2021) achieved the best performance from the LSTM network with an accuracy of 94.95%. WANG et al. (2021) also employed the LSTM model with different types of word embedding techniques such as FastText, Word2vec, and GloVe but the model's performance is not satisfactory in their rumor related dataset. Some of the studies applied the BiLSTM model which is an extension of the LSTM architecture. A BiLSTM model can also learn long-term dependencies and reserve contextual information in both the forward and backward directions. Hossain et al. (2020) used the BiLSTM model to classify tweet-misconception pairs related to COVID-19. Other studies had implemented BiLSTM for the classification of COVID-19 misinformation (Kumar et al. 2021; Dharawat et al. 2020; Boukouvalas et al. 2020). Dharawat et al. (2020) came up with an accuracy of 96.6% from the BiLSTM model using pre-trained GloVe embedding and this is the highest accuracy among the studies that employed BiLSTM for the classification purpose. Other two studies Wani et al. (2021) and Yang et al. (2021) employed an attention-based BiLSTM (Zhou et al. 2016) model in the fake news classification task. This architecture includes a BiLSTM layer followed by an attention layer. In Wani et al. (2021), the authors used both GloVe and FastText embeddings as input to the classification model. They achieved relatively higher performance using FastText embeddings with an accuracy of 94.71% from this model. Moreover, Cui and Lee (2020 deployed a model called BiGRU for the classification of healthcare misinformation related to COVID-19. BiGRU is a variant of RNN that consists of two GRU (Chung et al. 2014) models. Like BiLSTM, it can also learn long-term dependencies in both forward and backward directions with only the input and forget gates. The authors used word embeddings to the BiGRU model which was initialized by GloVe embedding. But they did not achieve good results using BiGRU due to their imbalanced data.

BERT is a newer DL method that has been extensively used for dealing with NLP tasks. Several exciting studies focused on BERT and its variants for classification purposes. For instance, Chen (2020) proposed a fine-grained classification method based on the BERT pre-training model to classify the rumors of COVID-19. The author fine-tuned the pre-trained BERT model for classification purposes. This study demonstrates that the multi-head attention mechanism used in BERT is capable to produce outstanding results. The author reported accuracy of 99.20% in the classification results using the BERT model. According to the author, the BERT model has the ability to pay attention to the corresponding word in a sentence from different angles, and a stronger potential to capture the distance information for a specific word after adding the positional embedding. Therefore, the BERT model can better express contextual semantics and show higher performance in terms of accuracy. Alam et al. (2021) proposed a multilingual model called mBERT to analyze the COVID-19 disinformation. The authors trained this model with combined English and Arabic tweets. They achieved good performance scores in both monolingual and multilingual settings using the mBERT model. Shahi and Nandini (2020) performed a BERT-based classification of real or fake news on COVID-19 by introducing a multilingual cross-domain dataset. Kumar et al. (2021) conducted a fine-grained classification of misinformation in COVID-19 tweets. The authors also applied several transformer language models including three variants of the BERT model (e.g., Distil-Bert, BERT-base, and BERT-large), three variants of the RoBERTa model (e.g., Distil-RoBERTa, RoBERTa-base, and RoBERTa-large), and two variants of ALBERT model (e.g., Albert-base and Albert-large) to perform a systematic analysis. They performed fine-tuning on these pre-trained models to get them ready for their classification task. Among all the adopted models, Roberta-large

appeared the best performing model with an F1-score of 76% as it was trained on a larger corpus compared with the other models. Bang et al. (2021) presented a model with robust loss and influence-based cleansing for the COVID-19 fake-news detection task. They fine-tuned transformers-based language models (LM) (e.g., ALBERT-base, BERT-base, BERT-large, RoBERTa-base, and RoBERTa-large) with robust loss functions (e.g., symmetric cross-entropy (SCE), the generalized cross-entropy (GCE), and curriculum loss (CL)). Among all of them, RoBERTa-large using cross-entropy (CE) loss function achieved a good accuracy of 98.13% on the Fake News-19 test set. For influence-based cleaning, they fine-tuned a pre-trained RoBERTa-large model with the FakeNews-19 train set. With 99% data cleansing, their best model achieved a 61.10% accuracy score and 54.33% weighted F1-score on Tweets-19. In Medina Serrano et al. (2020), the authors fine-tuned three transformer models e.g., XLNet base, BERT base, and RoBERTa base for the classification of user comments associated with COVID-19 misinformation videos. Among these models, RoBERTa showed the best performance in test data. Chen et al. (2021) used different variants of the pre-trained transformer language models (e.g., BERT, RoBERTa, and ALBERT) along with the CT-BERT model for the classification of COVID-19 fake news. They also proposed a robust classification model called Robust-COVID-Twitter-BERT (Ro-CT-BERT) which performs a feature-level fusion on the features extracted from the CT-BERT and RoBERTa models. This model involves adversarial training to improve the robustness and generalization ability in the fake news detection task. The authors achieved an accuracy of 99.02% with the same F1-score from the Ro-CT-BERT model which outperformed all other models in the classification performance. Wani et al. (2021) fine-tuned the pre-trained BERT and DistilBERT models for their classification task. For domain adaptation, the authors further trained BERT and DistilBERT as a language model (LM) with a corpus of COVID-19 tweets. They also used CT-BERT and Covid-bert-base[6] models which have domain adaptive pre-training on COVID corpus. Among the adopted models, the BERT model having LM pre-training achieved the highest accuracy of 98.41% in the classification of COVID-19 fake news.

Ayoub et al. (2021) performed fine-tuning on pre-trained BERT and DistilBERT models for the classification of COVID-19-related claims. The authors trained these models with both original and augmented datasets. In the knowledge distillation process from BERT, they also trained a logistic regression model with DistilBERT. In the classification task, BERT showed relatively higher performance than DistilBERT in both original and augmented data. The augmented BERT achieved an accuracy of 99.4% which is slightly higher than the accuracy (97.2%) obtained from the augmented DistilBERT model. Hossain et al. (2020) employed the Sentence-BERT (SBERT) (Reimers and Gurevych 2019) and SBERT (DA) models for their classification purpose. The SBERT model is a modification of pre-trained BERT architecture that uses siamese and triplet networks to extract semantically meaningful sentence embeddings. On the other hand, SBERT (DA) uses the SBERT representation with domain adaptive pre-training on COVID-19 tweets. In this work, the authors utilized COVID-Twitter-BERT embedding for domain adaptation purposes. The study Alam et al. (2021) showed the classification of COVID-19 disinformation both in English and Arabic languages adopting binary and multiclass classification settings. The authors fine-tuned the pre-trained BERT, RoBERTa, and ALBERT models for English language experiments. BERT outperformed all other models in the case of the English language. For Arabic language experiments, they employed the AraBERT model (Antoun et al. 2003) which is pre-trained on a large corpus of 70 million Arabic sentences. Due to the smaller size of the Arabic dataset used in the training, AraBERT did not perform very well in their study. Some other studies experimented with BERT for the classification of COVID-19 fake news (Koirala 2020), COVID-19 disinformation (Song et al. 2020), COVID-19 rumor (WANG et al. 2021), and COVID-19 misinformation (Boukouvalas et al. 2020; Dharawat et al. 2020).

Other methods. Apart from the above methods, researchers also applied other DL methods for classification purposes. For example, Song et al. (2020) proposed a classification-aware neural topic model called CANTM for topic modeling tasks by taking into account the classification information regarding COVID-19 disinformation. They accumulated the properties of the BERT with a VAE model to build up a robust classification system. The authors also used the SCHOLAR model (Card et al. 2018) for their experiment. SCHOLAR uses the functionality of the VAE framework in document modeling tasks. In the classification of COVID-19 disinformation, CANTM outperformed other baseline models in terms of accuracy and F1-score. It also achieved the best perplexity score (the measurement of how well a probability distribution or probability model predicts a sample) in the topic modeling task among all the models.

Some studies Cui and Lee (2020); Dharawat et al. (2020) employed attention-based models for the classification of COVID-19 misinformation. The authors used two models based on the attention mechanism namely HAN (Yang et al. 2016) and dEFEND (Shu et al. 2019a) for their purposes. HAN learns the hierarchical structure of the documents by using two levels of attention mechanisms applied at

[6] https://huggingface.co/deepset/covid_bert_base/tree/main.

the word and sentence levels. It uses a bidirectional GRU network for word and sentence level encoding procedures. An attention mechanism is used after the word encoder to extract the contextually important words and form a sentence vector by aggregating the representations of the informative words. A sentence encoder then works on the derived sentence vectors and generates a document vector. Another attention mechanism is used after the sentence encoder to measure the importance of sentences in the classification of a document. The dEFEND framework builds upon the HAN architecture. It involves the HAN on text content and a co-attention mechanism between the text content and user comments to classify misinformation. In Cui and Lee (2020) and Dharawat et al. (2020), dEFEND showed higher performance scores than HAN for its robustness. Wani et al. (2021) employed HAN with different word embeddings for the classification of COVID-19 fake news. They achieved 95% accuracy with FastText embeddings and 94.25% accuracy with GloVe embeddings from the HAN model.

In Zhou et al. (2020a), multi-modal information (e.g., textual and visual) of new articles on coronavirus was used for the detection of fake news. The authors adopted the SAFE model (Zhou et al. 2020b) which can jointly learn the textual and visual information along with their relationships to detect fake news. In SAFE architecture, a Text-CNN model is used to extract the textual features from the news articles and the visual features (e.g., images) are also extracted by the Text-CNN model while the visual information within the articles is first processed using a pre-trained image2sentence model. The authors achieved the best performance using the SAFE model among all the baseline methods employed.

Cui and Lee (2020) employed a model called SAME (Cui et al. 2019) for the classification of healthcare misinformation regarding COVID-19. SAME is a multi-modal system that uses news images, content, user profile information as well as users' sentiments to detect fake news. In this study, the authors skipped the visual part of the SAME model for their classification purpose as the majority of the news articles do not contain any cover images. They were not able to get satisfactory results from this model due to the imbalanced dataset. The authors also used a hybrid DL model called CSI (Ruchansky et al. 2017) for their experimental purpose. CSI explores news content, user responses to the news, and the sources that users promote for the detection of fake news. The authors' utilized GloVe embeddings as input features to the CSI model. Due to the imbalanced data, CSI also could not achieve good results in the classification task.

Paka et al. (2021) introduced Cross-SEAN which is a cross-stitch-based semi-supervised neural attention model. This model helps to reduce the dependency on the labeled data as it leverages unlabeled data. It uses tweet text, user metadata, tweet metadata, and external knowledge for each tweet as inputs. The cross-stitch unit is employed among tweet and user features for optimal sharing of parameters. They used sentence BERT to get contextual embedding of the external knowledge and Bi-LSTM with word embedding for encoded tweet text. As the similarity between tweet text and tweet features is close, they performed optimal sharing of information by concatenating one output of cross-stitch early in the network with the other afterward. They employed different types of the objective function. For supervised loss, they used maximum likelihood and adversarial training and virtual adversarial training for unsupervised loss. Compared with seven state-of-the-art models, they showed that it achieved a 95% F1-score on their CTF dataset and outperformed the best baseline by 9%.

Some other methods such as XLM-r and FastText were used to perform fine-grained disinformation analysis on Arabic tweets (Alam et al. 2021). In this study, the authors used these two models in both binary and multi-class classification settings. They achieved consistent and good results using FastText while XLM-r did not perform well as the amount of data was small and it was likely to overfit. In another study Yang et al. (2021), the authors used the FastText and Transformer model (Vaswani et al. 2017) in the classification of Chinese microblogs regarding COVID-19. The authors used them as baseline methods and achieved a macro F1-score of 92.7% from the transformer model outperforming the other.

Figure 6 represents the relationship between the feature extraction and DL methods (including combined DL methods) used in existing studies. Here, the bubbles contain the number of articles that employed the classification method (expressed on the X-axis) along with the feature extraction method (expressed on the Y-axis). The figure illustrates that a maximum of six studies employed the CNN model using the GloVe extracted features. Three studies used GloVe feature vectors with the LSTM model. Moreover, two studies applied LSTM with FastText extracted features, two studies applied HAN with GloVe extracted features, and the two more studies applied CNN with FastText extracted features. The total count of the studies that use other combinations of feature extraction and classification methods is also illustrated accordingly in the figure.

4.4.3 Combined methods

Some research works also used different combinations of traditional ML and DL techniques to increase the overall performance of classification (see Table 9).

Traditional ML with Traditional ML. In the study Al-Rakhami and Al-Amri (2020), the authors proposed an ensemble-learning-based framework where they used tweet-level and user-level features for justifying the credibility of the tweets. They used six traditional ML algorithms utilizing stacking-based ensemble learning for getting higher

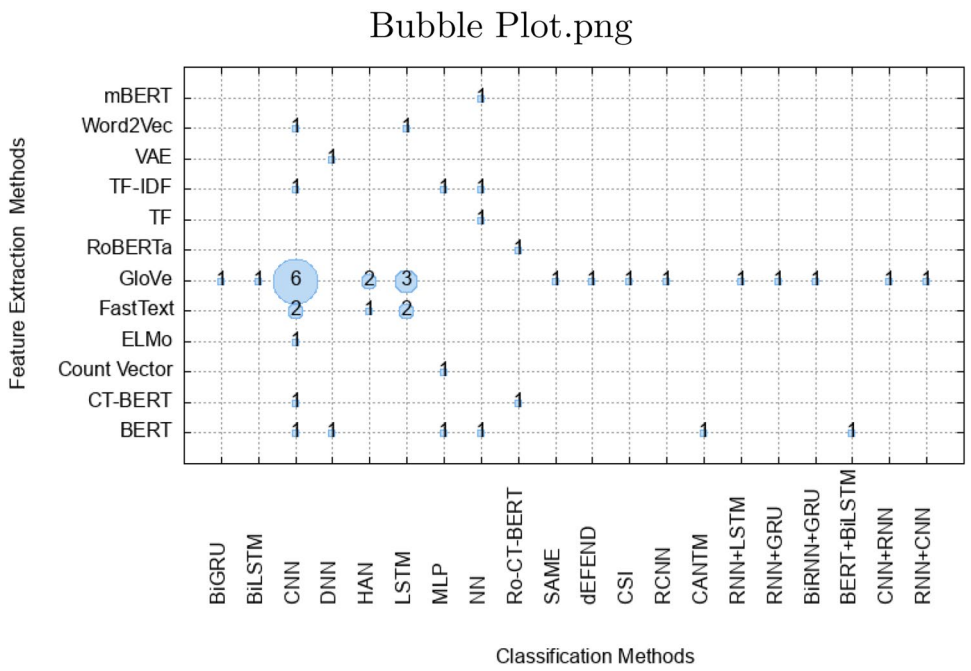

Fig. 6 Relationships between feature extraction and DL techniques

Table 9 Combined Models Used in Literature

Methods	Papers
ML+ML	
C4.5 + RF	Al-Rakhami and Al-Amri (2020)
C4.5 + kNN	Al-Rakhami and Al-Amri (2020)
SVM + RF	Al-Rakhami and Al-Amri (2020)
SVM + kNN	Al-Rakhami and Al-Amri (2020)
SVM + Bayes Net + kNN	Al-Rakhami and Al-Amri (2020)
C4.5 + Bayes Net + kNN	Al-Rakhami and Al-Amri (2020)
C4.5 + SVM + RF+ BayesNet + kNN + Naive Bayes	Al-Rakhami and Al-Amri (2020)
DL+DL	
RNN+LSTM	Elhadad et al. (2021)
RNN+GRU	Elhadad et al. (2021)
BiRNN+GRU	Elhadad et al. (2021)
BERTSCORE (DA)+BiLSTM	Hossain et al. (2020)
BERTSCORE (DA)+SBERT (DA)	Hossain et al. (2020)
BERT+BiLSTM	WANG et al. (2021)
CNN+RNN	Kumar et al. (2021)
RNN+CNN	Kumar et al. (2021)
C-LSTM	Kaliyar et al. (2021)

accuracy. For constructing the ensemble model, they carried out various experiments. They used the SVM+RF models for a level-0 weak learner and the C4.5 model as a meta-model for a level-1 weak learner. They also used different types of combinations for their experiment such as C4.5+RF, C4.5+kNN, SVM+kNN, SVM+ BN+kNN and C4.5+BN+kNN.

DL with DL. Kumar et al. (2021) proposed a CNN-RNN model (CNN layer stacked over the RNN layer) and an RNN-CNN model (a single BiLSTM layer is employed over the top of a 1D-CNN layer) with a word embedding in the first layer for the classification of misinformation in COVID-19 tweets. Kaliyar et al. (2021) proposed a hybrid model called C-LSTM for the classification of COVID-19 fake news. C-LSTM architecture consists of a CNN block followed by an LSTM network. The CNN block takes the word-embedding vector as input and performs the automatic feature extraction using different-sized kernels and filters. The LSTM network then takes the output of the CNN block and learns sequential information from the texts. The authors set optimal hyperparameters for the C-LSTM model that showed higher performance in the fake news classification task with an accuracy of 98.62% and an F1-score of 99.4%. Another study Elhadad et al. (2021) represents three DL models (e.g., RNN-LSTM, RNN-GRU, and BiRNN-GRU) which are the combinations of various recurrent neural networks. These models use pre-trained GloVe embedding in the first layer of each model and together constitute an ensemble DL system for detecting COVID-19 misleading information. The authors achieved very high performance from these models with more than 99% accuracy in every case. WANG et al. (2021) applied a bert-base-uncased pre-trained model containing 12 layers where they fed the hidden layer of BERT into BiLSTM. After fine-tuning the model, it achieved 72.95% in terms of F1-score which made this approach better for their rumor dataset compared to other

methods used in their study. Hossain et al. (2020) proposed a system that uses combinations of BERTSCORE (DA) with BiLSTM and BERTSCORE (DA) with SBERT (DA) models for detecting COVID-19 misinformation on social media. BERTSCORE (DA) represents BERTSCORE (Zhang et al. 2020) with domain-adaptive pre-training on COVID-19 tweets. The BERTSCORE (DA) with BiLSTM model uses BERTSCORE (DA) to retrieve relevant misconceptions and a BiLSTM model for classifying tweet-misconception pairs. On the other side, BERTSCORE (DA) with SBERT (DA) model uses the combination of BERTSCORE (DA) and the Sentence-BERT representation with domain-adaptive pre-training for the classification of tweet-misconception pairs.

4.5 Evaluation metrics

For evaluating the performances of models, different kinds of evaluation metrics are used such as accuracy, precision, recall, and F1-score. Many of these metrics are known by multiple names. Confusion Matrix, a tabular representation of a classification model, is used to get the necessary values for all of these metrics. This tabular representation is based on the performance of the test set which includes four parameters. They are true positive (TP), false positive (FP), true negative (TN), and false negative (FN) which are calculated based on the predicted class versus actual class (ground truth).

Accuracy. The ratio of accurately predicted instances to the total number of evaluated instances is known as accuracy. It is formally defined in Eq. 1.

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

Precision. It is also known as positive predictive value (PPV), which is defined as the correctly predicted positive instances from the total predicted instances in a positive class. It is formally defined in Eq. 2.

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

Recall. It is also known as true positive rate (TPR) or sensitivity, which is defined as the measurement of the fraction of positive instances that are correctly classified. It is formally defined in Eq. 3.

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

F1-score. It is the weighted average of Precision and Recall. It is formally defined in Eq. 4.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4.6 Evaluation results

In the existing research on COVID-19 misinformation classification, several traditional ML and DL methods have been employed. Among them, some are highly efficient in the classification of COVID-19 misinformation and achieve superior performance. Table 10 represents the best-performing models used in the existing studies on COVID-19 misinformation detection in terms of accuracy and F1-score. All the results are retrieved from the original articles. The best-performing model is typically chosen based on the accuracy metric but if there is no accuracy measure in the study, then F1-score is considered to choose the best model outperforming other models used in that study.

The CNN (Elhadad et al. 2021) model obtained 99.99% accuracy and 99.97% F1-score which are the highest among all the classification methods used in existing studies. This study employed a word-level representation of the texts in order to preserve the sequence of the words in a sentence which significantly improved the accuracy of the CNN model. Moreover, fine-tuning the hyperparameters of the CNN architecture to their optimum values helped to attain higher performance. This CNN model was trained and evaluated on the COVID-19-FAKES (Elhadad et al. 2020) dataset using an 80:20 train-test split. The NN (Elhadad et al. 2020) model achieved an accuracy of 99.68 % with the TF feature by using fivefold cross-validation instead of externally splitting the data into training, testing, and validation set. This model also showed an F1-score of 99.80% in the experiment. BERT (Ayoub et al. 2021) scored an accuracy of 99.4% using tenfold cross-validation on an augmented dataset, and this is the highest test accuracy obtained from any BERT models employed in the existing studies. Ro-CT-BERT (Chen et al. 2021) model outperformed all other methods used in this study with an accuracy of 99.02% from COVID-19 Fake News Dataset (Patwa et al. 2021). The C-LSTM (Kaliyar et al. 2021) model holds a hybrid architecture of CNN and LSTM networks which achieved an accuracy of 98.62% along with a 99.4% F1-score in the FN-COV dataset. MCNNet (Kaliyar et al. 2020) uses a multi-channel CNN architecture for the generalization purpose which showed an accuracy of 98.2% in the same dataset. In the study Dharawat et al. (2020), the dEFEND model that uses the co-attention mechanism got an accuracy of 98% in the binary classification task. In this study, the LR model attained an accuracy of 96.3% which outperformed other models used for the multi-class classification setting. An ensemble model named SVM+RF (Al-Rakhami and Al-Amri 2020) achieved a 97.8% accuracy score by applying tenfold cross-validation on the dataset. Other methods such as Cross-SEAN (Paka et al. 2021), C4.5 (Al-Rakhami and Al-Amri 2020), SVM (Patwa et al. 2021), XGBoost (Shi et al. 2020), kNN (Bandyopadhyay and

Table 10 Best performing models based on accuracy and F1-score

Problem tackled	Reference	Best model	Split ratio (%) Train, Validation, Test	Split count			Performance Metrics			
				Train	Validation	Test	A (%)	P (%)	R (%)	F1 (%)
Misleading	Elhadad et al. (2021)	CNN	80,20,–	5989	1497	–	99.99	99.93	100	99.97
	Elhadad et al. (2020)	NN (TF)	Fivefold C-V	–	–	–	99.68	99.87	99.73	99.80
Fake news	Kar et al. (2020)	*mBERT_NN* (monolingual)	80 , – , 20	1150	–	288	–	87.17	91.89	89.47
		mBERT_NN (multilingual)					–	76.47	86.66	81.25
	Koirala (2020)	CNN	54,26 ,20	1672	823	624	80	73	70	72
		BERT					80	–	–	–
	Bang et al. (2021)	RoBERTa-large (Fakenews-19)	– , – , –	6420	2140	2140	98.13	–	–	98.13
		BERT-large (Tweets-19)			60	200	61.10	–	–	54.33
	Hossain et al. (2020)	BERTSCORE (DA) + BiLSTM (SNLI)	–, –, –	–	–	–	–	44.20	45.30	43.10
		BERTSCORE (DA) + SBERT (DA) (MultiNLI)					–	55.90	50.90	50.20
		BERTSCORE (DA) + SBERT (DA) (MedNLI)					–	47.80	49.20	48.40
	Boukouvalas et al. (2020)	$BERT_{BASE}$ (DNN)	70, 30, –	392	168	–	87.50	84.70	90	87.30
		SVM/Gaussian (ICA)					81.20	85.90	76.30	80.30
		SparseICA-EBM ($\lambda = 100$)					69.10	74.40	67.90	64.40
	Al-Rakhami and Al-Amri (2020)	C4.5 (Meta-model)	Tenfold C-V	–	–	–	95.11	95.3	95.10	95.10
		SVM+RF (Ensemble-model)					97.80	–	–	–
	Dharawat et al. (2020)	LR(Multiclass)	80, –, 20	49,029	–	12,257	96.30	31.30	23.30	25
		dEFEND w.news (Binary)					98	92	68	75
	Chen et al. (2021)	Ro-CT-BERT	60 ,20 ,20	6420	2140	2140	99.02	99.02	99.02	99.02
	Kaliyar et al. (2021)	C-LSTM	–, –, –	–	–	–	98.62	99.20	98.9	99.40
	Wani et al. (2021)	BERT	60, 20, 20	6420	2140	2140	98.41	–	–	–
	Ayoub et al. (2021)	BERT	Tenfold C-V	–	–	–	99.40	99.40★	99.35★	99.4★
	Ng and Carley (2021)	BoW + LR (Story Validity Classification)	80, –, 20	5385	–	1346	–	–	–	89
	Mahlous and Al-Laith (2021)	LR (Count Vector)	Fivefold C-V	–	–	–	–	93.40	93.30	93.30
	Kaliyar et al. (2020)	MCNNet	–, –, –	–	–	–	98.20	97.50	98.70	98.10
	Yang et al. (2021)	TextCNN	70, 10, 20	1473	210	421	–	–	–	93.80
	Shahi and Nandini (2020)	BERT	– ,– ,–	–	–	–	–	78	75	76
	Bandyopadhyay and Dutta (2020)	kNN (k-5)	– , –, –	–	–	–	89	–	–	91
	Paka et al. (2021)	Cross-SEAN	80, –,20	–	–	–	95.40	94.60	96.10	95.30
	Patwa et al. (2021)	SVM	60 ,20 ,20	6420	2140	2140	93.46	93.48	93.46	93.46
	Zhou et al. (2020a)	SAFE	80 ,– ,20	–	–	–	–	75.15★	75.30★	75.25★
	Cui and Lee (2020)	dEFEND	75 , – , 25	2152	–	717	–	89.65	48.47	58.14
	Kumar et al. (2021)	RoBERTa-large	– , – , –	–	–	–	–	73.75	73.50	76

Table 10 (continued)

Problem tackled	Reference	Best model	Split ratio (%) Train, Validation, Test	Split count			Performance Metrics			
				Train	Validation	Test	A (%)	P (%)	R (%)	F1 (%)
Rumor	Chen (2020)	BERT	Tenfold C-V	–	–	–	99.20	99.17	98.13	98.34
	Alkhalifa et al. (2020)	CNN with CT-BERT	97, 2, 1	9206	150	140	–	78	–	–
	Shi et al. (2020)	XGBoost	–, –, –	–	–	–	91	94	85	89
	WANG et al. (2021)	BERT+BiLSTM	80,–,20	6102	–	1077	–	73.19	73.27	72.95
	Alsudias and Rayson (2020)	LR (COUNT VECTOR)	Tenfold C-V	–	–	–	84.03	81.04	80.03	80.50
	Cheng et al. (2021)	DNN	Fivefold C-V	–	–	–	–	–	–	85.98★
Conspiracy	Medina Serrano et al. (2020)	RoBERTa (Classification of Users Comments)	80, –, 20	2582	–	645	–	–	–	90.30★
		SVM (Classification of YouTube Videos)	Tenfold C–V	–	–	–	89.40	–	–	–
Disinformation	Song et al. (2020)	CANTM	Fivefold C-V	–	–	–	63.34	–	–	55.48
	Alam et al. (2021)	BERT (En) (Binary)	Tenfold C-V	–	–	–	–	–	–	85.6★
		mBERT (En) (Multiclass)					–	–	–	53.48★
		mBERT (Ar) (Binary)					–	–	–	83.96★
		FastText (Ar) (Multiclass)					–	–	–	69.52★

En = English; *Ar* = Arabic; *A* = Accuracy; *P* = Precision; *R* = Recall; *F1* = F1-Score ★ = Macro Average (Calculated)

Dutta 2020), SparseICA-EBM (Boukouvalas et al. 2020), CANTM (Song et al. 2020) scored an accuracy of 95.40%, 95.11%, 93.46%, 91%, 89%, 69.1%, 63.34%, respectively, in the existing studies.

Several studies did not use the accuracy metric in the performance evaluation of the classification models. These studies considered other metrics such as F1-score, precision, and recall in the evaluation of performance. In the study Yang et al. (2021), the TextCNN model showed a macro F1-score of 93.8% in the CHECKED dataset. A train, validation, and test split of 70:10:20 was used in this study. LR model (Mahlous and Al-Laith 2021) achieved an F1-score of 93.3% from the dataset named Arabic Fake News corpora, which is the highest F1-score obtained in this study. RoBERTa (Medina Serrano et al. 2020) came up with an F1-score of 90.30%. The mBERT_NN (Kar et al. 2020) model got an F1-score of 89.47% in the monolingual setting (tweets in English) which is 8.22% higher than in the multilingual setting (tweets in English, Hindi, and Bengali). Some other models such as DNN (Cheng et al. 2021), BERT (Alam et al. 2021), SAFE (Zhou et al. 2020a), BERT+BiLSTM (WANG et al. 2021), BERTSCORE (DA) + SBERT (DA) (MultiNLI) (Hossain et al. 2020) scored 85.98%, 85.6%, 75.25%, 72.95%, and 50.20% F1-scores, respectively.

5 Open issues and future research directions

During this COVID-19 pandemic, the propagation of misinformation through various platforms has already become a global concern. It has opened the door for researchers to come up with different ideas to solve this problem. Accordingly, researchers around the world are working on various research works on misinformation detection and classification related to COVID-19. In our systematic survey, we have presented the impact, characteristics, and detection of COVID-19 misinformation along with the research methodologies of the existing efforts. Researchers have proposed and implemented various techniques for the detection and classification of misinformation about COVID-19. Some of them are very efficient to classify misinformation with high accuracy value and some are not that much which can be taken into consideration for further improvements. Moreover, the number of notable works on COVID-19 misinformation detection is still not that big. Thus, we have pointed out several findings from these research works and the promising research directions for the future.

5.1 Quality of the datasets

It is observed that there is still a lack of benchmark datasets that include all relevant features related to COVID-19 misinformation. Besides, most of the studies have utilized the data that is mainly collected from social network platforms (e.g., Twitter, Facebook, etc.) and some other reliable sources. The majority of the datasets do not contain data from various sources. Moreover, class distribution in some datasets was observed to be imbalanced which affected the overall performance of the classification. Koirala (2020) showed that an increase in the weight of the minority class can handle this problem. A promising direction is to create a comprehensive, well-annotated, and large-scale benchmark dataset on COVID-19 misinformation which can be used by scholars to conduct further research in this domain. Furthermore, future researchers may employ and investigate different sampling techniques to handle the class imbalance problem and demonstrate their effect on classification performance.

5.2 Classifying multilingual misinformation

Detection of COVID-19 misinformation in multiple languages is still a challenging task because it requires multilingual data and also more pre-processing tasks. It was observed that most of the studies only used English language data for building up a classification system. Only a few studies utilized multilingual data to classify misinformation in multiple languages (Kar et al. 2020; Shahi and Nandini 2020; Alam et al. 2021). So, it can be a great scope for future researchers to work with multilingual misinformation data related to COVID-19. Moreover, there exists only one study in non-English languages on the classification of misinformation. Alsudias and Rayson (2020) used Arabic data for their classification model. So, researchers worldwide can conduct a similar type of study by considering the misinformation related to COVID-19 in their respective languages.

5.3 Pre-processing and feature selection for large volume data

In misinformation classification, data pre-processing is an underrated step. It was observed that most of the researchers give more focus on the method and often neglect the data pre-processing step. Elhadad et al. (2021) showed that with proper data-pre-processing approaches, the performance of the classification model can be improved significantly. Generally, in the pre-processing phase, punctuation marks, tags, URLs, special characters, and stop-words are eliminated, and Part of Speech (PoS) tagging, word stemming, case-folding, etc., are performed. In the future, researchers may

work on dataset-specific pre-processing tasks. As the number of studies working with large volume COVID-19 misinformation data is relatively small, it was noticed that there are no efficient techniques for the selection of the important features on large-scale data. Future research can focus on proposing methods how to extract the most significant features from large volume data by minimizing the feature vector size effectively.

5.4 Employing sentiment or emotion analysis

Existing studies on COVID-19 misinformation detection entirely focus on detecting the authenticity of the tweets or news articles but ignore the sentiments or emotions associated with them. The sentiment analysis from the texts can play a significant role in the detection of misinformation (Bhutani et al. 2019). It may be an interesting investigation for future researchers to extract the sentiments of misinformative facts related to COVID-19 and utilize them to build up a robust classification system. Besides, different emotions associated with the texts can also be a great consideration for making an emotion-based classification system (Ghanem et al. 2020).

5.5 Multi-modality-based detection system

There is still a lack of study on COVID-19 misinformation detection that used multimodalities such as texts, images, and videos altogether. Although individual modality is very important, it is not sufficient alone. Different modalities can help to gain different aspects of content and derived information from different modalities complement each other to detect misinformation (Singhal et al. 2019; Song et al. 2021). The similarity between the image and the text is very important which can be a piece of additional information for a comprehensive outcome. Thus, a study can be done by incorporating multimodal features to make a robust misinformation detection system. Though these multimodal systems can perform well in detecting misinformation, they can increase training and model size overhead, training cost, and complexity as the classifiers have always been trained with another classifier. In today's competitive age, it is worthwhile to research those open issues and researchers can make contributions to solve these problems.

5.6 Cross-domain misinformation studies

Cross-domain studies on misinformation such as analyzing the different kinds of sources of the information, topics can assist the current models to acquire better results. Current studies emphasize mostly distinguishing misinformation from real information using the content of the information. Content information is important to realize semantic information. But, it is difficult to detect newly emerged misinformation using content information only (Sitaula et al. 2020; Shu et al. 2019b). So, analyzing false news across sources of information, topics, and URLs allow one to obtain a better understanding of the information and also helps to identify its unique characteristics, which can further assist to detect misinformation early.

5.7 Unsupervised learning-based techniques

All the existing works on COVID-19 misinformation detection are supervised which requires an extensive amount of time and a pre-annotated misinformation dataset to train a model. Obtaining a benchmark misinformation dataset on COVID-19 is also time-consuming and labor-intensive work as the process needs careful checking of the contents. It is also required to check other additional proof such as authoritative reports, fact-checking websites, and news reports. Leveraging a crowdsourcing approach to obtain annotations could relieve the burden of expert checking, but the annotation quality may suffer (Kim et al. 2018). As misinformation is intentionally spread to mislead people, individual human workers alone may not have the domain expertise to differentiate between real information and misinformation (Charles F. Bond and DePaulo 2006). So it would be interesting to consider semi-supervised or unsupervised models having limited or unlabeled data. Besides, unsupervised models can be more practical in the real-life situation because it is easy to get unlabeled data.

5.8 Ensemble and hybrid learning-based techniques

Different kinds of ensemble and hybrid learning techniques can help to build more complex and effective models for extracting better features. It uses several weak classifiers to make one strong classifier that can do more accurate predictions. In the case of the misinformation detection system, different variants of ensemble methods can significantly boost the overall performance of the system (Al-Rakhami and Al-Amri 2020). Again, hybrid classifiers (i.e., traditional ML with traditional ML and DL with DL) have been used for improving the predictions of the classification task in some existing literature, e.g., Elhadad et al. (2021); Al-Rakhami and Al-Amri (2020); Kumar et al. (2021); Hossain et al. (2020). Other combinations (traditional ML with DL) of the hybrid classifier can be used for building up a robust classification system of COVID-19 misinformation.

5.9 Addressing the overfitting problem

ML algorithms face the overfitting problem when these models learn the noise and inaccurate information in the

data. These types of characteristics impact the execution of the model in real-life situations and produce biased results. A perfect combination of the dropout layer with other layers, and the use of different kinds of regularization methods (e.g., weight decay) can reduce this problem. However, these processes need much investigation by the 'Trial and Error' method. So, researchers can work to solve this problem in this area.

5.10 Reinforcement learning for misinformation studies

Reinforcement Learning (RL) is a type of ML technique where an agent learns to achieve a goal in an interactive and uncertain environment. The computer employs a 'Trial and Error' method and the agent gets feedback from its actions and experiences. The studies considered in this survey either use traditional ML or DL algorithms to detect and classify misinformation. However, the training of ML models requires labeled data and DL models also need a large amount of labeled data. Furthermore, manual annotation is time-consuming and expensive. Moreover, annotated data may be outdated due to the dynamic nature of the news article or information. So, it is a major challenge to get new high-quality labeled data to train those models. Thus, RL can be a good option to detect misinformative facts.

6 Conclusion

In this article, we presented a systematic survey outlining the existing research on COVID-19 misinformation classification and detection. The primary goal of this survey is to represent the current state of the relevant research domain to convey up-to-date knowledge to the researchers. In this study, we have discussed different types of misinformation related to COVID-19 and surveyed the existing techniques to detect COVID-19 misinformation by focusing on the pre-processing & feature extraction methods and the classification performance. From our survey, it is conspicuous that DL-based classification methods are more efficient in classifying COVID-19 misinformation compared to the traditional ML methods. Traditional ML algorithms also performed well in the misinformation classification tasks, despite occasional degradations in performance. After analyzing the existing studies in this area, we discovered major research gaps and open issues such as scarcity of benchmark datasets, unavailability of multilingual and multimodal information, inappropriate selection of feature extraction techniques, inadequate classification model accuracies, and so on, which should be investigated further in the future. We believe that this survey article will provide valuable insights into the development of a robust classification system for detecting COVID-19 misinformation and help researchers throughout the world to come up with new strategies to combat the spread of misinformation about the COVID-19 pandemic.

Appendix A Appendix

A.1 Abbreviation list

ML	Machine learning
DL	Deep learning
NLP	Natural language processing
C-V	Cross-validation
PCA	Principle component analysis
ICA	Independent component analysis
BoW	Bag of words
TF-IDF	Term frequency-inverted document frequency
LIWC	Linguistic inquiry and word count
RST	Rhetorical Structure Theory
SVM	Support vector machine
NB	Naive Bayes
MNB	Multinomial Naive Bayes
BNB	Bernoulli Naive Bayes
kNN	k-nearest neighbors
DT	Decision tree
RF	Random forest
ERF	Ensemble random forest
LR	Logistic regression
GDBT	Gradient boost
BN	Bayes net
MLP	Multi-layer perceptron
NN	Neural network
CNN	Convolutional neural network
RCNN	Recurrent convolutional neural network
RNN	Recurrent neural network
LSTM	Long short-term memory
BiLSTM	Bidirectional LSTM
GRU	Gated recurrent unit
BiGRU	Bidirectional GRU
BERT	Bidirectional Encoder Representations from Transformers
CT-BERT	COVID-Twitter-BERT
RoBERTa	Robustly optimized BERT approach
ALBERT	A Lite BERT
mBERT	Multilingual BERT
VAE	Variational autoencoder
SCHOLAR	Sparse Contextual Hidden and Observed Language Autoencoder
HAN	Hierarchical attention networks
dEFEND	Explainable FakE news detection
SAFE	Similarity-aware FakE news detection

SAME	Sentiment-aware multi-modal embedding
CSI	Capture, score, and integrate
CANTM	Classification aware neural topic model

A.2 Dataset link

[i] https://github.com/mohaddad/COVID-FAKES
[ii] https://github.com/DebanjanaKar/Covid19_FakeNews_Detection
[iii] https://gautamshahi.github.io/FakeCovid/
[iv] https://github.com/sshaar/clef2020-factchecking-task1#data-annotation-process
[v] https://github.com/JuanCarlosCSE/YouTube_misinfo
[vi] https://github.com/ucinlp/covid19-data
[vii] https://zoisboukouvalas.github.io/Code.html
[viii] https://competitions.codalab.org/competitions/26655
[ix] https://sites.google.com/view/counter-covid19-misinformation
[x] https://github.com/apurvamulay/ReCOVery
[xi] https://github.com/TIMAN-group/covid19_misinformation
[xii] https://github.com/Gautamshahi/Misinformation_COVID-19
[xiii] https://data.gesis.org/tweetscov19/
[xiv] https://github.com/cuilimeng/CoAID
[xv] https://github.com/lopezbec/COVID19_Tweets_Dataset
[xvi] https://gitlab.com/bigirqu/ArCOV-19
[xvii] https://doi.org/10.17635/lancaster/researchdata/394
[xviii] https://github.com/firojalam/ COVID-19-tweets-for-check-worthiness
[xix] https://github.com/cyang03/CHECKED
[xx] https://ieee-dataport.org/open-access/coronavirus-COVID-19-tweets-dataset
[xxi] https://crisisnlp.qcri.org/covid19
[xxii] https://zenodo.org/record/5652342#.YZDSPGDP3IU
[xxiii] https://github.com/SarahAlqurashi/COVID-19-Arabic-Tweets-Dataset
[xxiv] https://github.com/echen102/COVID-19-TweetIDs
[xxv] https://github.com/yemen2016/FakeNewsDetection
[xxvi] https://github.com/bigheiniu/X-COVID
[xxvii] https://github.com/byew/rumor_detection
[xxviii] https://github.com/MickeysClubhouse/COVID-19-rumor-dataset
[xxix] https://github.com/williamscott701/Cross-SEAN
[xxx] https://www.kaggle.com/gpreda/covid19-tweets
[xxxi] https://github.com/sociocom/covid19_dataset

Author contributions Authors' contributions—**A.R. Sanaullah** was involved in conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft, visualization. **Anupam Das** helped in conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft, visualization. **Anik Das** contributed to conceptualization, methodology, validation, writing—original draft, writing—review & editing, supervision, project administration. **Muhammad Ashad Kabir** was involved in conceptualization, methodology, writing—original draft, writing—review & editing, validation, supervision, project administration. **Kai Shu** helped in validation, writing—review & editing.

Funding The authors did not receive support from any organization for the submitted work.

Data availability Availability of data and materials—data and materials are available upon request to authors after publication for research purposes only.

Code availability Not applicable.

Declarations

Conflict of interest All the authors declare that they have no conflict of interest. The authors have no relevant financial or non-financial interests to disclose.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

References

Ahinkorah BO, Ameyaw EK, Hagan JE et al (2020) Rising above misinformation or fake news in Africa: another strategy to control COVID-19 spread. Front Commun 5:45. https://doi.org/10.3389/fcomm.2020.00045

Akon S, Bhuiyan A (2020) COVID-19: Rumors and Youth Vulnerabilities in Bangladesh

Al-Rakhami M, Al-Amri A (2020) Lies Kill, facts save: detecting COVID-19 misinformation in Twitter. IEEE Access 8:155961–155970. https://doi.org/10.1109/ACCESS.2020.3019600

Alam F, Dalvi F, Shaar Set al (2021) Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms. In: ICWSM

Alkhalifa R, Yoong T, Kochkina Eet al (2020) QMUL-SDS at CheckThat! 2020: determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions. arXiv:2008.13160

Alqurashi S, Alhindi A, Alanazi E (2020) Large Arabic Twitter Dataset on COVID-19. arXiv:2004.04315

Alsudias L, Rayson P (2020) COVID-19 and Arabic Twitter: how can Arab World Governments and Public Health Organizations Learn from Social Media? In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Association for Computational Linguistics

Antoun W, Baly F, Hajj H (2003) AraBERT, Transformer-based Model for Arabic Language Understanding, p 00104

Ayoub J, Yang X, Zhou F (2021) Combat COVID-19 infodemic using explainable natural language processing models. Inf Process Manag. https://doi.org/10.1016/j.ipm.2021.102569

Bale JM (2007) Political paranoia v. political realism: on distinguishing between bogus conspiracy theories and genuine conspiratorial politics. Patterns prejud 41(1):45–60. https://doi.org/10.1080/00313220601118751

Banda JM, Tekumalla R, Wang G et al (2021) A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. Epidemiologia 2(3):315–324. https://doi.org/10.3390/epidemiologia2030024

Bandyopadhyay S, Dutta S (2020) Analysis of Fake News in Social Medias for Four Months during Lockdown in COVID-19- A Study. Xeno Journal of Biomedical Sciences 1(1), 1–6. https://doi.org/10.20944/preprints202006.0243.v1

Bang Y, Ishii E, Cahyawijaya Set al (2021) Model Generalization on COVID-19 Fake News Detection, vol 1402 CCIS. https://doi.org/10.1007/978-3-030-73696-5_13

Bernard D (2020) A Man Drank a Bottle of Rubbing Alcohol for COVID-19. https://www.medpagetoday.com/infectiousdisease/covid19/86094, Accessed 22 Apr 2020

Bhutani B, Rastogi N, Sehgal Pet al (2019) Fake News Detection Using Sentiment Analysis. In: 2019 12th international conference on contemporary computing, IC3 2019. IEEE, pp 1–5, https://doi.org/10.1109/IC3.2019.8844880

Boukouvalas Z, Mallinson C, Crothers Eet al (2020) Independent Component Analysis for Trustworthy Cyberspace during High Impact Events: An Application to Covid-19. arXiv:2006.01284

Busari S, Adebayo B (2020) Nigeria records chloroquine poisoning after Trump endorses it for coronavirus treatment - CNN. https://edition.cnn.com/2020/03/23/africa/chloroquine-trump-nigeria-intl/index.html, Accessed 03 Aug 2021

Card D, Tan C, Smith NA (2018) Neural models for documents with metadata. ACL 2018–56th Annual Meeting of the Association for Computational Linguistics. Proceedings of the conference (Long Papers) 1:2031–2040. https://doi.org/10.18653/v1/p18-1189

Charles F, Bond J, DePaulo BM (2006) Accuracy of deception judgments. Personal Soc Psychol Rev 10(3):214–234. https://doi.org/10.1207/s15327957pspr1003_2

Chen B, Chen B, Gao Det al (2021) Transformer-Based Language Model Fine-Tuning Methods for COVID-19 Fake News Detection, vol 1402 CCIS. https://doi.org/10.1007/978-3-030-73696-5_9

Chen E, Lerman K, Ferrara E (2020) Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. JMIR Publ Health Surveill. https://doi.org/10.2196/19273

Chen S (2020) Research on Fine-Grained Classification of Rumors in Public Crisis—Take the COVID-19 incident as an example. E3S Web of conferences 179:02,027. https://doi.org/10.1051/e3sconf/202017902027

Cheng M, Nazarian S, Bogdan P (2020) VRoC: Variational Autoencoder-Aided Multi-Task Rumor Classifier Based on Text. In: proceedings of The web conference 2020. Association for Computing Machinery, New York, NY, USA, WWW '20, pp 2892–2898, https://doi.org/10.1145/3366423.3380054

Cheng M, Wang S, Yan X et al (2021) A COVID-19 rumor dataset. Front Psychol. https://doi.org/10.3389/fpsyg.2021.644801

Chou WYS, Oh A, Klein WM (2018) Addressing health-related misinformation on social media. Jama 320(23):2417–2418. https://doi.org/10.1001/jama.2018.16865

Chung J, Gulcehre C, Cho Ket al (2014) Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In: NIPS 2014 Workshop on Deep Learning, December 2014

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/BF00994018

Cui L, Lee D (2020) Coaid: Covid-19 healthcare misinformation dataset. arXiv:2006.00885

Cui L, Wang S, Lee D (2019) Same: sentiment-aware multi-modal embedding for detecting fake news. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM 2019, pp 41–48, https://doi.org/10.1145/3341161.3342894

Devlin J, Chang MW, Lee Ket al (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp 4171–4186, https://doi.org/10.18653/v1/N19-1423

Dharawat A, Lourentzou I, Morales Aet al (2020) Drink bleach or do what now? Covid-HeRA: a dataset for risk-informed health decision making in the presence of COVID19 misinformation. arXiv:2010.08743

Dimitrov D, Baran E, Fafalios Pet al (2020) TweetsCOV19—A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. International conference on information and knowledge management, proceedings (June):2991–2998. https://doi.org/10.1145/3340531.3412765

Douglas KM, Sutton RM, Callan MJ et al (2016) Someone is pulling the strings: hypersensitive agency detection and belief in conspiracy theories. Think Reason 22(1):57–77. https://doi.org/10.1080/13546783.2015.1051586

Elhadad M, Li K, Gebali F (2021) An Ensemble Deep Learning Technique to Detect COVID-19 Misleading Information, vol 1264 AISC. https://doi.org/10.1007/978-3-030-57811-4_16

Elhadad MK, Li KF, Gebali F (2020) Detecting Misleading Information on COVID-19. IEEE Access 8:165201–165215. https://doi.org/10.1109/access.2020.3022867

Fernandez M, Alani H (2018) Online misinformation: challenges and future directions. In: Companion proceedings of the web conference 2018. International world wide web conferences steering committee, Republic and Canton of Geneva, CHE, WWW '18, p 595-602, https://doi.org/10.1145/3184558.3188730

Fetzer JH (2004) Disinformation: the use of false information. Minds Mach 14(2):231–240. https://doi.org/10.1023/B:MIND.0000021683.28604.5b

Fritsche I, Moya M, Bukowski M et al (2017) The great recession and group-based control: converting personal helplessness into social class in-group trust and collective action. J Soc Issues 73(1):117–137. https://doi.org/10.1111/josi.12207

F.R.S. KP (1901) LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2(11):559–572. https://doi.org/10.1080/14786440109462720

Gao Z, Yada S, Wakamiya Set al (2020) NAIST COVID: Multilingual COVID-19 Twitter and Weibo Dataset. arXiv:2004.08145

Ghanem B, Rosso P, Rangel F (2020) An emotional analysis of false information in social media and news articles. ACM Trans Internet Technol 20(2):1–17. https://doi.org/10.1145/3381750

Haouari F, Hasanain M, Suwaileh Ret al (2020) ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection. arXiv:2010.08768

Hérault J, Jutten C (1987). Space or time adaptive signal processing by neural network models. https://doi.org/10.1063/1.36258

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hossain T, Logan IV RL, Ugarte Aet al (2020) COVIDLIES: detecting COVID-19 Misinformation on Social Media. In: proceedings of the 1st workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics, https://doi.org/10.18653/v1/2020.nlpcovid19-2.11

Huang B, Carley KM (2020) Disinformation and Misinformation on Twitter during the Novel Coronavirus Outbreak. arXiv:2006.04278

Ji Y, Eisenstein J (2014) Representation learning for text-level discourse parsing. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, pp 13–24, https://doi.org/10.3115/v1/P14-1002

Joulin A, Grave E, Bojanowski Pet al (2017) Bag of tricks for efficient text classification. In: proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Valencia, Spain, pp 427–431

Kaliyar R, Goswami A, Narang P (2020) MCNNet: Generalizing Fake News Detection with a Multichannel Convolutional Neural Network using a Novel COVID-19 Dataset. In: ACM International Conference Proceeding Series, p 437, https://doi.org/10.1145/3430984.3431064

Kaliyar R, Goswami A, Narang P (2021) A hybrid model for effective fake news detection with a novel COVID-19 dataset. In: ICAART 2021—Proceedings of the 13th international conference on agents and artificial intelligence, pp 1066–1072, https://doi.org/10.5220/0010316010661072

Kar D, Bhardwaj M, Samanta Set al (2020) No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection. arXiv:2010.06906

Karlova N, Fisher K (2013) A social diffusion model of misinformation and disinformation for understanding human information behaviour. Inf Res 18

Kim J, Tabibian B, Oh Aet al (2018) Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. In: Proceedings of the eleventh ACM international conference on web search and data mining. Association for Computing Machinery, New York, NY, USA, WSDM '18, pp 324–332, https://doi.org/10.1145/3159652.3159734

Koirala A (2020) Covid-19 fake news classification using deep learning. Master's thesis, Asian Institute of Technology, Thailand

Kumar S, Pranesh RR, Carley K (2021) A Fine-Grained Analysis of Misinformation in COVID-19 Tweets. https://doi.org/10.21203/rs.3.rs-588650/v1

Lai S, Xu L, Liu K et al (2015) Recurrent convolutional neural networks for text classification. Proc Natl Conf Artif Intell 3:2267–2273. https://doi.org/10.5555/3060832.3061023

Lamsal R (2020) Design and analysis of a large-scale COVID-19 tweets dataset. Appl Intell (October). https://doi.org/10.1007/s10489-020-02029-z

Lazer D, Baum M, Benkler Y et al (2018) The science of fake news: addressing fake news requires a multidisciplinary effort. Science 359(6380):1094–1096. https://doi.org/10.1126/science.aao2998

Li HOY, Bailey A, Huynh D et al (2020) YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ Glob Health. https://doi.org/10.1136/bmjgh-2020-002604

Li Q, Zhang Q, Si Let al (2019) Rumor detection on social media: datasets, methods and opportunities. In: proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda. Association for Computational Linguistics, Hong Kong, China, pp 66–75, https://doi.org/10.18653/v1/D19-5008

Li Y, Jiang B, Shu Ket al (2020b) Toward a multilingual and multimodal data repository for covid-19 disinformation. In: 2020 IEEE international conference on big data (Big Data), IEEE, pp 4325–4330, https://doi.org/10.1109/BigData50022.2020.9378472

Lin X, Liao X, Xu Tet al (2019) Rumor Detection with Hierarchical Recurrent Convolutional Neural Network. In: Tang J, Kan MY, Zhao Det al (eds) Natural Language Processing and Chinese Computing. Springer International Publishing, Cham, pp 338–348, https://doi.org/10.1007/978-3-030-32236-6_30

Liu P, Qiu X, Huang X (2016) Recurrent neural network for text classification with multi-task learning. In: proceedings of the twenty-fifth international joint conference on artificial intelligence. AAAI Press, IJCAI'16, pp 2873–2879

Loomba S, de Figueiredo A, Piatek SJ et al (2021) Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Human Behav 5(3):337–348. https://doi.org/10.1038/s41562-021-01056-1

Lopez C, Gallemore C (2020) An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic. https://doi.org/10.21203/rs.3.rs-95721/v1

Losee RM (1997) A discipline independent definition of information. J Am Soc Inf Sci 48(3):254–269. https://doi.org/10.1002/(SICI)1097-4571(199703)48:3<254::AID-ASI6>3.0.CO;2-W

Madani Y, Erritali M, Bouikhalene B (2021) Using artificial intelligence techniques for detecting covid-19 epidemic fake news in Moroccan tweets. Results Phys 25(104):266. https://doi.org/10.1016/j.rinp.2021.104266

Mahlous A, Al-Laith A (2021) Fake News Detection in Arabic Tweets during the COVID-19 Pandemic. International Journal of Advanced Computer Science and Applications 12(6), 778–788https://doi.org/10.14569/IJACSA.2021.0120691

Mann W, Thompson S (1988) Rhetorical structure theory: toward a functional theory of text organization. Text 8:243–281. https://doi.org/10.1515/text.1.1988.8.3.243

Medina Serrano JC, Papakyriakopoulos O, Hegelich S (2020) NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In: proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020. Association for Computational Linguistics

Mejova Y, Kalimeri K (2020) Advertisers Jump on Coronavirus Bandwagon: Politics, News, and Business. arXiv:2003.00923

Micallef N, He B, Kumar Set al (2020) The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. arXiv:2011.05773

Mikolov T, Sutskever I, Chen Ket al (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems—Volume 2. Curran Associates Inc., Red Hook, NY, USA, NIPS'13, pp 3111–3119

Moher D, Liberati A, Tetzlaff J et al (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLOS Med 6(7):1–6. https://doi.org/10.1371/journal.pmed.1000097

Müller M, Salathé M, Kummervold PE (2020) COVID-Twitter-BERT: a natural language processing model to analyse COVID-19 Content on Twitter. arXiv:2005.07503

Ng LHX, Carley KM (2021) The coronavirus is a bioweapon: classifying coronavirus stories on fact-checking sites. Comput Math Organ Theory 27(2):179–194. https://doi.org/10.1007/S10588-021-09329-W

Nguyen MH, Gruber J, Fuchs Jet al (2020) Changes in Digital Communication During the COVID-19 Global Pandemic: Implications for Digital Inequality and Future Research. Social Media + Society 6(3). https://doi.org/10.1177/2056305120948255

Paka WS, Bansal R, Kaushik A et al (2021) Cross-sean: a cross-stitch semi-supervised neural attention model for covid-19 fake news detection. Appl Soft Comput 107(107):393. https://doi.org/10.1016/j.asoc.2021.107393

Papadopoulos T, Baltas KN, Balta ME (2020) The use of digital technologies by small and medium enterprises during COVID-19: implications for theory and practice. Int J Inf Manag 55(102):192. https://doi.org/10.1016/j.ijinfomgt.2020.102192

Patwa P, Sharma S, Pykl Set al (2021) Fighting an Infodemic: COVID-19 Fake News Dataset, vol 1402 CCIS. https://doi.org/10.1007/978-3-030-73696-5_3

Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates 71(2001):2001

Pennington J, Socher R, Manning C (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1532–1543, https://doi.org/10.3115/v1/D14-1162

Peters M, Neumann M, Iyyer Met al (2018) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 2227–2237, https://doi.org/10.18653/v1/N18-1202

Petratos PN (2021) Misinformation, disinformation, and fake news: cyber risks to business. Bus Horizons 64(6):763–774. https://doi.org/10.1016/j.bushor.2021.07.012

Preda G (2020) COVID19 Tweets. https://www.kaggle.com/gpreda/covid19-tweets, Accessed: 01 Aug 2021

van Prooijen JW, Douglas KM (2018) Belief in conspiracy theories: basic principles of an emerging research domain. Eur J Soc Psychol 48(7):897–908. https://doi.org/10.1002/ejsp.2530

Qazi U, Imran M, Ofli F (2020) Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. SIGSPATIAL Spec 12(1):6–15. https://doi.org/10.1145/3404820.3404823

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, pp 3980–3990, https://doi.org/10.18653/v1/D19-1410

Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for idf. J Doc 60:503–520. https://doi.org/10.1108/00220410410560582

Ruchansky N, Seo S, Liu Y (2017) CSI: a hybrid deep model for fake news detection. Int Conf Inf Knowl Manag Proc Part F1318:797–806. https://doi.org/10.1145/3132847.3132877

Shahi GK, Nandini D (2020) FakeCovid—A Multilingual Cross-domain Fact Check News Dataset for COVID-19. In: workshop proceedings of the 14th international AAAI conference on web and social media, https://doi.org/10.36190/2020.14

Shahi GK, Dirkson A, Majchrzak TA (2021) An exploratory study of covid-19 misinformation on twitter. Online Soc Netw Media 22(100):104. https://doi.org/10.1016/j.osnem.2020.100104

Shahsavari S, Holur P, Wang T et al (2020) Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. J Comput Soc Sci. https://doi.org/10.1007/s42001-020-00086-5

Shi A, Qu Z, Jia Qet al (2020) Rumor detection of covid-19 pandemic on online social networks. In: 2020 IEEE/ACM symposium on edge computing (SEC), IEEE, pp 376–381, https://doi.org/10.1109/SEC50012.2020.00055

Shimizu K (2020) 2019-ncov, fake news, and racism. Lancet 395(10225):685–686. https://doi.org/10.1016/S0140-6736(20)30357-3

Shu K, Cui L, Wang Set al (2019a) Defend: explainable fake news detection. Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (May):395–405. https://doi.org/10.1145/3292500.3330935

Shu K, Wang S, Liu H (2019b) Beyond news contents: the role of social context for fake news detection. In: proceedings of the twelfth ACM international conference on web search and data mining. Association for Computing Machinery, New York, NY, USA, WSDM '19, pp 312–320, https://doi.org/10.1145/3289600.3290994

Singhal S, Shah RR, Chakraborty Tet al (2019) Spotfake: a multimodal framework for fake news detection. In: 2019 IEEE Fifth international conference on multimedia big data (BigMM), pp 39–47, https://doi.org/10.1109/BigMM.2019.00-44

Sitaula N, Mohan CK, Grygiel Jet al (2020) Credibility-Based Fake News Detection, Springer International Publishing, Cham, pp 163–182. https://doi.org/10.1007/978-3-030-42699-6_9

Song C, Ning N, Zhang Y et al (2021) A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf Process Manag 58(1):102437. https://doi.org/10.1016/j.ipm.2020.102437

Song X, Petrak J, Jiang Yet al (2020) Classification Aware Neural Topic Model and its Application on a New COVID-19 Disinformation Corpus. arXiv:2006.03354

Spring M (2021) Social media firms fail to act on Covid-19 fake news. https://www.bbc.com/news/technology-52903680, Accessed: 14 Nov 2021

Su Q, Wan M, Liu X et al (2020) Motivations, methods and metrics of misinformation detection: an nlp perspective. Nat Lang Process Res 1:1–13. https://doi.org/10.2991/nlpr.d.200522.001

Su Z, McDonnell D, Wen J et al (2021) Mental health consequences of COVID-19 media coverage: the need for effective crisis communication practices. Global Health 17(1):1–8. https://doi.org/10.1186/s12992-020-00654-4

Swami V, Coles R, Stieger S et al (2011) Conspiracist ideation in britain and austria: evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories. Br J Psychol 102(3):443–463. https://doi.org/10.1111/j.2044-8295.2010.02004.x

Tasnim S, Hossain MM, Mazumder H (2020) Impact of rumors and misinformation on covid-19 in social media. J Prevent Med Public Health 53(3):171–174. https://doi.org/10.3961/JPMPH.20.094

Tausczik Y, Pennebaker J (2010) The psychological meaning of words: Liwc and computerized text analysis methods. J Lang Soc Psychol 29:24–54. https://doi.org/10.1177/0261927X09351676

tutor2u (2020) Misleading Information. https://www.tutor2u.net/psychology/topics/misleading-information, Accessed: 27 Feb 2021

UNICEF (2021) COVID-19: frequently asked questions. https://www.unicef.org/stories/novel-coronavirus-outbreak-frequently-asked-questions, Accessed: 18 July 2021

Van Prooijen JW, Douglas KM (2017) Conspiracy theories as part of history: the role of societal crisis situations. Memory Stud 10(3):323–333. https://doi.org/10.1177/1750698017701615

Vaswani A, Shazeer N, Parmar Net al (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'17, pp 6000–6010

Wang H, Gan J, Chen Jet al (2021) Automatic detecting for covid-19-related rumors data on internet. In: 2021 9th international conference on communications and broadband networking. Association for Computing Machinery, New York, NY, USA, ICCBN 2021, pp 22–26, https://doi.org/10.1145/3456415.3456420

Wang Y, McKee M, Torbica A et al (2019) Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 240(112):552. https://doi.org/10.1016/j.socscimed.2019.112552

Wani A, Joshi I, Khandve Set al (2021) Evaluating Deep Learning Approaches for Covid19 Fake News Detection, vol 1402 CCIS. https://doi.org/10.1007/978-3-030-73696-5_15

WHO (2020) Novel Coronavirus. Situation Report–205 205(6):1–19

Wu L, Morstatter F, Carley KM et al (2019) Misinformation in social media: definition, manipulation, and detection. SIGKDD Explor Newsl 21(2):80–90. https://doi.org/10.1145/3373464.3373475

Yang C, Zhou X, Zafarani R (2021) CHECKED: Chinese COVID-19 fake news dataset. Soc Netw Anal Mining 11(1):1–8. https://doi.org/10.1007/s13278-021-00766-8

Yang Z, Yang D, Dyer Cet al (2016) Hierarchical attention networks for document classification. In: proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for Computational Linguistics, San Diego, California, pp 1480–1489, https://doi.org/10.18653/v1/N16-1174

Zarocostas J (2020) How to fight an infodemic. Lancet (London, England) 395(10225):676. https://doi.org/10.1016/S0140-6736(20)30461-X

Zhang H, Kuhnle A, Smith JDet al (2018) Fight under uncertainty: restraining misinformation and pushing out the truth. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 266–273, https://doi.org/10.1109/ASONAM.2018.8508402

Zhang T, Kishore V, Wu Fet al (2020) Bertscore: evaluating text generation with bert. In: international conference on learning representations

Zhou L, Burgoon JK, Nunamaker JF et al (2004) Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group Decis Negot 13(1):81–106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6f

Zhou P, Shi W, Tian Jet al (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: proceedings of the 54th annual meeting of the association for computational linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Berlin, Germany, pp 207–212, https://doi.org/10.18653/v1/P16-2034

Zhou X, Mulay A, Ferrara Eet al (2020a) ReCOVery: a multimodal repository for COVID-19 news credibility research, association for computing machinery, New York, NY, USA, pp 3205–3212. https://doi.org/10.1145/3340531.3412880

Zhou X, Wu J, Zafarani R (2020b) SAFE: similarity-aware multi-modal fake news detection. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12085 LNAI(1):354–367. https://doi.org/10.1007/978-3-030-47436-2_27