# A Survey of COVID-19 Misinformation: Datasets, Detection Techniques and Open Issues

A.R. Sana Ullah[a], Anupam Das[a], Anik Das[b], Muhammad Ashad Kabir[c,*], Kai Shu[d]

[a]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chottogram, Bangladesh
[b]Department of Computer Science, St. Francis Xavier University, NS, Canada
[c]School of Computing and Mathematics, Charles Sturt University, NSW, Australia
[d]Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

## 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 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 identify various COVID-19 misinformation datasets, and review different data processing, feature extraction and classification techniques to detect COVID-19 misinformation. At the end, the challenges and limitations in detecting COVID-19 misinformation using machine learning techniques and the future research directions are discussed.

*Keywords:* COVID-19, Misinformation, Fake news, Rumor, Disinformation, Machine learning, Deep learning, Detection, Classification, Survey

## 1. Introduction

Misinformation is a piece of false information or inaccurate information that is intentionally created to get more attention from people [1]. There are many terms related to misinformation such as fake news, rumor, false information, misleading information and disinformation [2]. 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 [3, 4, 5, 6, 7].

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 [8]. People are now getting more

*Corresponding author
*Email addresses:* u1604050@student.cuet.ac.bd (A.R. Sana Ullah), u1604054@student.cuet.ac.bd (Anupam Das), x2021gmg@stfx.ca (Anik Das), akabir@csu.edu.au (Muhammad Ashad Kabir), kshu@iit.edu (Kai Shu)

involved in these platforms, especially on Facebook, Twitter, Instagram, etc., and expressing their thoughts, 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 [9]. They gather information about COVID-19 from any news media or social media platforms 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, affecting people's mental health, daily lives and behaviors [10]. 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 [11]. World Health Organization (WHO) called this situation as '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 [12, 13].

It is now a global concern to combat the spread of COVID-19 misinformation in online platforms. It has already gained a great deal of attention from researchers all around the world. A significant number of research works (e.g., [14, 15, 16]) have applied various machine learning techniques for detecting COVID-19 misinformation in online platforms. As there are still many challenging issues in the existing studies that need further investigations [17], 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 our inclusion criteria. We include papers that aim to detect COVID-19 misinformation using either traditional machine learning or deep learning 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, 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, Section 6 concludes the paper.

## 2. COVID-19 misinformation

### 2.1. Misinformation types

According to Fetzer et al. [18], misinformation is "false, mistaken, or misleading information". Others define misinformation as inaccurate information, which is created to misguide the readers [1, 19] or "any claim of fact that is currently false due to lack of scientific evidence" [20]. 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. Fig. 1 depicts the categorization of COVID-19 misinformation within the scope of this survey.

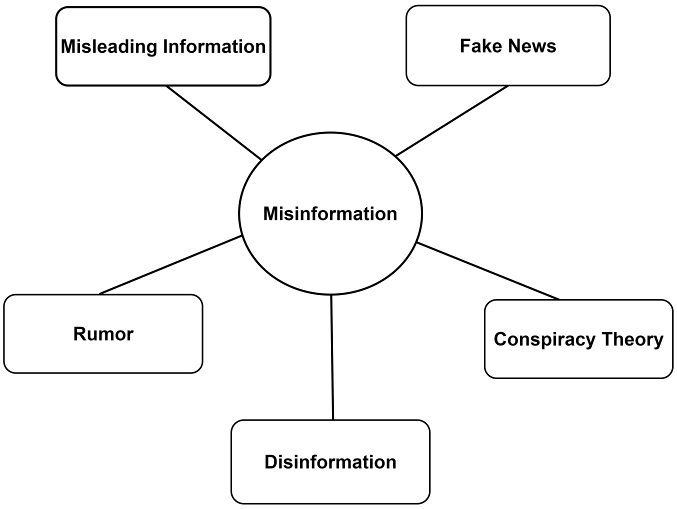

Figure 1: Types of COVID-19 misinformation

**Fake news** is a modified version of an original news which is used to misguide the people or manipulate public opinion using traditional mass media and online social media [21]. It is also known as fabricated information which differs in organizational procedure or purpose but looks similar to news media content [22]. 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 [3]. These kinds of news are unreliable and create misconception 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 [23, 24, 25]. These are created for doing harm to the people by the help of internet access [26, 25]. People believe in conspiracy theories during societal crises, such as natural disasters, financial crises or diseases, wars and terrorist attacks [27, 28]. For example, many conspiracy theories are created during 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" [29].

**Rumor** is basically a story of uncertain or doubtful truth. It spreads online very quickly [30]. Sometimes, it is called "false rumor" or "fake news" when a rumor's veracity value is false [31]. Many kinds of rumors are circulating during this COVID-19 pandemic. For example, among the rumors spread in the beginning of coronavirus infection in Bangladesh are: "Coronavirus would not come in Bangladesh as its temperature is more that 30 degrees", and "Drinking 3 cups of tea daily can get rid of coronavirus" [32].

**Misleading information** is defined as incorrect information which is given to an eyewitness following an event [33]. 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 [34].

**Disinformation** is treated as a part of misinformation [35, 36]. Inaccurate information is referred to as "Misinformation" whereas deceptive information is referred to as "Disinformation" [37]. It creates misconception among the people. One recent disinformation related to COVID-19 is that drinking pure alcohol can kill coronavirus [38], which is really misguiding and injurious to health.

### *2.2. Impact of COVID-19 misinformation*

Since the beginning of COVID-19 pandemic, misinformation has become a major issue worldwide. The main reason behind this is the substantial increase of internet use during this pandemic for different purposes (e.g., communication [39], business [40, 41], health related information [42] 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 [43, 44, 45]. Facebook, a popular social networking site, has reported that approximately 90 million pieces of contents during the March and April of 2020 are related to COVID-19 misinformation [46]. Li et al. [42] 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 [47]. Furthermore, some misinformation might create a serious threat by misleading the general population [43].

## 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) [48] method where we illustrate step by step systematic approach for selecting the articles.

### *3.1. Search scope*

In this survey, we have searched in three prominent databases such as Scopus, Web of Science and Google Scholar. Scopus and Web of Science are the popular authenticated 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 query string/keyword-based searching method in our study. Our query string/keyword includes COVID-19 related misinformation, fake news, rumors, misleading information related studies that have used detection, classification and clustering techniques using ML algorithms. The search keywords and query strings are listed in the Table 1. We have searched the different formatted query strings on these databases between July 18, 2021 and July 24, 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 rumour detection |

### *3.3. Selection criteria*

For the selection of the papers for our systematic review, we have defined five inclusion criteria: (i) 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) article should either employ any traditional machine learning (ML) and/or deep learning (DL) model(s) to classify misinformation, or present a dataset related to COVID-19 misinformation, (iv) article employs classification model(s) must have presented performance evaluation of the adopted model(s), and (v) article must be written in English.

Fig. 2 shows the systematic selection process of the articles using PRISMA [48]. A total of 260 papers were found in the “identification” phase of our study by searching the databases. After removing 38 duplicate articles, 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 which were included and analysed in this survey.

## 4. Analysis

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

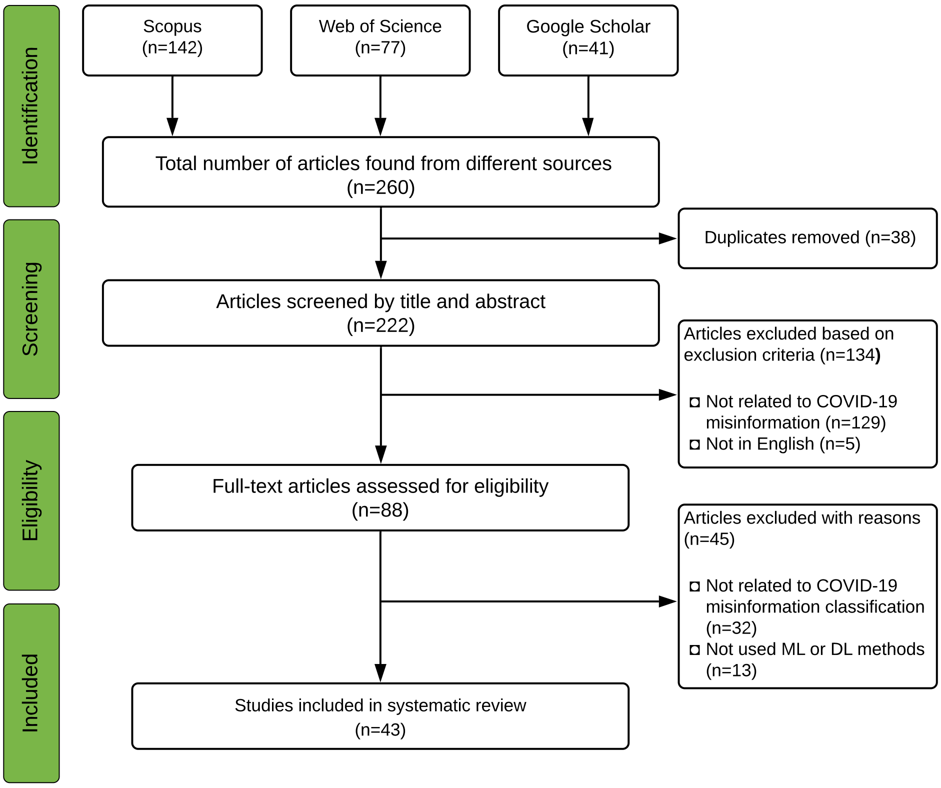
Identification
Screening
Eligibility
Included
Scopus (n=142)
Web of Science (n=77)
Google Scholar (n=41)
Total number of articles found from different sources (n=260)
Duplicates removed (n=38)
Articles screened by title and abstract (n=222)
Articles excluded based on exclusion criteria (n=134)
□ Not related to COVID-19 misinformation (n=129)
□ Not in English (n=5)
Full-text articles assessed for eligibility (n=88)
Articles excluded with reasons (n=45)
□ Not related to COVID-19 misinformation classification (n=32)
□ Not used ML or DL methods (n=13)
Studies included in systematic review (n=43)

Figure 2: Prisma flow diagram for the systematic selection and evaluation of the articles

### *4.1. Dataset description*

Relevant and sufficient training data is considered as the basis to achieve precise results from any ML-based misinformation detection systems. To perform the misinformation classification task, data from various platforms such as social media, news websites, fact-checking sites, 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 the 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.

#### *4.1.1. Data sources*

Studies included in this review paper cover data from multiple sources. The articles [14, 34] 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 [16, 52, 72, 53, 54, 56, 57, 58, 73, 76, 62, 78, 79, 80, 81, 82, 83, 84, 60, 63, 85, 68] used Twitter platform as their data source. Several Twitter APIs such as streaming API, Tweepy API, etc, are generally used to collect the tweets from this platform.

Table 2: Summary of the datasets

| Type | Paper | Dataset Name | Data Source | | | | Dataset link | Language | Instances | Labels |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | SP | FCW | NW | O | | | | |
| Misleading | [34] | COVID-19-FAKES | ✓ | ✓ | - | ✓ | i | En,Ar | 7,486 | 2 |
| Fake News | [16] | Indic-covidemic tweet dataset | ✓ | - | - | - | ii | En,B,H | 1,438 | 2 |
| | [49] | FakeCovid | - | ✓ | - | - | iii | En, H, G and Other 37 | 5,182 | 2 |
| | [50] | Koirala | - | - | ✓ | - | - | En | 3,119 | 2 |
| | [51] | COVID-19 Twitter Fake News (CTF) | ✓ | ✓ | - | ✓ | xxxi | En | 45,261 | 2 |
| | [52] | Madani et al. | ✓ | - | - | - | - | En | 2000 | 2 |
| | [53] | COVID19-Lies | ✓ | - | - | - | vi | En | 6,761 | 3 |
| | [54] | COVID-19 Twitter Data | ✓ | - | - | - | vii | En | 560 | 2 |
| | [55] | MM-COVID | ✓ | ✓ | ✓ | ✓ | xxvi | En, S, P, H, It, F | 11,173 | 2 |
| | [56] | Al-Rakhami et al. | ✓ | - | - | - | - | En | 409,484 | 2 |
| | [57] | Fake news dataset | ✓ | - | - | ✓ | - | En | 19,873 | 2 |
| | [58] | Kumar et al. | ✓ | - | - | - | - | En | 1,970 | 4 |
| | [59] | COVID-19 Fake News Dataset | ✓ | ✓ | - | - | viii | En | 10,700 | 2 |
| | [60] | Counter-covid19-misinformation | ✓ | - | - | - | ix | En | 155,468 | 3 |
| | [61] | ReCOVery | ✓ | - | ✓ | - | x | En | news 2,029<br>tweets 140,820 | 2 |
| | [62] | Covid-HeRA | ✓ | - | - | - | xi | En | 61,286 | 5 |
| | [63] | Misinformation COVID-19 | ✓ | ✓ | - | - | xii | En | 1,500 | 2 |
| | [21] | CoAID | ✓ | ✓ | ✓ | - | xiv | En | 298,778 | 2 |
| | [64] | ArCOV-19 | ✓ | ✓ | - | - | xvi | Ar | 9,414 | 3 |
| | [65] | FN-COV | - | - | ✓ | - | - | En | 69,976 | 2 |
| | [66] | Ayoub et al. | - | ✓ | ✓ | ✓ | - | En | 984 | 2 |
| | [67] | Ng et al. | - | ✓ | - | - | - | En | 6,731 | 5 |
| | [68] | Arabic Fake News corpora | ✓ | - | - | - | xxv | Ar | 36,066 | 2 |
| | [69] | CHECKED | ✓ | - | - | - | xix | C | 2,104 | 2 |
| Rumor | [15] | Chen | - | ✓ | - | - | - | En | 3,737 | 3 |
| | [70] | Wang et al. | - | ✓ | - | - | xxvii | En | 7,179 | 3 |
| | [71] | Shi et al. | ✓ | - | - | - | - | En | 1,537 | 2 |
| | [72] | CLEF dataset | ✓ | - | - | - | iv | En | 962 | 2 |
| | [73] | COVID-19 Arabic tweets | ✓ | - | - | - | xvii | Ar | 2,000 | 3 |
| | [74] | COVID-19-rumor-dataset | ✓ | ✓ | ✓ | - | xxviii | En | 6,834 | 3 |
| Conspiracy | [75] | YouTube_misinfo | ✓ | - | - | - | v | En | videos 180<br>comments 151,567 | 2 |
| Disinformation | [76] | COVID-19 Infodemic Twitter Dataset | ✓ | - | - | - | xviii | En, Ar | 722 | 2 |
| | [77] | COVID-19 Disinformation corpus | - | ✓ | - | - | - | En | 1,293 | 10 |
| Unlabeled | [78] | TweetsCOV19 | ✓ | - | - | - | xiii | En | 8,151,524 | - |
| | [79] | COV19Tweets Dataset | ✓ | - | - | - | xx | En | Over 310 million | - |
| | [51] | CTF | ✓ | ✓ | - | ✓ | xxix | En | 21.85 million | - |
| | [80] | GeoCoV19 | ✓ | - | - | - | xxi | En, S and other 60 | 524,353,432 | - |
| | [81] | COVID-19 Twitter Chatter Dataset | ✓ | - | - | - | xxii | En, F, S, G and others | 800,064,296 | - |
| | [82] | COVID-19-Arabic-Tweets-Dataset | ✓ | - | - | - | xxiii | Ar | 3,934,610 | - |
| | [83] | COVID-19 Twitter dataset | ✓ | - | - | - | xxiv | En, S, I, F, P and other 62 | 123,113,914 | - |
| | [84] | COVID19_Tweets_Dataset | ✓ | - | - | - | xv | En, S, P and other 63 | 785,118,723 | - |
| | [85] | COVID19 Tweets | ✓ | - | - | - | xxx | En | 179,108 | - |
| | [86] | 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.

The article [75] used video data which were collected from Youtube using YouTube's Data API. In the study [61], the dataset includes news articles as well as the 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. Another study [64] 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. Cui and Lee [21] 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 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. [86] introduced a multilingual dataset containing microblogs related to COVID-19 from Twitter and Chinese social media platform Weibo.

On the other hand, article [15] utilized the data fetched from various Chinese rumor refuting platforms such as Sina News, Baidu, 360 rumor refuting platform etc, article [49] used data that is collected from different fact-checking websites by getting references from Poynter and Snopes, article [50] made a dataset by scraping the data from various news and blog sites using Webhose.io API, study [77] used the data gathered from IFCN Poynter website, article [69, 71] used a dataset containing microblogs related to COVID-19 which were crawled from the popular Chinese social media platform Weibo, and the study [59] 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. In the study [74], the authors introduced a COVID-19 rumor dataset that contains rumors regarding COVID-19 from a wide range of sources. The rumors are collected from various news sites (e.g., CNN, BBC News), fact-checking websites (e.g., Poynter, FactCheck ), and Twitter platform. This dataset also includes some metadata of the rumors which are source website, date of publication, reposts or retweets, etc. Wang et al. [70] collected their rumor data from Snopes. Ng et al. [67] collected fact-checked stories regarding coronavirus to make a dataset. The stories are curated from popular fact-checking websites such as Poynter, Snopes, and PolitiFact. In the study [65], a dataset has been used that contains news articles regarding COVID-19 published worldwide. Ayoub et al. [66] 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. [55] 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 Figure 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, Figure 4 represents the number of datasets against their application purposes, such as Fake news, Rumor, Disinformation, Conspiracy theory, and Misleading information.

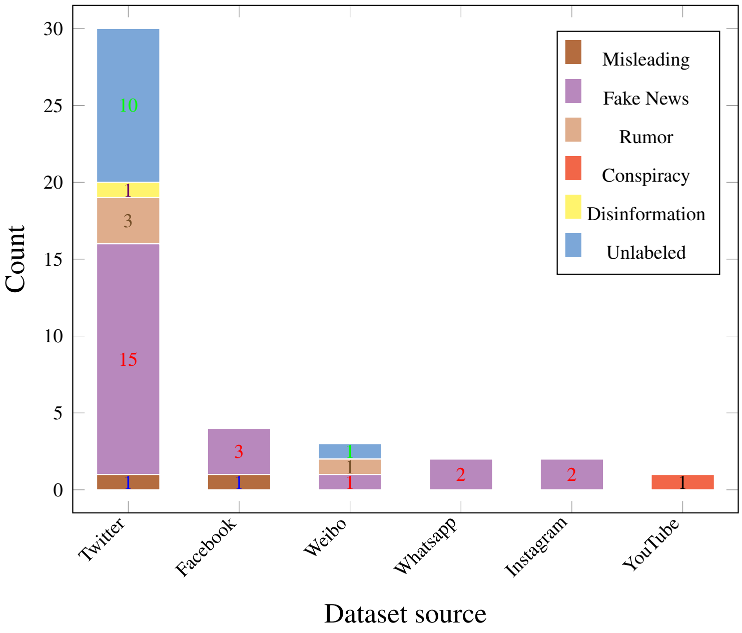

Figure 3: Datasets from various social media platforms

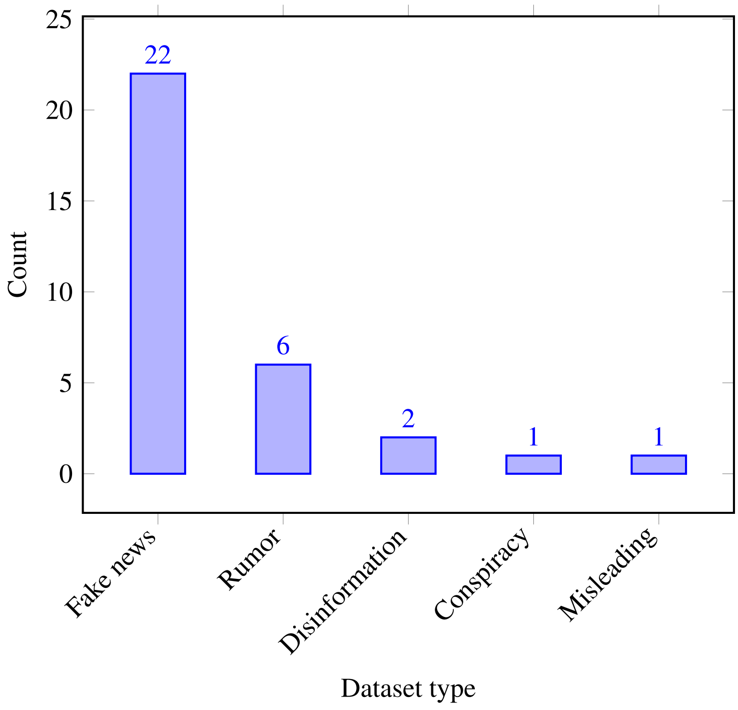

Figure 4: Count of different dataset types

#### 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. In the misleading category, the studies [34, 14] 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 [16, 49, 50, 57, 59, 69, 63, 21, 65, 66, 68, 51, 55, 52] 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 [54, 56, 61] used different names to represent class labels. ‘Unreliable’ or ‘Non-credible’ is used to represent fake news, and ‘Reliable’ or ‘Credible’ is used to represent true news pieces about COVID-19. In one study [53], 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 [64] also used a dataset of three classes labeled as 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 can not be labeled as one of the two earlier cases then it is labeled as Other. Another study [60] 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 can not be categorized in the prior two classes. Ng et al. [67] 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 can not be verified at all. Other two studies [58, 62] under this category used four and five class labels respectively to organize their data.

In the rumor category, the study [71, 72] used a dataset labeled as Rumor and Non-rumor or real. If a tweet needs check-worthiness for the topic, it is labeled as Rumor, otherwise, it is labeled as Non-rumor or real. In the study [15], the author labeled the COVID-19 rumor data into three categories where health-related rumors are labeled as health rumor, scientific rumors are labeled as science rumor and the rumors about the society are labeled as society rumor. Alsudias et al. [73] organized their data using three-class labels which are True, False, and Unrelated. The tweets which represent correct information are labeled as ‘True’, where the tweets containing rumor or false information are labeled as ‘False’ and irrelevant tweets are labeled as ‘Unrelated’. In the study [74], 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, a study [75] 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, the study [77] 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 [76] 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.

Table 3 represents the datasets along with their corresponding class labels as well as the studies that introduced or used them.

Table 3: Datasets and their class labels

| Dataset name | Class labels | Used in |
|---|---|---|
| COVID-19-FAKES | Real, Misleading | [34, 14] |
| Indic-covidemic tweet dataset | Fake, Non-Fake (Real) | [16] |
| FakeCovid | False (Fake), Others | [49] |
| Koirala | Fake, True (Real) | [50] |
| Madani et al. [52] | Fake, Real | [52] |
| CTF | Fake, Genuine (Real) | [51] |
| COVID-19 Twitter Data | Reliable, Unreliable | [54] |
| MM-COVID | Fake, Real | [55] |
| Al-Rakhami et al. [56] | Credible, Non-credible | [56] |
| Fake news dataset | Fake, Real | [57] |
| Kumar et al. [58] | Irrelevant, Conspiracy, True (Real), False (Fake) | [58] |
| COVID-19 Fake News Dataset | Fake, Real | [59, 87, 88, 66, 89] |
| Counter-covid19-misinformation | Misinformation, Counter-misinformation, Irrelevant | [60] |
| ReCOVery | Reliable, Unreliable | [61] |

Table 3: Datasets and their class labels (continued)

| Dataset name | Class labels | Used in |
|---|---|---|
| Covid-HeRA | Real, Refutes/Rebuts, Highly severe, Possibly severe, Not severe | [62] |
| COVID19-Lies | Agree, Disagree, No Stance | [53] |
| Misinformation COVID-19 | False (Fake) and Partially False (Partially fake) | [63] |
| CoAID | Fake, True (Real) | [21, 90] |
| ArCOV-19 | False (Fake), True (Real), Other | [64] |
| FN-COV | Fake, Real | [90, 65] |
| Ayoub et al. [66] | Fake, True (Real) | [66] |
| Xian Ng et al. [67] | True (Real), Partially true (Partially real), Partially false (Partially fake), False (Fake), Unknown | [67] |
| Arabic Fake News corpora | Fake, Not Fake (Real) | [68] |
| CHECKED | Fake, Real | [69] |
| Chen | Health rumor, Science rumor, Society rumor | [15] |
| Wang et al. [70] | Fake (Rumor), Real, Unverified | [70] |
| Shi et al. [71] | Rumor, Real | [71] |
| CLEF dataset | Rumor, Non-rumor | [72] |
| COVID-19 Arabic tweets | True (Real), False (Rumor), Unrelated | [73] |
| COVID-19-rumor-dataset | True (Real), False (Rumor), Unverified | [74] |
| YouTube_misinfo | Conspiracy, Agreement | [75] |
| COVID-19 Infodemic Twitter Dataset | Yes (Not trustworthy), No (Trustworthy) | [76, 88] |
| COVID-19 Disinformation corpus | PubAuthAction, CommSpread, GenMedAdv, PromActs, Consp, VirTrans, VirOrgn, PubRec, Vacc, None | [77] |

### *4.1.3. Dataset language and availability*

Among the labeled datasets, majority of them contain data only in English language, except the studies [14, 34, 64, 68, 69, 73]. The datasets [64, 68, 73] contain the tweets only in Arabic language while the dataset used in these studies [14, 34] contains data in two different languages - English and Arabic. The dataset [69] contains data in Chinese language. Some studies [16, 49, 55, 76] also introduced multilingual datasets containing data in multiple languages. Datasets used in the studies [15, 50, 77, 56, 57, 58, 65, 66, 67, 70, 52] have not been made publicly available. Additionally, we have collected some unlabeled datasets that are vast in size. All of these datasets [78, 79, 80, 81, 82, 83, 84, 85, 55, 86] are publicly available to use. The datasets [80, 81, 83, 84, 55, 86] are multilingual while others [79, 85, 52, 82] are monolingual containing data in English (first three) and Arabic, respectively. The dataset proposed by Paka et al. [51] contains both labeled and unlabeled data in 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. Some pre-processing steps are discussed in the following subsections.

Tokenization, stop-word removal both are the most common methods performed during the data pre-processing step. In the tokenization process, the entire text or paragraphs are split into small units, called token, whereas removal of the stopword is the process of eliminating the words which do not provide much context. These steps are performed effectively in the studies [77, 75, 54]. Patwa et al. [59] represented their data by performing tokenization and removing stop-words with non alphanumeric characters and unnecessary links. Hossain et al. [53] conducted this tokenization process using NLTK Library. Dutta et al. [57] deleted incomplete news and communal news in their pre-processing steps as it had no need for misinformation detection. They also removed data which had no relation with COVID-19 because of their specific research on COVID-19 misinformation analysis.

Kumar et al. [58] used the NLTK Library for text processing and removed stopwords. They also removed unnecessary tweets, usernames etc. from their data and performed lemmatization technique for converting a word to its roots which helps to extract feature in the next step. In the study [73], 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. [34] used some steps such as text parsing, data cleaning and stop-word removal, POS tagging, stemming for data pre-processing. In data cleaning process, they applied regular expression to get the combination of english alphabets and numbers and eliminate others. They also transformed the digit into the text. In another study [14], the authors removed links, symbols, stop-words, html encoding and repeated words and performed POS tagging and stemming.

Alkhalifa et al. [72] presented different kinds of pre-processing techniques such as Segment2Token, Segment2Root, Word2id and padding. Shahi et al. [49] used a python based library named “langdetect” to identify different kinds of languages to assign respective language 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 [62], reserved tokens such URL, mentions, retweets etc. are filtered out from the tweet data. Alam et al. [76] 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 is used in the study [16]. Ng et al. [67] conducted stemming and lemmatization on the words to find their grammatical roots and removed special characters from the textual contents of the stories.

In the study [89], the authors 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. Kaliyar et al. [65] conducted several pre-processing tasks which include removal of HTML tags, special characters and numbers, and conversion of text characters into lowercase and number words into numeric forms. Wani et al. [87] 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 the study [66], 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, stop-words removal, etc. Mahlous et al. [68] 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.

The pre-processing techniques used in COVID-19 misinformation studies are reported in Table 4.

Table 4: Data pre-processing techniques used in existing research

| Techniques | Explanation | Papers |
|---|---|---|
| Tokenization | Splitting the text into smaller units, known as 'Token' | [54, 77, 58, 14, 59, 72, 49, 50, 75, 53, 62, 76, 73, 89, 66] |
| Stop-words Removal | Removing the words which do not provide much context and hold less useful information | [54, 77, 58, 14, 59, 72, 49, 50, 75, 34, 57, 76, 73, 89, 87, 66] |
| Case-folding | Converting the characters of a sentence into lower case | [65, 87] |
| Stemming | Converting a word to its grammatical roots so that they can be presented in one term only | [14, 34, 87, 67, 68] |
| Lemmatization | Transforming a word to its root which is also known as 'lemma' depending on the context | [58, 66, 67, 68] |
| POS tagging | Assigning one of the part-of-speech to a given word | [34, 14] |
| Data Augmentation | Increasing the data by modifying existing data | [16, 66] |
| Others | Removing HTML tags, URLs and other special characters from texts | [65, 87, 67, 68, 89] |

[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 the important information. In the text categorization, a document generally consists of a large number of words, phrases which creates a high computational burden in the learning process. Also, it is difficult to learn from high dimensional data. Besides, classifier's accuracy can decrease for taking irrelevant features. By 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 described in this subsection.

Table 5: Feature extraction methods used in the literature

| Methods | Papers |
|---|---|
| BERT | |
| Pre-trained BERT | [16, 72, 53, 77, 62, 67, 70, 74] |
| mBERT | [16] |
| COVID-Twitter-BERT | [72, 53, 62] |
| RoBERTa | [89] |
| GloVe | [21, 14, 50, 53, 58, 62, 87, 70] |
| ELMo | [72] |
| Word2Vec | [73, 70] |
| FastText | [73, 87, 70] |
| BoW | [21, 75, 67] |
| Count Vector | [73, 50, 68] |
| TF | [34] |
| TF-IDF | [34, 72, 50, 75, 53, 59, 73, 62, 66, 68, 88] |
| PCA | [54] |
| ICA | [54] |
| LIWC | [75, 61] |
| RST | [61] |
| VAE | [74] |

**PCA** [91] is a method which 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. [54], 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 N-dimensional sub-space and reduced the dimension.

**ICA** [92] 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 mutual orthogonality of the components and the issue of the variance among the datapoints. In the study [54], 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 as a feature in this representation. This method has been adopted for vector representation of texts in several studies [21, 75]. Ng et al. [67] 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** [93] is a feature extraction method which 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 the study [34], 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 [75] used this standard TF-IDF features to pre-process the comments for getting better performance from their classification model. Ayoub et al. [66] applied TF-IDF method to represent the texts into a vector space and extract relevant features. The authors used these features as input to the machine learning classification models. In the study [68], 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. [53] 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 another study [72], 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 [50, 73, 59, 62] 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 [50, 68], the authors used this technique to represent the textual contents into a vector space containing the counts of the terms present in the texts. In the study [73], the authors 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** [94] stands for Linguistic Inquiry and Word Count. It is a psycholinguistic lexicon and can count the words in the article. It count the words based on the one or more of 93 linguistic, psychological and topical categories. 93 features are derived here and classified within a traditional statistical learning framework. [95]. Zhou et al. [61] used this lexicon for their study to extract valuable features from the news articles. In the similiar way, Serrano et al. [75] used LR model usng LIWC's lexicon-derived frequencies as features at the time of training of one of the simpler baseline models.

**RST** [96] stands for Rhetorical Structure Theory which points out the relationship between the parts of text and represents them as a content of a tree. In the study [61], the authors used a pre-trained RST parser [97] 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 them in a traditional statistical learning framework.

**Word2vec** [98] is a word embedding technique developed by Google based on shallow neural network. There are two types of Word2Vec. One is Skip-gram and another is Continuous Bag of Words (CBOW). In CBOW method, it 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 small amount of data and is found to represent rare words well. Wang et al. [70] used GoogleNews-vectors-negative300 [3] which is 100 billion words trained on the Google News dataset. In the study [73], 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** [99] is an extension of word2vec model developed by facebook AI research lab. As a technique of extracting n-gram feature, the generated vector for a word includes the sum of this character 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 the study [73], 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. [70] chose crawl-300d-2M4 embedding [4], a 2 million word vectors which is trained with subword information on Common Crawl (600B tokens). Wani et al. [87] 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) [100], a technique for NLP pre-training is developed by Google. It is deeply bidirectional because it learns text representation for both directions for 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

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

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

model (mBERT) which covers 102 languages and it is trained on Wikipedia. In the study [16], 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. [62] used pre-trained BERT embeddings [100] to determine how close each tweet by using the centroid of its respective category based on their document vectors. The studies [72, 77, 70, 88] also used pre-trained BERT embedding and different pretrained BERT-based models to transform their features into a word-level vector representation. Hossain et al. [53] executed a contextualized word embedding using the pre-trained BERT model to find the semantic similarity between the tweets and misconceptions. In another study [67], the authors used the pre-trained BERT to generate contextualized word embeddings of the stories related to COVID-19. Cheng et al. [74] used BERT to convert the rumor texts into a contextualized vector form. They also used an LSTM-based variational autoencoder (VAE) [101] 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 [72, 53, 62], the authors used COVID-Twitter-BERT [102] 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** [103] stands for global vectors for word representation developed by Stanford as an open source project. It is an unsupervised learning algorithm which 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 [14, 58], the authors applied an embedding layer of dimension 300 using 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. [62] used 100-dimensional pre-trained GloVe embeddings with different classifiers as text representation. Other studies [21, 87] also employed the same dimensional pre-trained GloVe embeddings for their feature extraction process. The study [50] also used 300-dimensional GloVe vectors for their word embedding purpose. The author applied GloVe to create an embedding matrix of words with the indices of tokenized words. Wang et al. [70] chose the GloVe.840b.300d.3 [5] which is trained on Common Crawl consisting of 2 million words. Hossain et al. [53] 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.

**ELMo** [104] stands for Embeddings from Language Models developed in 2018 by AllenNLP. It is a deep contextualized word representation which does not use fixed embedding for each word but for creating word representations, it employs a deep, bi-directional LSTM model. Unlike other traditional word embeddings such as word2vec and

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

GloVe, it analyses 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 the study [72], 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 mostly used classification strategy for classifying COVID-19 misinformation compared to multi-class classification.

Table 6: Classification strategies used in the literature

| Strategy | Papers |
|---|---|
| Binary class | [21, 14, 16, 49, 72, 50, 75, 54, 34, 56, 59, 57, 76, 61, 71, 51, 88, 90, 69, 68, 65, 66, 89, 87] |
| Multi-class | [15, 53, 73, 77, 58, 76, 62, 70, 74, 67] |

#### 4.4.1. *Traditional machine learning 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 of COVID-19 misinformation. The studies that have used traditional ML methods are shown in Table 7.

Based on the ICA model [92], Boukouvalas et al. [54] 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 [105] by using different kinds of popular kernel methods, e.g., Gaussian, RBF, Polynomial. Using SVM model with Gaussian kernel method, an accuracy of 81.2% was reported in their study. Bang et al. [88] used SVM [105] model for setting the baseline of their experiment which is trained on TF-IDF feature and cross-entropy (CE) as loss function. This achieved an accuracy and F1-score of 93.32% both.

Elhadad et al. [34] 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. In this study, the highest accuracy using kNN is 99.36% which was achieved by selecting character level features using 5-fold cross validation with significant precision and recall scores (99.47% and 99.73%) respectively. Kar et al. [16] 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

Table 7: Traditional machine learning methods used in the literature

| Methods | Papers |
|---|---|
| SVM | [21, 16, 75, 54, 34, 56, 59, 73, 62, 61, 67, 68, 88] |
| LR | [21, 75, 34, 59, 73, 62, 61, 66, 67, 68] |
| RF | [21, 16, 75, 56, 62, 61, 66, 68, 34] |
| DT | [34, 59, 61, 66] |
| NB | [56, 73, 61, 67, 68] |
| MNB | [75, 34] |
| BNB | [34] |
| kNN | [34, 56, 57, 61] |
| XGBoost | [34, 68, 71] |
| GDBT | [59] |
| C4.5 | [56] |
| Perceptron | [34] |
| BN | [56] |
| Linear Classifier | [53] |

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.

Serrano et al. [75] 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 LR model based on LIWC's lexicon-derived frequencies [106] as features and another is 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 SVM model is trained using different kernel methods such as linear, sigmoid, and RBF kernel. Among all the six features, comments with conspiracy feature got slightly better accuracy. Alsudias and Rayson [73] 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.

Dutta et al. [57] used kNN based classifier method 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 in social medias. They got a decent accuracy using this classifier. In another study [56], 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 10 fold cross validation.

Cui et al. [21] used different classification methods on their own created dataset as baselines for the comparative analysis of 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 concatenate 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 another study [61], extensive experiments are conducted using ReCOVery dataset which included the baseline performances using either single-modal or multi- modal information of news articles for predicting news credibility and allow future methods to be compared to. 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. [62] performed experiments with several multiclass classification models on their own created new benchmark dataset- "Covid-HeRA". They used RF, SVM, LR model with BOW and 100-dimensional pre-trained GloVe embeddings and achieved a very good accuracy above 95%. In one study [59], the authors performed 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 TF-IDF feature. Hossain et al. [53] 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 in the texts related to COVID-19 and also these texts are linguistically different from the tweets.

Ayoub et al. [66] experimented with three machine learning algorithms (e.g., LR, RF and DT) using TD-IDF features. The 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. Ng et al. [67] 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 et al. [68] employed some machine learning 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, XGB, etc. 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.

Shi et al. [71] introduced a model using XGBoost ensemble learning algorithm where 16 basic features of four types such as text characteristic, user-related, interaction-based and emotion-based feature 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 Figure 5. In this figure, the bubbles contain the number of articles that employed the classification method (expressed in X-axis) and the feature extraction method (expressed in Y-axis).

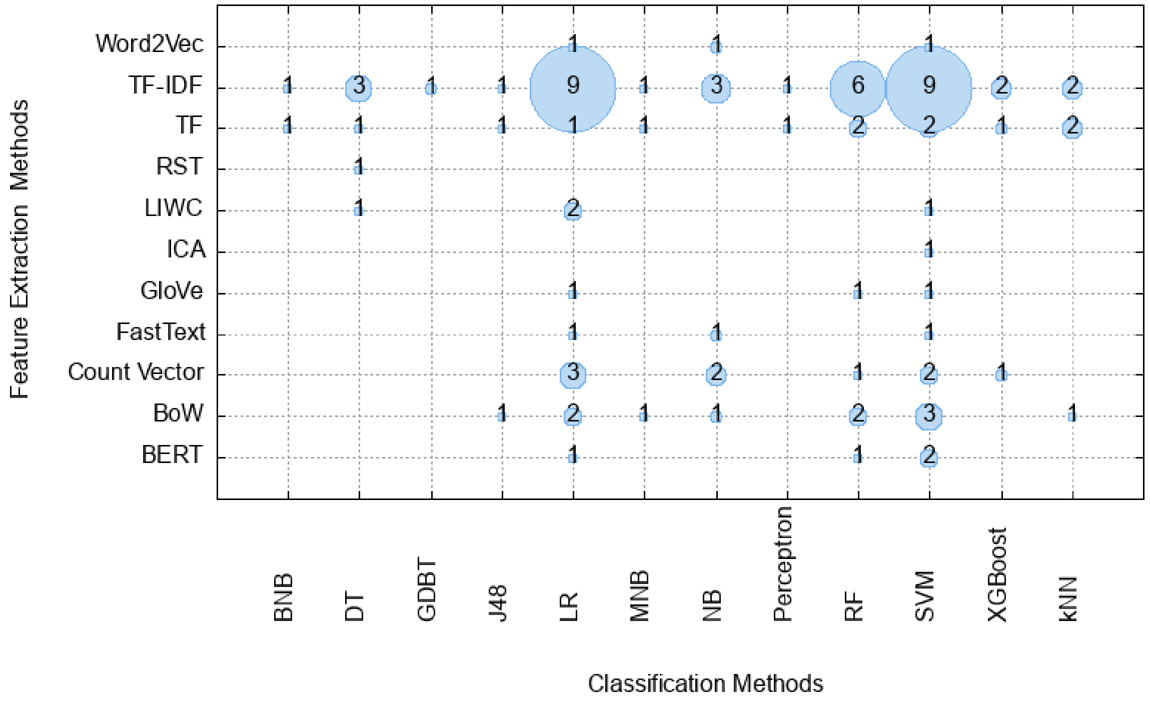

Figure 5: Relationships between feature extraction and traditional machine learning techniques

### *4.4.2. Deep learning 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 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.

Table 8: Deep learning methods used in the literature

| Methods | Papers |
|---|---|
| NN | [34, 16] |
| DNN | [74] |
| MLP | [16, 68] |
| Transformer | [69] |
| BERT-base | [15, 49, 50, 75, 77, 54, 58, 76, 62, 89, 87, 66, 88, 70] |
| BERT-large | [58, 89, 88] |
| Distil-BERT | [58, 87, 66] |
| mBERT | [76] |
| AraBERT | [76] |
| RoBERTa-base | [75, 58, 76, 89, 88] |
| RoBERTa-large | [58, 89, 88] |
| Distil-RoBERTa | [58] |
| ALBERT-base | [76, 58, 89, 88] |
| ALBERT-large | [58, 89] |
| ALBERT-xlarge | [89] |
| CT-BERT | [89, 87] |
| Covid-bert-base | [87] |
| Ro-CT-BERT | [89] |
| XLNet | [75] |
| CNN | [21, 50, 14, 72, 62, 87, 70] |
| RCNN | [14] |
| MCNNet | [90] |
| TextCNN | [15, 58, 61, 69] |
| TextRNN | [15, 69] |
| Att-TextRNN | [69] |
| LSTM | [50, 54, 58, 87, 70] |
| BiLSTM | [53, 54, 58, 62, 70] |
| BiLSTM-Attention | [87] |
| BiGRU | [21] |

Table 8: Deep learning methods used in the literature (continued)

| Methods | Papers |
|---|---|
| Sequential Model | [14] |
| SBERT | [53] |
| SBERT (DA) | [53] |
| XLM-r | [76] |
| FastText | [76, 69] |
| SCHOLAR | [77] |
| SAFE | [61] |
| SAME | [21] |
| HAN | [21, 62, 87] |
| Cross-SEAN | [51] |
| dEFEND | [21, 62] |
| CSI | [21] |
| CANTM | [77] |

**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. [34] implemented a NN model with different feature extraction techniques such as TF, TF-IDF, 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 NN classification model. In another study [16], 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 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 et al. [68] 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 [14] used a sequential model with a GloVe embedding vector to detect misleading information related to COVID-19. Cheng et al. [74] proposed a system for COVID-19 rumor veracity classification based on deep neural networks (DNN). The authors used an LSTM-based variational autoencoder (VAE) [101] 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 it's other variants for the classification purposes. For example, Cui and Lee [21] 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 another study [14], 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 and were able to obtain high accuracy in results. Wang et al. [70] 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. [72] 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) [102] embedding which is pre-trained on COVID-19 Twitter data. Another study [50] 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 class 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. [87] 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. [90] 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 channels features into a single vector and uses some dropout layers to provide generalization cabability 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 FN-COV dataset. This study reports an accuracy of 98.2% and an F1-score of 98.1% with MCNNet from this dataset. 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. [62] 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 [14] 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 [15], fake news [61, 69] and misinformation [58] 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 [15] obtained the highest performance with an accuracy and F1 score of 98.40% and 97.24% respectively, among the studies that adopted the TextCNN model. Elhadad et al. [14] used a 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 [107]. In this study, RCNN performed very well with an accuracy of 99.997%.

**RNN** has the ability to capture better contextual information from the texts. Therefore, various studies utilized RNN and it's other variants for the classification of COVID-19 misinformation. In particular, the studies [15, 69] used the TextRNN [108] model to classify COVID-19 rumors and fake news respectively. TextRNN model uses different LSTM layers inside its architecture. In the study [15], 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 [109] 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 [50, 54, 58, 87]. Among these studies, the study [87] achieved the best performance from LSTM network with an accuracy of 94.95%. Wang et al. [70] also employed 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 both in the forward and backward directions. Hossain et al. [53] used the BiLSTM model to classify tweet-misconception pairs related to COVID-19. Other studies implemented BiLSTM for the classification of COVID-19 misinformation [54, 58, 62]. BiLSTM using pre-trained GloVe embedding came up with an accuracy of 96.6% which is the highest among the studies that employed BiLSTM for the classification purpose [62]. Other two studies [87, 69] employed an attention-based BiLSTM [110] model in the fake news classification task. This architecture includes a BiLSTM layer followed by an attention layer. In the study [87], 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, one study [21] 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 [111] 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 exiting studies focused on BERT and its variants for classification purposes. For instance, Chen [15] 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 purpose. This study demonstrates that the multiheaded attention mechanism used in BERT is capable to produce outstanding results. This study reported an accuracy of 99.20% in the classification results using the BERT model. Alam et al. [76] 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 mBERT model. Shahi and Nandini [49] performed a BERT-based classification of real or fake news on COVID-19 by introducing a multilingual cross-domain dataset. Kumar et al. [58] conducted a fine-grained classification of misinformation in COVID-19 tweets. The authors also applied several transformer language models including three variants of BERT model (e.g., Distil-Bert, BERT-base and BERT-large), three variants of RoBERTa model (e.g., Distil-RoBERTa, RoBERTa-base, RoBERTa-large), and two variants of ALBERT model (e.g., Albert-base, 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. [88] 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 FakeNews-19 train set. By 99% data cleansing percentage, their best model achieved 61.10% accuracy score and 54.33% weighted F1-score on Tweets-19. In another study [75], 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. [89] used different variants of the pre-trained transformer language models (e.g., BERT, RoBERTa, 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. [87] 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. [66] 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

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

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. [53] employed the Sentence-BERT (SBERT) [112] 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 purpose. The study [76] 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 model for English language experiments. BERT outperformed all other models in case of English language. For Arabic language experiments, they employed AraBERT [113] model which is pre-trained on a large corpus of 70 million Arabic sentences. Due to the smaller size of 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 [50], COVID-19 disinformation [77], COVID-19 rumor [70] and COVID-19 misinformation [54, 62].

**Other methods.** Apart from the above methods, researchers also applied other DL methods for the classification purposes. For example, Song et al. [77] proposed a classification aware neural topic model called CANTM for topic modeling task 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 [114] model 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 [21, 62] employed attention-based models for the classification of COVID-19 misinformation. The authors used two models based on attention mechanism namely HAN [115] and dEFEND [116] for their purposes. HAN learns the hierarchical structure of the documents by using two levels of attention mechanisms applied at the word and sentence level. 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 the studies [21, 62], dEFEND showed higher performance scores than HAN for its robustness. Wani et al. [87] 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 one study [61], 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 [117] model 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 [21] employed a model called SAME [118] for the classification of healthcare misinformation regarding COVID-19. SAME is a multi-modal system which uses news image, 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 does 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 [119] 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. [51] introduced Cross-SEAN which is a cross-stitch based semi-supervised neural attention model. This model helps to reduce the dependency on the labelled data as it leverages unlabelled 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 are close, they performed optimal sharing of information by concatenating one output of cross-stitch early in the network with the other afterwards. They employed different types of 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 95% F1-score on their CTF dataset and outperformed the best baseline by 9%.

Some other methods such as XLM-r, FastText were used to perform fine-grained disinformation analysis on Arabic tweets [76]. 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 [69], the authors used the FastText and Transformer [120] model 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 transformer model outperforming the other.

Figure 6 represents the relationship between the feature extraction and deep learning methods (including combined DL methods) used in existing studies. In this figure, the bubbles contain the number of articles that employed the classification method (expressed in X-axis) and the feature extraction method (expressed in Y-axis).

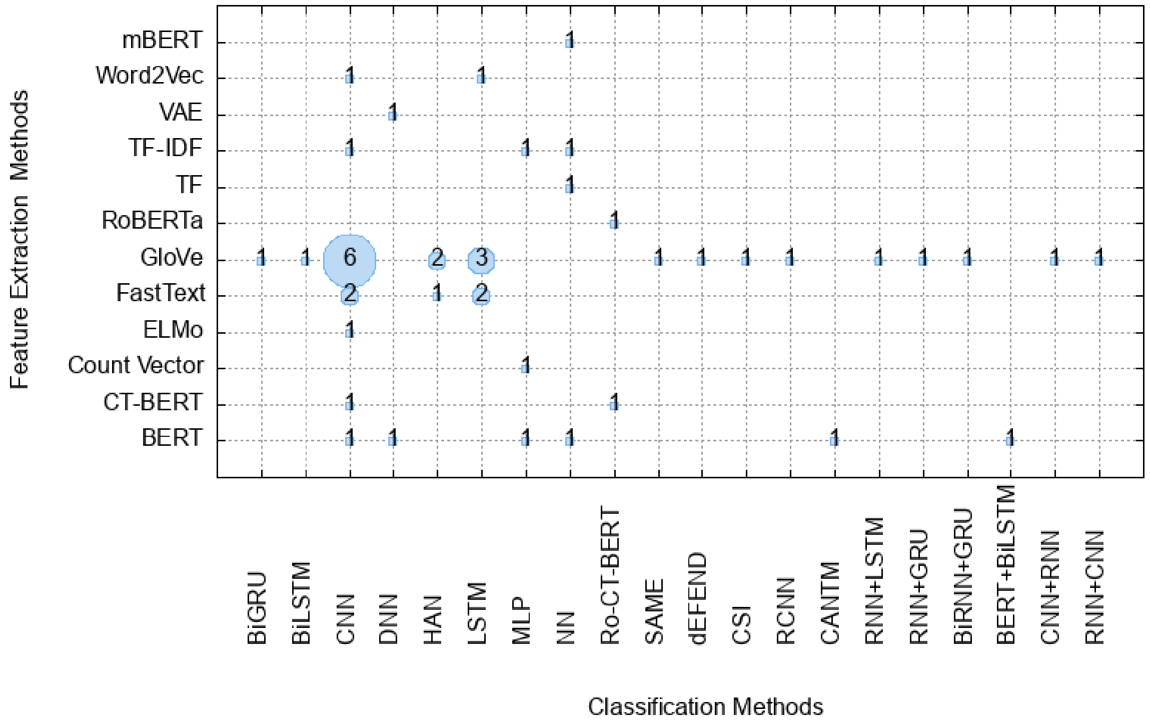

Figure 6: Relationships between feature extraction and deep learning techniques

### *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 [56], 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 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 combination 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. [58] proposed a CNN-RNN model (CNN layer stacked over the RNN layer) and a RNN-CNN model (a single BiLSTM layer is employed over the top of a 1D-CNN layer) with an word embedding in the first layer for the classification of misinformation in COVID-19 tweets. Kaliyar et al. [65] 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

Table 9: Combined Models Used in Literature

| Methods | Papers |
|---|---|
| **ML+ML** | |
| C4.5 + RF | [56] |
| C4.5 + kNN | [56] |
| SVM + RF | [56] |
| SVM + kNN | [56] |
| SVM + Bayes Net + kNN | [56] |
| C4.5 + Bayes Net + kNN | [56] |
| C4.5 + SVM + RF+ BayesNet + kNN + Naive Bayes | [56] |
| **DL+DL** | |
| RNN+LSTM | [14] |
| RNN+GRU | [14] |
| BiRNN+GRU | [14] |
| BERTSCORE (DA)+BiLSTM | [53] |
| BERTSCORE (DA)+SBERT (DA) | [53] |
| BERT+BiLSTM | [70] |
| CNN+RNN | [58] |
| RNN+CNN | [58] |
| C-LSTM | [65] |

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 [14] represents three DL models (e.g., RNN-LSTM, RNN-GRU, 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 cases. Wang et al. [70] applied 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. [53] 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 [121] with domain-adaptive pretraining 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, F1 Score etc. Many of these metrics are known by multiple names. Confusion Matrix, a tabular representation of a classification model, are 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 Equation 1.

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

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

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

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

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

**F1-score.** It is the weighted average of Precision and Recall. It is formally defined in Equation 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 [14] model obtained 99.99% accuracy and 99.97% F1-score which are the highest among all the classification methods used in existing studies. This model was trained and evaluated on the COVID-19-FAKES [34] dataset using an 80:20 train-test split. The NN [34] model achieved an accuracy of 99.68 % with the TF feature by using 5-fold 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 [66] scored an accuracy of 99.4% using 10-fold 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 [89] model outperformed all other methods used in this study with an accuracy of 99.02% from COVID-19 Fake News Dataset [59]. The C-LSTM [65] 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 [90] uses a multi-channel CNN architecture for the generalization purpose which showed an accuracy of 98.2% in the same dataset. In the study [62], the dEFEND model that uses 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 [56] achieved a 97.8% accuracy score by applying 10-fold cross-validation on the dataset. Other methods such as Cross-SEAN [51], C4.5 [56], SVM [59], XGBoost [71], kNN [57], SparseICA-EBM [54], CANTM [77] 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 [69], 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 [68] model 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 [75] came up with an F1-score of 90.30%. The mBERT_NN [16] 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 [74], BERT [76], SAFE [61], BERT+BiLSTM [70], BERTSCORE (DA) + SBERT (DA) (MultiNLI) [53] 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 the 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 of 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

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

| Problem tackled | Reference | Best Model | Split ratio (%) | Split count | | | Performance Metrics | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Train, Validation, Test | Train | Validation | Test | A (%) | P (%) | R (%) | F1 (%) |
| Misleading | [14] | CNN | 80,20,- | 5989 | 1497 | - | 99.99 | 99.93 | 100 | 99.97 |
| | [34] | NN (TF) | 5-fold C-V | - | - | - | 99.68 | 99.87 | 99.73 | 99.80 |
| Fake news | [16] | mBERT_NN (monolingual) | 80 , - , 20 | 1150 | - | 288 | - | 87.17 | 91.89 | 89.47 |
| | | mBERT_NN (multilingual) | | | | | - | 76.47 | 86.66 | 81.25 |
| | [50] | CNN | 54,26 ,20 | 1672 | 823 | 624 | 80 | 73 | 70 | 72 |
| | | BERT | | | | | 80 | - | - | - |
| | [88] | RoBERTa-large (Fakenews-19) | - , - , - | 6420 | 2140 | 2140 | 98.13 | - | - | 98.13 |
| | | BERT-large (Tweets-19) | | | 60 | 200 | 61.10 | - | - | 54.33 |
| | [53] | 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 |
| | [54] | $BERT_{BASE}$ (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 |
| | [56] | C4.5 (Meta-model) | 10-fold C-V | - | - | - | 95.11 | 95.3 | 95.10 | 95.10 |
| | | SVM+RF (Ensemble-model) | | | | | 97.80 | - | - | - |
| | [62] | LR (Multiclass) | 80, -, 20 | 49,029 | - | 12,257 | 96.30 | 31.30 | 23.30 | 25 |
| | | dEFEND w.news (Binary) | | | | | 98 | 92 | 68 | 75 |
| | [89] | Ro-CT-BERT | 60 ,20 ,20 | 6420 | 2140 | 2140 | 99.02 | 99.02 | 99.02 | 99.02 |
| | [65] | C-LSTM | -, -, - | - | - | - | 98.62 | 99.20 | 98.9 | 99.40 |
| | [87] | BERT | 60, 20, 20 | 6420 | 2140 | 2140 | 98.41 | - | - | - |
| | [66] | BERT | 10-fold C-V | - | - | - | 99.40 | 99.40★ | 99.35★ | 99.4★ |
| | [67] | BoW + LR (Story Validity Classification) | 80, -, 20 | 5385 | - | 1346 | - | - | - | 89 |
| | [68] | LR (Count Vector) | 5-fold C-V | - | - | - | - | 93.40 | 93.30 | 93.30 |
| | [90] | MCNNet | -, -, - | - | - | - | 98.20 | 97.50 | 98.70 | 98.10 |
| | [69] | TextCNN | 70, 10, 20 | 1473 | 210 | 421 | - | - | - | 93.80 |
| | [49] | BERT | - ,- ,- | - | - | - | - | 78 | 75 | 76 |
| | [57] | kNN (k-5) | - , -, - | - | - | - | 89 | - | - | 91 |
| | [51] | Cross-SEAN | 80, -,20 | - | - | - | 95.40 | 94.60 | 96.10 | 95.30 |
| | [59] | SVM | 60 ,20 ,20 | 6420 | 2140 | 2140 | 93.46 | 93.48 | 93.46 | 93.46 |
| | [61] | SAFE | 80 ,- ,20 | - | - | - | - | 75.15★ | 75.30★ | 75.25★ |
| | [21] | dEFEND | 75 , - , 25 | 2152 | - | 717 | - | 89.65 | 48.47 | 58.14 |
| | [58] | RoBERTa-large | - , - , - | - | - | - | - | 73.75 | 73.50 | 76 |
| Rumor | [15] | BERT | 10-fold C-V | - | - | - | 99.20 | 99.17 | 98.13 | 98.34 |
| | [72] | CNN with CT-BERT | 97, 2, 1 | 9206 | 150 | 140 | - | 78 | - | - |
| | [71] | XGBoost | -, -, - | - | - | - | 91 | 94 | 85 | 89 |
| | [70] | BERT+BiLSTM | 80,-,20 | 6102 | - | 1077 | - | 73.19 | 73.27 | 72.95 |
| | [73] | LR (COUNT VECTOR) | 10-fold C-V | - | - | - | 84.03 | 81.04 | 80.03 | 80.50 |
| | [74] | DNN | 5-fold C-V | - | - | - | - | - | - | 85.98★ |
| Conspiracy | [75] | RoBERTa (Classification of Users Comments) | 80, -, 20 | 2582 | - | 645 | - | - | - | 90.30★ |
| | | SVM (Classification of YouTube Videos) | 10-fold C-V | - | - | - | 89.40 | - | - | - |
| Disinformation | [77] | CANTM | 5-fold C-V | - | - | - | 63.34 | - | - | 55.48 |
| | [76] | BERT (En) (Binary) | 10-fold C-V | - | - | - | - | - | - | 85.6★ |
| | | mBERT (En) (Multiclass) | | | | | - | - | - | 53.48★ |
| | | mBERT (Ar) (Binary) | | | | | - | - | - | 83.96★ |
| | | FastText (Ar) (Multiclass) | | | | | - | - | - | 69.52★ |

En=English, Ar=Arabic, ★ = Macro average calculated, A=Accuracy, P=Precision, R=Recall, F1=F1-Score

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 data sets was observed to be imbalanced which affected the overall performance of the classification. Abhishek Koirala [50] 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 the 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 [16, 49, 76]. 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 used Arabic data for their classification model [73]. So, researchers from all around the world can conduct similar type of studies 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. [14] 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, 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 on 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 [122]. 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 [123].

### *5.5. Multi-modality based detection system*

There is still lack of the study on COVID-19 misinformation detection that used multimodality 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 contents and derived information from different modalities complement each other to detect misinformation [124, 125]. The similarity between the image and the text is very important which can be an 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, it 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 source of the information, topics can assist the current models to acquire more better result. Current studies emphasize mostly on distinguishing misinformation from real information using content of the information. Content information is important to realize the semantic information. But, it is difficult to detect newly emerged misinformation using content information only [126, 127]. So, analyzing false news across sources of the information, topics, and URLs allow one to obtain a better understanding of the information and also help 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, news reports etc. Leveraging a crowdsourcing approach to obtain annotations could relieve the burden of expert checking, but the annotation quality may suffer [128]. 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 [129]. 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 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 week classifier to make one strong classifier which can do more accurate prediction. In the case of misinformation detection system, different variants of ensemble methods can significantly boost up the overall performance of the system [56]. 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., [56, 77, 14, 53, 58]. 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. This types of characteristics impact the execution of the model in real-life situations and produces biased results. Perfect combination of dropout layer with other layers, use of different kinds of regularization methods (e.g., weight decay) can reduce this problem. Although, these processes need much investigation by '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 an goal in an interactive and uncertain environment. The computer employs 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 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 paper, we presented a systematic survey outlining existing research works on COVID-19 misinformation classification and detection. In particular, we have discussed different kinds of misinformation and existing techniques to detect COVID-19 misinformation focusing on pre-processing and feature extraction methods, classification and detection performance. Comparing with the existing techniques, deep learning based techniques appeared more efficient and effective techniques to classify misinformation accurately, compared to the tranditional machine learning techniques. Although sometimes the performance degrades, traditional machine learning techniques also perform well in the misinformation classification task. We also discussed the limitations of the existing studies and presented several research directions for future investigations. We believe that our survey can provide important insights to build up robust classification systems for detecting misinformation related to COVID-19 and help researchers around the world to come up with new strategies to fight against the spread of misinformation during this pandemic.

## Competing interests statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] M. Fernandez, H. Alani, Online misinformation: Challenges and future directions, in: Companion Proceedings of the The Web Conference 2018, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 595–602. doi:10.1145/3184558.3188730.

[2] L. Wu, F. Morstatter, K. M. Carley, H. Liu, Misinformation in social media: Definition, manipulation, and detection, SIGKDD Explor. Newsl. 21 (2) (2019) 80–90. doi:10.1145/3373464.3373475.

[3] Q. Su, M. Wan, X. Liu, C.-R. Huang, Motivations, methods and metrics of misinformation detection: An nlp perspective, Natural Language Processing Research 1 (2020) 1–13. doi:10.2991/nlpr.d.200522.001.

[4] Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic literature review on the spread of health-related misinformation on social media, Social Science & Medicine 240 (2019) 112552. doi:10.1016/j.socscimed.2019.112552.

[5] S. Dhoju, M. M. U. Rony, M. A. Kabir, N. Hassan, A large-scale analysis of health journalism by reliable and unreliable media, in: L. Ohno-Machado, B. Séroussi (Eds.), 17th World Congress on Medical and Health Informatics (MedInfo), Vol. 264 of Studies in Health Technology and Informatics, IOS Press, 2019, pp. 93–97. doi:10.3233/SHTI190190.

[6] S. Dhoju, M. Main Uddin Rony, M. Ashad Kabir, N. Hassan, Differences in health news from reliable and unreliable media, in: Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 981–987. doi:10.1145/3308560.3316741.

[7] F. Afsana, M. A. Kabir, N. Hassan, M. Paul, Automatically assessing quality of online health articles, IEEE Journal of Biomedical and Health Informatics 25 (2) (2021) 591–601. doi:10.1109/JBHI.2020.3032479.

[8] B. Huang, K. M. Carley, Disinformation and Misinformation on Twitter during the Novel Coronavirus Outbreak, arXiv (2020) 1–19arXiv:2006.04278.

[9] UNICEF, COVID-19: Frequently asked questions, https://www.unicef.org/stories/novel-coronavirus-outbreak-frequently-asked-questions, accessed: 2021-07-18.

[10] Z. Su, D. McDonnell, J. Wen, M. Kozak, J. Abbas, S. Šegalo, X. Li, J. Ahmad, A. Cheshmehzangi, Y. Cai, L. Yang, Y. T. Xiang, Mental health consequences of COVID-19 media coverage: the need for effective crisis communication practices, Globalization and Health 17 (1) (2021) 1–8. doi:10.1186/s12992-020-00654-4.

[11] S. Busari, B. Adebayo, 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: 2021-08-03.

[12] WHO, Novel Coronavirus, Situation Report – 205 205 (6) (2020) 1–19.

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

[14] M. Elhadad, K. Li, F. Gebali, An Ensemble Deep Learning Technique to Detect COVID-19 Misleading Information, Vol. 1264 AISC, 2021. doi:10.1007/978-3-030-57811-4_16.

[15] S. Chen, Research on Fine-Grained Classification of Rumors in Public Crisis —— Take the COVID-19 incident as an example, E3S Web of Conferences 179 (2020) 02027. doi:10.1051/e3sconf/202017902027.

[16] D. Kar, M. Bhardwaj, S. Samanta, A. P. Azad, No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet DetectionarXiv:2010.06906.

[17] F. Afsana, M. A. Kabir, N. Hassan, M. Paul, Towards domain-specific characterization of misinformation (2020). arXiv:2007.14806.

[18] J. H. Fetzer, Disinformation: The Use of False Information, Minds and Machines 14 (2) (2004) 231–240. doi:10.1023/B:MIND.0000021683.28604.5b.

[19] H. Zhang, A. Kuhnle, J. D. Smith, M. T. Thai, 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), 2018, pp. 266–273. doi:10.1109/ASONAM.2018.8508402.

[20] W.-Y. S. Chou, A. Oh, W. M. Klein, Addressing health-related misinformation on social media, Jama 320 (23) (2018) 2417–2418.

[21] L. Cui, D. Lee, Coaid: Covid-19 healthcare misinformation dataset, arXiv (2020) 1–11arXiv:2006.00885.

[22] D. Lazer, M. Baum, Y. Benkler, A. Berinsky, K. Greenhill, F. Menczer, M. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. Sloman, C. Sunstein, E. Thorson, D. Watts, J. Zittrain, The science of fake news: Addressing fake news requires a multidisciplinary effort, Science 359 (6380) (2018) 1094–1096. doi:10.1126/science.aao2998.

[23] J. M. Bale, Political paranoia v. political realism: On distinguishing between bogus conspiracy theories and genuine conspiratorial politics, Patterns of prejudice 41 (1) (2007) 45–60.

[24] V. Swami, R. Coles, S. Stieger, J. Pietschnig, A. Furnham, S. Rehim, M. Voracek, 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, British Journal of Psychology 102 (3) (2011) 443–463.

[25] K. M. Douglas, R. M. Sutton, M. J. Callan, R. J. Dawtry, A. J. Harvey, Someone is pulling the strings: Hypersensitive agency detection and belief in conspiracy theories, Thinking & Reasoning 22 (1) (2016) 57–77.

[26] J.-W. van Prooijen, K. M. Douglas, Belief in conspiracy theories: Basic principles of an emerging research domain, European journal of social psychology 48 (7) (2018) 897–908.

[27] I. Fritsche, M. Moya, M. Bukowski, P. Jugert, S. de Lemus, O. Decker, I. Valor-Segura, G. Navarro-Carrillo, The great recession and group-based control: Converting personal helplessness into social class in-group trust and collective action, Journal of Social Issues 73 (1) (2017) 117–137.

[28] J.-W. Van Prooijen, K. M. Douglas, Conspiracy theories as part of history: The role of societal crisis situations, Memory studies 10 (3) (2017) 323–333.

[29] S. Shahsavari, P. Holur, T. Wang, T. Tangherlini, V. Roychowdhury, Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news, Journal of Computational Social Science 3. doi:10.1007/s42001-020-00086-5.

[30] X. Lin, X. Liao, T. Xu, W. Pian, K.-F. Wong, Rumor Detection with Hierarchical Recurrent Convolutional Neural Network, in: J. Tang, M.-Y. Kan, D. Zhao, S. Li, H. Zan (Eds.), Natural Language Processing and Chinese Computing, Springer International Publishing, Cham, 2019, pp. 338–348.

[31] Q. Li, Q. Zhang, L. Si, Y. Liu, 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, 2019, pp. 66–75. doi:10.18653/v1/D19-5008.

[32] S. Akon, A. Bhuiyan, COVID-19: Rumors and Youth Vulnerabilities in Bangladesh, 2020.

[33] tutor2u, Misleading Information, https://www.tutor2u.net/psychology/topics/misleading-information, accessed: 2021-02-27.

[34] M. K. Elhadad, K. F. Li, F. Gebali, Detecting Misleading Information on COVID-19, IEEE Access 8 (2020) 165201–165215. doi:10.1109/access.2020.3022867.

[35] R. M. Losee, A discipline independent definition of information, Journal of the American Society for Information Science 48 (3) (1997) 254–269. doi:10.1002/(SICI)1097-4571(199703)48:3<254::AID-ASI6>3.0.CO;2-W.

[36] L. Zhou, J. K. Burgoon, J. F. Nunamaker, D. Twitchell, Automating Linguistics-Based Cues for Detecting Deception in Text-Based

Asynchronous Computer-Mediated Communications, Group Decision and Negotiation 13 (1) (2004) 81–106. doi:10.1023/B:GRUP.0000011944.62889.6f.

[37] N. Karlova, K. Fisher, A social diffusion model of misinformation and disinformation for understanding human information behaviour, Inf. Res. 18.

[38] D. Bernard, A Man Drank a Bottle of Rubbing Alcohol for COVID-19, https://www.medpagetoday.com/infectiousdisease/covid19/86094, accessed: 2020-04-22.

[39] M. H. Nguyen, J. Gruber, J. Fuchs, W. Marler, A. Hunsaker, E. Hargittai, Changes in Digital Communication During the COVID-19 Global Pandemic: Implications for Digital Inequality and Future Research, Social Media + Society 6 (3). doi:10.1177/2056305120948255.

[40] T. Papadopoulos, K. N. Baltas, M. E. Balta, The use of digital technologies by small and medium enterprises during COVID-19: Implications for theory and practice, International Journal of Information Management 55 (2020) 102192. doi:10.1016/j.ijinfomgt.2020.102192.

[41] P. N. Petratos, Misinformation, disinformation, and fake news: Cyber risks to business, Business Horizonsdoi:10.1016/j.bushor.2021.07.012.

[42] H. O.-Y. Li, A. Bailey, D. Huynh, J. Chan, YouTube as a source of information on COVID-19: a pandemic of misinformation?, BMJ Global Health 5 (5). doi:10.1136/bmjgh-2020-002604.

[43] B. O. Ahinkorah, E. K. Ameyaw, J. E. Hagan, A.-A. Seidu, T. Schack, Rising Above Misinformation or Fake News in Africa: Another Strategy to Control COVID-19 Spread, Frontiers in Communication 5 (2020) 45. doi:10.3389/fcomm.2020.00045.

[44] Y. Mejova, K. Kalimeri, Advertisers jump on coronavirus bandwagon: Politics, news, and business, arXiv preprint arXiv:2003.00923.

[45] K. Shimizu, 2019-ncov, fake news, and racism, The lancet 395 (10225) (2020) 685–686.

[46] M. Spring, Social media firms fail to act on covid-19 fake news, BBC News.

[47] S. Tasnim, M. M. Hossain, H. Mazumder, Impact of rumors and misinformation on covid-19 in social media, Journal of preventive medicine and public health 53 (3) (2020) 171–174.

[48] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, T. P. Group, Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement, PLOS Medicine 6 (7) (2009) 1–6. doi:10.1371/journal.pmed.1000097.

[49] G. K. Shahi, D. Nandini, FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19, arXivarXiv:2006.11343, doi:10.36190/2020.14.

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

[51] W. S. Paka, R. Bansal, A. Kaushik, S. Sengupta, T. Chakraborty, Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection, Applied Soft Computing 107 (2021) 107393.

[52] Y. Madani, M. Erritali, B. Bouikhalene, Using artificial intelligence techniques for detecting covid-19 epidemic fake news in moroccan tweets, Results in Physics 25 (2021) 104266.

[53] T. Hossain, COVID LIES : Detecting COVID-19 Misinformation on Social Media.

[54] Z. Boukouvalas, C. Mallinson, E. Crothers, N. Japkowicz, A. Piplai, S. Mittal, A. Joshi, T. Adalı, Independent Component Analysis for Trustworthy Cyberspace during High Impact Events: An Application to Covid-19 (2020) 1–9arXiv:2006.01284.

[55] Y. Li, B. Jiang, K. Shu, H. Liu, Toward A Multilingual and Multimodal Data Repository for COVID-19 Disinformation, Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020 (2020) 4325–4330doi:10.1109/BigData50022.2020.9378472.

[56] M. Al-Rakhami, A. Al-Amri, Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter, IEEE Access 8 (2020) 155961–155970. doi:10.1109/ACCESS.2020.3019600.

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

[58] S. Kumar, K. M. Carley, A Fine-Grained Analysis of Misinformation in COVID-19 Tweets.

[59] P. Patwa, S. Sharma, S. PYKL, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an Infodemic: COVID-19 Fake News DatasetarXiv:2011.03327.

[60] N. Micallef, B. He, S. Kumar, M. Ahamad, N. Memon, The Role of the Crowd in Countering Misinformation: A Case Study of the

COVID-19 InfodemicarXiv:2011.05773.

[61] X. Zhou, A. Mulay, E. Ferrara, R. Zafarani, ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research, Association for Computing Machinery, New York, NY, USA, 2020, p. 3205–3212.

[62] A. Dharawat, I. Lourentzou, A. Morales, C. Zhai, Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformationarXiv:2010.08743.

[63] An exploratory study of COVID-19 misinformation on Twitter, Online Social Networks and Media 22 (2021) 100104. doi:10.1016/j.osnem.2020.100104.

[64] F. Haouari, M. Hasanain, R. Suwaileh, T. Elsayed, ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation DetectionarXiv:2010.08768.

[65] R. Kaliyar, A. Goswami, P. Narang, 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, Vol. 2, 2021, pp. 1066–1072.

[66] J. Ayoub, X. Yang, F. Zhou, Combat COVID-19 infodemic using explainable natural language processing models, Information Processing and Management 58 (4). doi:10.1016/j.ipm.2021.102569.

[67] L. H. X. Ng, K. M. Carley, “The coronavirus is a bioweapon”: classifying coronavirus stories on fact-checking sites, Computational and Mathematical Organization Theory 27 (2) (2021) 179–194. doi:10.1007/S10588-021-09329-W.

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

[69] C. Yang, X. Zhou, R. Zafarani, CHECKED: Chinese COVID-19 fake news dataset, Social Network Analysis and Mining 11 (1) (2021) 1–8. arXiv:2010.09029, doi:10.1007/s13278-021-00766-8.

[70] H. WANG, J. GAN, J. CHEN, Z. OUYANG, Automatic detecting for covid-19-related rumors data on internet, in: 2021 9th International Conference on Communications and Broadband Networking, ICCBN 2021, Association for Computing Machinery, New York, NY, USA, 2021, p. 22–26. doi:10.1145/3456415.3456420.

[71] A. Shi, Z. Qu, Q. Jia, C. Lyu, Rumor detection of covid-19 pandemic on online social networks, in: 2020 IEEE/ACM Symposium on Edge Computing (SEC), IEEE, 2020, pp. 376–381.

[72] R. Alkhalifa, T. Yoong, E. Kochkina, A. Zubiaga, M. Liakata, QMUL-SDS at CheckThat! 2020: Determining COVID-19 Tweet Check-Worthiness Using an Enhanced CT-BERT with Numeric Expressions (2020) 22–25arXiv:2008.13160.

[73] L. Alsudias, P. Rayson, COVID-19 and Arabic Twitter : How can Arab World Governments and Public Health Organizations Learn from Social Media?, ACL 2020 Workshop NLP-COVID Submission (2020) 1–9.

[74] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, P. Bogdan, A COVID-19 Rumor Dataset, Frontiers in Psychology 12. doi:10.3389/fpsyg.2021.644801.

[75] J. Carlos, M. Serrano, O. Papakyriakopoulos, S. Hegelich, NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube, ACL 2020 Workshop NLP-COVID (2018).

[76] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. D. S. Martino, A. Abdelali, H. Sajjad, K. Darwish, P. Nakov, Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to ArmsarXiv:2007.07996.

[77] X. Song, J. Petrak, Y. Jiang, I. Singh, D. Maynard, K. Bontcheva, Classification Aware Neural Topic Model and its Application on a New COVID-19 Disinformation CorpusarXiv:2006.03354.

[78] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch, S. Dietze, TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic, International Conference on Information and Knowledge Management, Proceedings (June) (2020) 2991–2998. arXiv:2006.14492, doi:10.1145/3340531.3412765.

[79] R. Lamsal, Design and analysis of a large-scale COVID-19 tweets dataset, Applied Intelligence (October). doi:10.1007/s10489-020-02029-z.

[80] U. Qazi, M. Imran, F. Ofli, GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information, arXiv (2020) 1–3arXiv:2005.11177.

[81] J. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, K. Artemova, E. Tutubalina, G. Chowell, A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration, 2020.

[82] S. Alqurashi, A. Alhindi, E. Alanazi, Large Arabic Twitter Dataset on COVID-19, arXiv (2020) 2–4arXiv:2004.04315.

[83] E. Chen, K. Lerman, E. Ferrara, Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set (Preprint), 2020. doi:10.2196/preprints.19273.

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

[85] G. Preda, COVID19 Tweets, https://www.kaggle.com/gpreda/covid19-tweets, accessed: 2021-08-01.

[86] Z. Gao, S. Yada, S. Wakamiya, E. Aramaki, NAIST COVID: Multilingual COVID-19 Twitter and Weibo DatasetarXiv:2004.08145.

[87] A. Wani, I. Joshi, S. Khandve, V. Wagh, R. Joshi, Evaluating Deep Learning Approaches for Covid19 Fake News Detection, Vol. 1402 CCIS, 2021. doi:10.1007/978-3-030-73696-5_15.

[88] Y. Bang, E. Ishii, S. Cahyawijaya, Z. Ji, P. Fung, Model generalization on covid-19 fake news detection, arXiv preprint arXiv:2101.03841.

[89] B. Chen, B. Chen, D. Gao, Q. Chen, C. Huo, X. Meng, W. Ren, Y. Zhou, Transformer-Based Language Model Fine-Tuning Methods for COVID-19 Fake News Detection, Vol. 1402 CCIS, 2021. doi:10.1007/978-3-030-73696-5_9.

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

[91] K. P. F.R.S., Liii. on lines and planes of closest fit to systems of points in space, Philosophical Magazine Series 1 2 559–572.

[92] J. Hérault, C. Jutten, Space or time adaptive signal processing by neural network models, 1987.

[93] S. Robertson, Understanding inverse document frequency: On theoretical arguments for idf, Journal of Documentation - J DOC 60 (2004) 503–520. doi:10.1108/00220410410560582.

[94] J. Pennebaker, M. Francis, R. Booth, Linguistic inquiry and word count (liwc).

[95] X. Zhou, A. Jain, V. V. Phoha, R. Zafarani, Fake news early detection: A theory-driven model, Digital Threats: Research and Practice 1 (2). doi:10.1145/3377478.

[96] W. MANN, S. Thompson, Rethorical structure theory: Toward a functional theory of text organization, Text 8 (1988) 243–281. doi:10.1515/text.1.1988.8.3.243.

[97] Y. Ji, J. Eisenstein, 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, 2014, pp. 13–24. doi:10.3115/v1/P14-1002.

[98] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems 26.

[99] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, arXiv preprint arXiv:1607.01759.

[100] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1 (Mlm) (2019) 4171–4186. arXiv:1810.04805.

[101] M. Cheng, S. Nazarian, P. Bogdan, VRoC: Variational Autoencoder-Aided Multi-Task Rumor Classifier Based on Text, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 2892–2898. doi:10.1145/3366423.3380054.

[102] M. Müller, M. Salathé, P. E. Kummervold, COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on TwitterarXiv:2005.07503.

[103] J. Pennington, R. Socher, C. Manning, 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, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.

[104] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, 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, 2018, pp. 2227–2237. doi:10.18653/v1/N18-1202.

[105] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1007/BF00994018.

[106] Y. Tausczik, J. Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods, Journal of Language and Social Psychology 29 (2010) 24–54. doi:10.1177/0261927X09351676.

[107] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, Proceedings of the National Conference on Artificial Intelligence 3 (2015) 2267–2273.

[108] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, AAAI Press, 2016, p. 2873–2879.

[109] S. Hochreiter, J. Urgen Schmidhuber, Long Shortterm Memory, Neural Computation 9 (8) (1997) 17351780.

[110] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, 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, 2016, pp. 207–212. doi:10.18653/v1/P16-2034.

[111] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (December). arXiv:1412.3555.

[112] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference (2020) 3982–3992arXiv:1908.10084, doi:10.18653/v1/d19-1410.

[113] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based Model for Arabic Language Understanding, arXivarXiv:2003.00104.

[114] D. Card, C. Tan, N. A. Smith, Neural models for documents with metadata, ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) 1 (2018) 2031–2040. arXiv:1705.09296, doi:10.18653/v1/p18-1189.

[115] P. H. Seo, Z. Lin, S. Cohen, X. Shen, B. Han, Hierarchical Attention Networks, ArXiv (2016) 1480–1489arXiv:1606.02393.

[116] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, Defend: Explainable fake news detection, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (May) (2019) 395–405. doi:10.1145/3292500.3330935.

[117] X. Zhou, J. Wu, R. Zafarani, 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) (2020) 354–367. arXiv:2003.04981, doi:10.1007/978-3-030-47436-2_27.

[118] L. Cui, S. Wang, D. Lee, Same: Sentiment-aware multi-modal embedding for detecting fake news, Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019 (2019) 41–48doi:10.1145/3341161.3342894.

[119] N. Ruchansky, S. Seo, Y. Liu, CSI: A hybrid deep model for fake news detection, International Conference on Information and Knowledge Management, Proceedings Part F1318 (2017) 797–806. arXiv:1703.06959, doi:10.1145/3132847.3132877.

[120] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 2017-Decem (Nips) (2017) 5999–6009. arXiv:1706.03762.

[121] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv (2019) 1–43arXiv:1904.09675.

[122] B. Bhutani, N. Rastogi, P. Sehgal, A. Purwar, Fake News Detection Using Sentiment Analysis, 2019 12th International Conference on Contemporary Computing, IC3 2019 (2019) 1–5doi:10.1109/IC3.2019.8844880.

[123] B. Ghanem, P. Rosso, F. Rangel, An Emotional Analysis of False Information in Social Media and News Articles, ACM Transactions on Internet Technology 20 (2) (2020) 1–17. arXiv:1908.09951, doi:10.1145/3381750.

[124] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, Spotfake: A multi-modal framework for fake news detection, in: 2019

IEEE Fifth International Conference on Multimedia Big Data (BigMM), 2019, pp. 39–47. doi:10.1109/BigMM.2019.00-44.

[125] C. Song, N. Ning, Y. Zhang, B. Wu, A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks, Information Processing & Management 58 (1) (2021) 102437. doi:https://doi.org/10.1016/j.ipm.2020.102437.

[126] N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, R. Zafarani, Credibility-Based Fake News Detection, Springer International Publishing, Cham, 2020, pp. 163–182. doi:10.1007/978-3-030-42699-6_9.

[127] K. Shu, S. Wang, H. Liu, 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, WSDM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 312–320. doi:10.1145/3289600.3290994.

[128] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, M. Gomez-Rodriguez, 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, WSDM '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 324–332. doi:10.1145/3159652.3159734.

[129] J. Charles F. Bond, B. M. DePaulo, Accuracy of Deception Judgments, Personality and Social Psychology Review 10 (3) (2006) 214–234. doi:10.1207/s15327957pspr1003_2.

# 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://github.com/thepanacealab/covid19_twitter

[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