IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 November 2022, accepted 2 January 2023, date of publication 3 February 2023, date of current version 16 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242234

SURVEY

Machine Learning Techniques for Sentiment Analysis of COVID-19-Related Twitter Data

NIKLAS BRAIG[1], **ALINA BENZ**[1], **SOEREN VOTH**[1], **JOHANNES BREITENBACH**[1], **AND RICARDO BUETTNER**[1,2], **(Senior Member, IEEE)**
[1]Chair of Information Systems and Data Science, University of Bayreuth, 95447 Bayreuth, Germany
[2]Fraunhofer FIT, 95444 Bayreuth, Germany

Corresponding author: Johannes Breitenbach (johannes.breitenbach@uni-bayreuth.de)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant 491183248, in part by the Federal Ministry for Economic Affairs and Climate Action [Bundesministerium für Wirtschaft und Klimaschutz (BMWK)] under Grant 16KN088849, and in part by the Open Access Publishing Fund of the University of Bayreuth.

ABSTRACT On Twitter, COVID-19 is a highly discussed topic. People worldwide have used Twitter to express their viewpoints and feelings during the pandemic. Previous research has focused on particular topics such as the public's sentiment during the lockdown, their opinion on governmental measures, or their stance towards COVID-19 vaccines. However, until today, there is no comprehensive overview that presents possible areas of application for sentiment analysis of COVID-19 Twitter data. Therefore, this study reveals how sentiment analysis can provide relevant insights for managing the pandemic by applying a behavioral and social science lens. In this context, our systematic literature review focuses on machine learning-based sentiment analysis techniques and compares the best-performing classification algorithms for COVID-19-related Twitter data. We performed a search in five databases, which are: IEEE Xplore DL, ScienceDirect, SpringerLink, ACM DL, and AIS Electronic Library. This search resulted in 40 papers published between October 2019 and January 2022 that used sentiment analysis to evaluate the public opinion on COVID-19-related topics, which we further investigated. Our research indicates that the best performing models in terms of accuracy are ensemble models that comprise various machine learning classifiers. Especially BERT and RoBERTa models provide the most promising results when fine-tuned on Twitter data. Our study aims to combine machine learning-based sentiment analysis and insights from social and behavioral science to provide decision-makers and public health experts with guidance on the application of sentiment analysis in the fight against the spread of COVID-19.

INDEX TERMS Behavioral science, COVID-19, deep learning, machine learning, sentiment analysis, social science, twitter.

I. INTRODUCTION

Researchers from a variety of disciplines have rushed to provide solutions to mitigate the impact of the COVID-19 pandemic that has so far claimed the lives of more than five million people [1], [2]. Since the outbreak of COVID-19, social media has become pivotal in staying connected with family, friends, and colleagues, but also to stay informed and discuss new policy updates and regulations [3]. Governments and other organizations, such as the World Health Organization (WHO), have used social media as a direct communication channel to disseminate information and manage the crisis [4], [5]. COVID-19 is, until today, one of the most discussed topics on Twitter. In the period from January to May 2020 alone, more than 120 million messages related to COVID-19 were published on Twitter [6]. This abundance of unfiltered Twitter data offers promising opportunities for public health research [7]. Besides, compared to other social media platforms, its uncomplicated access via the Twitter API makes it the go-to platform for researchers [8], [9], [10]. A frequently used approach to extract and analyze large

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

volumes of fuzzy, user-generated data is called sentiment analysis [11].

Especially in the context of COVID-19, prior studies have shown that sentiment analysis could support governments and health authorities in their fight against COVID-19 in multiple ways: Twitter users' "digital footprints" provide insights into their emotional and behavioral state and, thus, could reveal relevant information about the public's acceptance of containment measures, vaccine readiness or trust in the government. In particular, vaccine hesitancy is a substantial threat to the success of public health campaigns. Previous studies have shown that sentiment analysis provides a powerful tool to study the concerns and sentiments causing vaccine hesitancy in real-time on Twitter [12]. These insights can be used to adjust communication strategies and to develop more target-oriented policy measures that aim at stimulating the vaccine uptake [12], [13], [14]. Moreover, it allows to identify and monitor trends, and misinformation [15].

Research from the past century highlights that managing a pandemic requires a substantial behavior change - not at least due to the adaption to physical distancing rules [1]. Hence, social and behavioral science research can provide valuable insights into responding effectively to COVID-19 while minimizing psychological distress. Prior studies have already combined well-known social and behavioral theories [16], [17], [18] with the domain of sentiment analysis [19], [20], [21].

In this study, we provide a comprehensive overview of possible sentiment analysis applications in the context of COVID-19-related Twitter data from a social and behavioral science lens. By drawing on the topics relevant to managing the COVID-19 pandemic [1], we note how sentiment analysis can generate solid data foundations to make well-informed decisions by gaining insights into the public's thinking, feeling, and acting. Since machine learning (ML) classifiers are used with increasing popularity for classifying tweets [22], [23], our primary focus lies on the ML algorithms that are used for sentiment classification. Therefore, we extend the current literature by synthesizing literature from behavioral and social science and sentiment analysis. Our goal is to examine how sentiment analysis can be applied in the context of COVID-19 and which ML classifiers offer the most promising results. We use the categories with origin in behavioral theory (Fig. 6) proposed by van Bavel et al. [1] as guidance to equip governments and public-health experts with a manual to use the power of data and ML to manage the pandemic.

To the best of our knowledge, there exists no such literature review. Existing literature reviews focused on multiple diseases at an early stage of COVID-19 [24], investigated different social media platforms, or had a single focus, such as vaccination [25], [26]. We examine the classification techniques used for sentiment analysis to a greater extent and, thereby, aim at accelerating the development of new and better approaches that can support governments and health authorities in their daily fight against COVID-19. In the face of possible new emerging virus variants, taking into account new insights from Twitter data could change the course of the current pandemic and future public health crises.

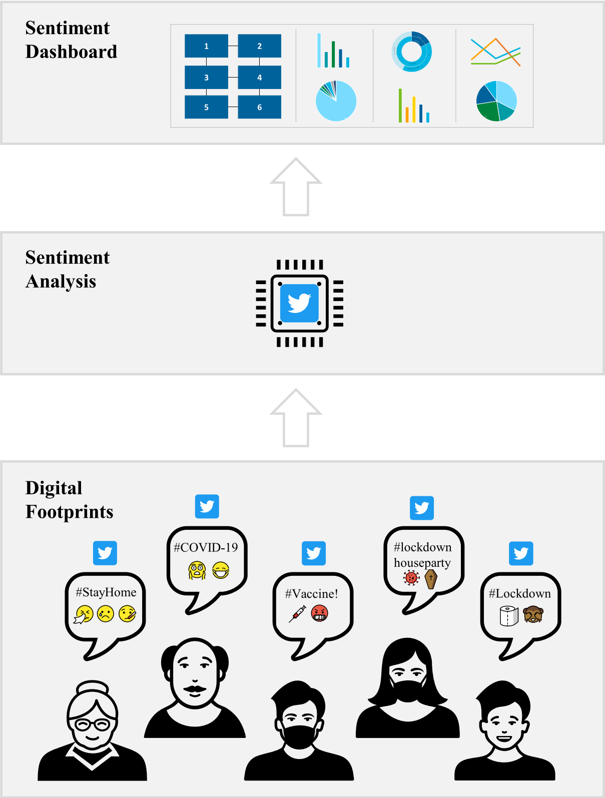


FIGURE 1. Sentiment analysis of COVID-19-related tweets can provide relevant insights for effectively managing the pandemic and its impacts.

This literature review is organized as follows: In Section II, we present our research methodology. Section III comprises a brief introduction to sentiment analysis and the techniques used for sentiment analysis of Twitter data. Section IV presents our findings from a social and behavioral science perspective, as well as from a technical perspective. In Section V, the results of the previous section are discussed in detail. Section VI summarizes our main findings, reiterates current challenges and limitations for sentiment analysis on COVID-19-related Twitter data, and offers suggestions for future directions of research.

II. METHODOLOGY

The research on sentiment analysis of COVID-19-related Twitter data is widely fragmented and primarily focuses on a specific topic of interest. This study applies the technique of a systematic literature review to present an overview of possible areas of application for sentiment analysis on COVID-19-related Twitter data. Since the management of the pandemic requires governments and health authorities to influence the behavior of the public in a certain way to contain the spread of the coronavirus, we apply a social and behavioral science lens. Apart from the social and behavioral perspective,

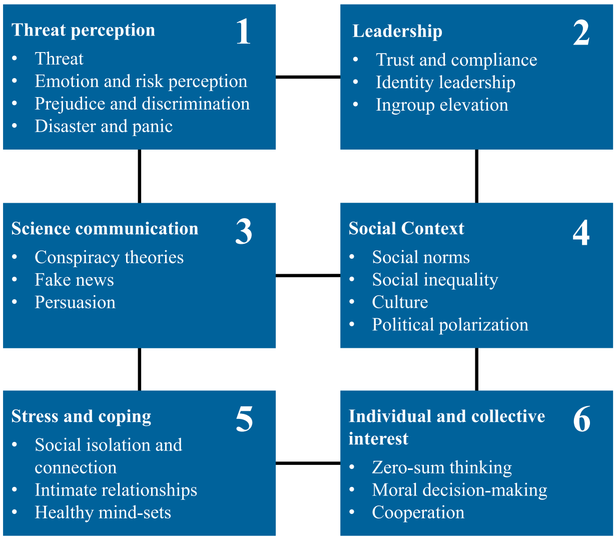


FIGURE 2. Relevant topics for managing the COVID-19 pandemic from a social and behavioral science perspective according to van Bavel et al. [1].

a particular focus lies on the classification techniques such as the used algorithms, feature extraction methods, and datasets. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for this research. The method was originally developed for health sciences and, thus, is well suited for our research topic. Due to its structured application, it was adopted in many other fields for reviewing existing literature. According to the PRISMA guidelines, the overall selection process of studies can be divided into four steps: Identification, Screening, Eligibility, and Inclusion. In the first two steps, we searched the literature in the five databases, which are further described in the following subsection, and evaluated our search hits based on our inclusion criteria. In the eligibility step, each article was evaluated utilizing a full-text search. In case an article met our inclusion criteria, it was included in the review and, thus, forwarded to the inclusion step [27], [28]. The four phases are described in more detail in the subsequent paragraphs and Fig. 3.

A. DATABASES

To collect literature for our review, we performed searches in five academically-renowned databases that were selected due to their scientific recognition and suitability for this type of research: (1) IEEE Xplore DL, (2) ScienceDirect, (3) SpringerLink, (4) ACM DL, and (5) AIS Electronic Library.

B. SEARCH STRATEGY

Our search term was developed in an iterative process. It comprises keywords related to our research question as well as the underlying framework of van Bavel et al. [1]. The search string can be broken down into four clusters that lead to relevant search hits for our literature review. The first cluster relates to the domain of sentiment analysis and captures its synonyms, such as "opinion analysis" and "opinion mining". To come up with the relevant synonyms we draw on the research by Maentylae et al. [29], Cambria et al. [30] and Alamoodi et al. [24]. The second cluster aims at capturing only research papers that are based on Twitter data or so-called Twitter "Tweets". Cluster three is set up in the fashion to include the specific technical characteristics of literature that falls into the spectrum of machine learning or deep learning. By comparing related literature and scrutinizing our search hits in each database, we refined and optimized our search string gradually [31], [32].

TABLE 1. Search string validation in IEEE Xplore DL.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Outcome
("Sentiment Analysis" OR "Opinion Analysis" OR "Opinion Mining")	("Twitter" OR "Tweet*")	("Machine Learning" OR "NLP" OR "ML" OR "Deep Learning" OR "Natural Language Processing" OR "DL")	("COVID-19" OR "Pandemic" OR "Corona")	Conferences: 62, Journals: 9, Early Access Articles: 4

We found out that the abbreviations of machine learning ("ML") and deep learning ("DL") did not provide additional relevant search hits and were thus excluded – as illustrated by the color red in Table 1. Only the individual keywords that provided additional relevant search results were included. For our last cluster that captures literature related to the COVID-19 pandemic, we used the keywords of the study by Yousefinaghani et al. [12] as guidance and further considered the term "pandemic" as highlighted by van Bavel et al. [1].

Table 1 shows the process exemplary for the IEEEXplore DL database. The terms displayed in green color were important for generating search hits and, therefore, included in the final search string. The terms stated in red did not lead to additional relevant search results and were, thus, excluded from the search string. In our search query, the individual keywords are connected with the boolean operators AND or OR.

We searched all five databases using the following keywords divided into four main clusters:
("Sentiment Analysis" OR "Opinion Analysis" OR "Opinion Mining") AND ("Twitter" OR "Tweet*") AND ("Machine Learning" OR "Deep Learning" OR "Natural Language Processing" OR "NLP") AND ("COVID-19" OR "Pandemic")
Due to the search limitations of ScienceDirect, the search string was slightly adapted, as illustrated in Appendix D.

C. INCLUSION AND EXCLUSION CRITERIA

In order to select the most relevant articles, we applied specific inclusion and exclusion criteria. First of all, to cover the time of the COVID-19 pandemic, the publication date of studies has to be between 01 October 2019 and 31 January 2022 (most recent). Secondly, we included journal pre-proofs. Thirdly, only international peer-reviewed journal articles in the five selected databases written in English were considered. Other types of publication, such as datasets or literature reviews, were not included.

D. SPECIFIC SELECTION CRITERIA

Further criteria address content-specific properties: Only articles that apply a (1) sentiment analysis approach, (2) that used Twitter datasets, and (3) that analyzed a topic related to COVID-19 were included in our literature review. Regarding relevancy, research papers that had a too narrow, domain-specific topic or were irrelevant for our review were excluded.

E. STUDY SELECTION

By following our search strategy in accordance with the PRISMA guidelines, the first article search resulted in 425 articles. We received the following search results sorted by the selected database: (1) IEEE Xplore DL: 9 articles, (2) ScienceDirect: 234 articles, (3) SpringerLink: 153 articles, (4) ACM DL: 18 articles, (5) AIS Electronic Library: 11 articles. One study from the ScienceDirect search was excluded because it was a duplicate. We carefully scrutinized the list of all retrieved publications and included only those that satisfied our inclusion criteria. Out of the initial 425 articles, 47 articles were excluded on the grounds of our exclusion criteria, resulting in 378 articles that were forwarded for screening. The three authors screened the studies' titles, keywords, and abstracts in detail on the basis of the previously illustrated eligibility criteria. The entire paper was inspected if the authors' opinion regarding the ex- or inclusion of an article diverged. After an in-depth screening of the studies, 40 articles were identified as relevant for answering our research question.

The identification process was based on a discussion by the three reviewers at an inter-reviewer reliability of 92%. The reliability was assessed after reviewing the articles. The reliability is the sum of 309 unanimously rejected and 38 accepted articles, divided by the total number of 378 articles examined. Fig. 3 illustrates our literature search approach.

F. DATA EXTRACTION

All three researchers thoroughly extracted relevant data from the included journal articles. In agreement, we determined the set of features that will be analyzed in order with the purpose of our study. General characteristics were extracted, such as title, author, publication year, published journal, and keywords (author).

The main part of our analysis concerns the methodological aspects of the included studies. This comprises information about the time period of the Twitter dataset, its language, the number of tweets of the training and evaluation dataset, the classification algorithms, and their accuracy and features. In addition, all studies were screened for their respective areas of application. Table 2 provides a detailed collection of the analyzed features.

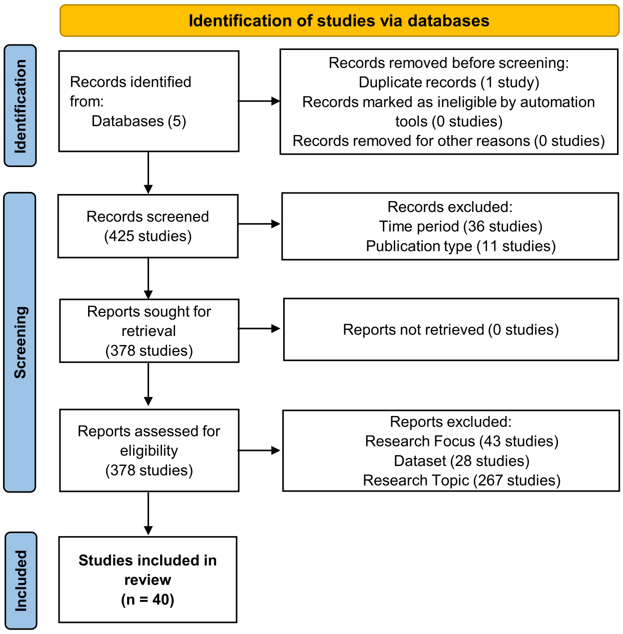

Identification of studies via databases
Identification
Records identified from:
Databases (5)
Records removed before screening:
Duplicate records (1 study)
Records marked as ineligible by automation tools (0 studies)
Records removed for other reasons (0 studies)
Screening
Records screened (425 studies)
Records excluded:
Time period (36 studies)
Publication type (11 studies)
Reports sought for retrieval (378 studies)
Reports not retrieved (0 studies)
Reports assessed for eligibility (378 studies)
Reports excluded:
Research Focus (43 studies)
Dataset (28 studies)
Research Topic (267 studies)
Included
Studies included in review (n = 40)

FIGURE 3. PRISMA flow diagram that illustrates our identification process of suitable literature [27].

TABLE 2. Elements extracted from the included literature.

Data item	Description
General characteristics	Title, authors, publication year, journal and keywords (author)
ML algorithm	Used sentiment analysis method in the research
Dataset	Origin, sample size, time period, language, hashtags
Performance Parameters	Accuracy/recall/precision/F1-Score to compare different approaches
Use-cases	Possible application areas
Research focus	Research foundation and research gap

III. FOUNDATIONS FOR SENTIMENT ANALYSIS ON TWITTER

A. TWITTER AND COVID-19

Twitter is a microblogging platform where people can express their feelings, emotions, and opinions in short messages, so-called "tweets", that can contain up to 280 characters. Each day, more than 500 million messages are posted on Twitter from users worldwide. This enables information to travel in real-time across the globe and disseminate across a large, diverse base of users [33]. From country leaders to scientists and healthcare organizations, Twitter has become an important medium to communicate policy updates and information [34]. One of the most active accounts during the pandemic was the one of WHO (Twitter: @WHO) [5]. Twitter has become the go-to platform for citizens to share their opinions and discuss pandemic-related information.

In the first 90 days, users already posted more than 520 million tweets with COVID-19-related hashtags. This represents approximately 8 million tweets per day [35]. In recent times, the microblogging network has evolved as an important pillar in public health research and monitoring [3], [7]. Studies show that analyzing Twitter messages can help to identify, monitor, or forecast diseases [36]. Due to its uncomplicated access compared to other platforms, most research in this field is based on Twitter data [8]. As one of the largest and openly available databases, researchers can select and download specific data based on keywords and hashtags using Twitter API. There is a myriad of COVID-19-related tweets, ranging from informative and morale-boosting statements to emotionally-laden opinions. A few examples that used the COVID-19 hashtag (#COVID-19) are:

> "This morning, I tested positive for COVID-19. I'm feeling fine – and I'll continue to work remotely this week while following public health guidelines. Everyone, please get vaccinated and get boosted."

> "New Zealand is back in lockdown because four people have got Covid 19. The biggest overreaction in the history of humanity since, I dunno, probably last week. I'm losing track now."

> "My uncle passed away yesterday from COVID-19. I have been begging him to take the vaccine for the past 3 months. He refused because of fears trumped up by WhatsApp and FB university. The virus may have killed him, but it was disinformation that led him to this demise."

B. SENTIMENT ANALYSIS ON TWITTER

Sentiment analysis denotes the techniques and methods used to automate the extraction and analysis of sentiments in a text. Principally, language can convey mostly two types of information: One is factual, objective information, and the other is subjective, emotionally-laden information [37]. The latter can comprise a variety of human moods that range from opinions or assessments to views, attitudes, and emotions [38]. A large bunch of research in sentiment analysis has focused on polarity classification [30]. Sentiments can be classified binary (positive or negative) or three-way (positive, negative, or neutral) [33]. Apart from detecting polarity, sentiment analysis can be applied to detect and model sentiment topics, opinions, emotions, and social or political orientations. Sentiment analysis and opinion mining (OM) are commonly used interchangeably in literature [11]. Sentiments can be predominantly analyzed on three levels: the document, sentence, and aspect (entity) level. At the document level, the text's sentiment is analyzed as a whole. At the sentence level, each sentence's polarity is assessed and classified accordingly. Furthermore, at the aspect level, the polarity of a particular object (entity) is identified [39]. Sentiment analysis on Twitter, which is also referred to as Twitter sentiment analysis, is a specialized subfield of sentiment analysis [33]. Due to the restricted length of tweets, there is no substantial difference between the document and sentence level. Thus, for Twitter sentiment analysis, the analysis can be applied either on the sentence (here: Tweet/message) level or on the entity level [23]. Identifying sentiments on Twitter comes not without challenges: In contrast to other types of text, such as blog articles or forums, Twitter messages contain a variety of linguistic particularities like the informal style of writing and length limitations [33].

C. OVERVIEW OF TECHNIQUES FOR SENTIMENT ANALYSIS

There are mainly three techniques for sentiment classification: lexicon-based (knowledge-based), machine learning (statistical methods), and hybrid approaches [40]. Lexicon-based approaches determine the polarity of a text based on the individual polarity of the words that are present in the text [40]. These kinds of methods rely on dictionaries, such as WordNet [41], that contain the polarity values of a large corpus of words [40]. The text's overall polarity can be calculated by adding up the individual polarity scores [42]. For instance, if more positive than negative words are included in a text, the overall polarity is positive [43]. In case the text contains an equal amount of positive and negative words, the overall sentiment is neutral [44]. The main advantage of this approach is that no labeled training data is required, and it can be easily adapted to different languages [42]. This type of approach is considered an unsupervised learning method [42]. The performance of lexicon-based methods mainly depends on the quality of the lexicon resource. Lexical-based approaches are, to a high degree, domain-specific, which means that words have a different meaning based on their context [43]. A shortcoming is their inability to deal with semantic rules and linguistic specificities, such as negation, slang, or sarcasm, which is prevalent in natural language texts [40].

ML classification techniques are used for sentiment analysis with increasing popularity [22], [23]. Machine learning approaches can be distinguished between supervised and unsupervised methods; however, supervised methods are most commonly used for Twitter sentiment analysis [23] and, therefore, illustrated in the following. As mentioned beforehand, ML-based approaches require training datasets that contain labeled sentiment classes [44]. On these training datasets, classification models are trained and optimized. ML models quantify natural language text based on their feature representation and predict a text's polarity based on its feature value [33]. Their performance depends on how well the selected features classify text [45]. Typical features are bag-of-words (BOW) methods like term frequency-inverse document frequency (TF-IDF) or n-gram, and word embeddings like Word2Vec or GloVe. Word2Vec and GloVe are pre-trained, unsupervised models that create a vector with a cluster of similar words [43], [46]. The learning datasets often contain unstructured data, so-called "noise", such as hashtags, abbreviations, stop-words, poorly structured sentences, spelling mistakes, punctuation, or non-dictionary words that can impair the performance [47]. Thus, prior to the feature extraction, a data pre-processing step is necessary, which is mostly done by applying natural language processing (NLP) techniques. Pre-processing can be separated into normalization, tokenization, and noise reduction. Noise reduction steps can contain, for instance, the replacement of negative mentions, removing URLs, removing capitalization, or replacing acronyms with their unabbreviated form. This ensures extracting only relevant features [48]. To normalize the dataset, stemming and lemmatization are frequently used. With these normalization algorithms, words are transformed into a standard form without a prefix or suffix to enable the identification of the same words with similar meanings [49]. Tokenization splits the text into a sequence of tokens, which makes it usable for ML classifiers models. The tokens are then defined by the BOW representation. N-grams can be chosen as a tokenization method as well [50]. In the pre-processing phase, the order of each step is important. Depending on the classifiers, there are different results with different pre-processing steps and combinations [51].

A major downside of ML methods is the limited availability of required labeled datasets. For ML classification models to work well, usually large datasets are required to optimize the model parameters. Thus, the performance is often related to the availability of labeled training data - which is not often the case for novel topics [23]. ML-based sentiment analysis models are easily adaptable to a certain domain or can be used for a certain purpose [45]. At the same time, domain dependency is one of their main limitations. The models perform well when applied to similar data as the training data (similar domain); however, their performance decreases when applied to an unrelated domain. This implies that the classifier has to be retrained when the domain is changed [45]. The most frequently used classifiers are: Naïve Bayes (NB) [52], Maximum Entropy (ME) [53], Support Vector Machine (SVM) [54], Logistic Regression (LR), and Random Forest (RF) [23], [55]. The performance of sentiment analysis classifiers is generally evaluated in an experimental context. Frequently used indicators are accuracy, precision, recall, and F1-Score [56]. The accuracy denotes the relationship between correct predictions and the total number of predictions. The correct predictions are comprised of true positives and true negatives. The equations for accuracy, recall, precision, and F1-Score are assumed according to Skolova and Lapalme [57].

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

To balance out the disadvantages of lexicon-based and ML approaches, so-called hybrid approaches were suggested. Hybrid approaches combine lexicon-based and ML-based approaches.

D. SUPERVISED MACHINE LEARNING TECHNIQUES FOR TWITTER SENTIMENT ANALYSIS

1) TRADITIONAL MACHINE LEARNING CLASSIFIERS

A characteristic of supervised ML techniques is their dependency on labeled prior data. Traditional ML techniques require that domain experts select the applied features to reduce the complexity of the algorithms. One of the most frequently used traditional ML algorithms is the NB algorithm, which is a probabilistic classifier. The technique is based on the assumption that a certain feature in a category is independent of other features' occurrences. The approach requires pre-assigned labels [43]. Using conditional probability and Bayes Theorem, the posterior probability of a class, i.e., that a selected feature belongs to a certain category, is calculated. The model is suitable for large datasets due to its calculation velocity and simplicity of implementation [46]. This ML model shows better performance on categorical data than on numerical data. SVM aim to find optimal boundaries between the different classes. The ideal separation is achieved when the distance between the individual classes is maximized. SVM creates a set of hyperplanes and uses linear regressions for the classification process. The best separators are the ones with the maximal distance to other classes [58]. These methods need training data and are highly effective for semi-structured or unstructured data [46], and suitable for Twitter sentiment analysis [58]. The performance decreases with ''noisy'' data. A downside of that approach is the long training time required for extensive datasets [43]. Besides SVM, another widely used classifier is the RF algorithm. RF consists of a set of individual tree predictors that operate in combination as an ensemble. The prediction of each tree is collected, and by applying a majority voting, the class that has received the most votes determines the categorization [55].

2) DEEP LEARNING CLASSIFIERS

Deep learning (DL) is a subfield of machine learning that mimics the learning process of human brains [59]. The original idea dates back to the concept of neural networks: Through experience, the composition and strength of neutronic connections of the brain can be adapted [60]. In the last few years, it has become the ''Gold Standard'' in ML, achieving cutting-edge performance on a variety of cognitive tasks [61]. DL has outperformed popular traditional ML techniques in many fields of application, such as in NLP [61]. By using artificial neural networks and many layers of activity vectors, new representations can learn to be discovered [62].

With every level of abstraction, by applying non-linear transformations, the particular representations are elevated to a more abstract level. This mechanism allows filtering for relevant inputs or features for classification tasks. The level of layers reflects the depth of the network. In contrast to models with fewer layers, DL models consist of large amounts of neurons and processing layers. DL models can detect structures from unstructured and unlabeled data by applying a back-propagation algorithm [63]. Another advantage of deep-learning models is that they do not rely on manually designed feature extraction because the ideal features can be extracted automatically [63]. Hence, they do not require the knowledge of a domain expert. The most common techniques used in DL are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT).

A CNN is a feed-forward network that was inspired by the mechanisms of visual perception. A CNN multi-layered architecture aims to lower the number of parameters in the process of a general feed-forward backpropagation training [64]. It filters and extracts features based on a learning process that consists mainly of three types of layers: convolutional, pooling, and fully connected layers [65]. The convolutional layer learns and extracts relevant features. The pooling layers perform feature reduction. The fully-connected layer links the features with the predicted target label class that is illustrated subsequently in the output layer [66]. CNNs can extract relevant features in an unsupervised manner [61].

RNNs are widely used in speech processing and NLP. RNNs are a class of artificial neural networks that allows storing previous sequences of outputs in hidden states that can influence the decision-making process. RNN can learn from the embedded sentence structures to decipher context-specific meaning. RNNs can save prior input in hidden layers, possessing a type of short-term memory [46]. One of the drawbacks of RNNs is their inability to deal with distant dependencies and their sensitivity to gradient explosion or decay [61]. New data inputs can cause the network to undervalue prior input. A solution for this decay in sensitivity provides the LSTM model that has memory blocks that can save prior states and interact with each other [67].

BERT is a model for machines to understand the meaning of ambiguous language in texts. It is a model developed by an expert team of Google and presented by Devlin et al. [68]. The basis of the model are the transformers, according to Vaswani et al. [69]. This enables BERT to understand context and ambiguity by processing a given word in the context of all other words in the sentence rather than being processed individually. In this context, bidirectional means that BERT can read text input in both directions simultaneously rather than sequentially, unlike other language models. The transformers also bring this capability to the model. Bidirectional learning makes it possible to train with a larger amount of data than with RNNs and CNNs, which require a sequence of data. Pre-training is done using Masked Language Models (MLMs) and Next Sentence Prediction (NSP). The training corpus that is used is BooksCorpus with 800M words and the English Wikipedia with 2,500M words. Through the pre-training methods, BERT understands the language as it is spoken by predicting a masked word by its context. Because of unlabeled learning, it continues learning as it operates and uses the pre-training as a base layer. In fine-tuning, BERT can be adapted to a specific field through supervised learning by pre-training it with task- specific inputs and outputs [68].

IV. RESULTS

The result section summarizes the key findings of our literature review and is structured as follows: In the first part, we review the included papers from a behavioral and social science perspective in terms of the applicability of sentiment analysis in the context of COVID-19. In the second part, we analyze the included studies from a technical perspective, focusing on the classification techniques.

In total, 40 papers were included in this literature review. As illustrated in Fig. 4, 38 out of 40 of the observed studies used English tweets as a basis for analysis. However, it is important to note that several papers used multiple datasets, including non-English datasets. Therefore, we observed 47 different datasets, not counting standard sentiment libraries. For instance, Behl et al. [70] used two datasets from Nepal and Italy for training their classifiers and one global self-extracted COVID-19 dataset for evaluation. According to our classification approach, this paper would denote three categories: Nepal, Italy, and Global. The category "not explicitly specified" contains all papers in which the country of origin of the dataset was not identifiable due to missing specifications. Besides English datasets, two papers used Spanish tweets, and one paper respectively used tweets in Portuguese, Arabic, French, Persian, and Indonesian. Fig. 5 describes the geographical location of the datasets. The majority of the included papers use datasets from the beginning of the pandemic, as illustrated in Fig. 8.

A. COMBINING SENTIMENT ANALYSIS WITH INSIGHTS FROM SOCIAL AND BEHAVIORAL SCIENCE

In the fight to contain the spread of COVID-19, governments worldwide have imposed preventive measures such as physical distancing rules. Since human beings are, by nature, social species, these measures come with high attached costs: The disruption of everyday habits and relationships has prompted feelings of anxiety, social isolation, and learning impairment [71]. The US Government has already announced to launch a strategy addressing the so-called "National Mental Health Crisis" in response to the toll COVID-19 takes on citizens [72]. Adapting to governmental preventive measures requires a substantial shift in human behavior. In order to reduce psychological distress, behavioral and social science can give best practice approaches on how to align human behavior with the advice from health experts [1]. Hence, social and behavioral sciences can provide valuable insights for managing the pandemic and its impact [73]. Previous research has identified six social and behavioral research topics that are relevant in the context of pandemics: Threat perception, leadership, science communication, social context, stress and coping, and individual and collective interests [1].

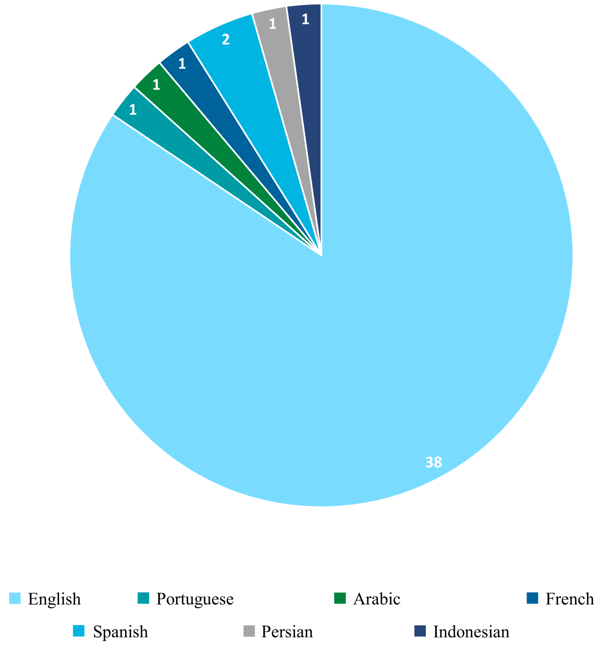


FIGURE 4. Overview of the included studies' Twitter datasets sorted by language.

This paper reviews possible sentiment analysis applications from a social and behavioral science lens. In this vein, we assign each of the 40 included papers to the six categories identified by van Bavel et al. [1] on the grounds of their research contribution or area of analysis. Van Bavel et al. [1] identified topics that are, from a social and behavioral perspective, important to contain the pandemic. We summarize the classification in Fig. 6. In each section, we briefly describe how behavioral and social science knowledge can advise on managing the pandemic. However, our main focus is to illustrate how sentiment analysis can be applied to collect relevant data to gain insights into the thinking, feeling, and acting of the public. Categories and subcategories without assignment will not be addressed and will be included in the discussion section.

Our literature review aims to provide decision-makers and public health experts with guidance on the application of sentiment analysis in the context of COVID-19. By using the classification of van Bavel et al. [1] as a point of reference, decision-makers are provided with examples of how sentiment analysis can be applied in the respective categories that can serve as a data basis for decision-making. The better the data foundations and the better the understanding of the public's behavior, the better decisions governments and health organizations can make to reduce the disruptive effects of prevention measures.

1) THREAT PERCEPTION

People respond in different ways to the imminent threat of the pandemic to human life. On a massive scale, the COVID-19 pandemic has triggered anxiety, stress, and depression across large swathes of the population [74]. In this sub-chapter, we will have a closer look at people's perceptions and reactions to threats and their emotional responses. Besides, we will examine the consequences of fear, which might lead to prejudice and discrimination against others, and panicking behavior [1]. For our analysis framework, we combined the first two subcategories, threat and emotion and risk perception, into one cluster "emotions". A natural human response is to switch into a defensive mode when facing a threat. On the one hand, fear can be a driver of efficiency when the threat is perceived surmountable; however, when fear outgrows one's coping mechanism and capabilities, an even stronger defensive reaction is triggered [1]. Eventually, people can develop an "optimism bias" when underestimating the likelihood of catching the virus [75]. Sentiment analysis allows investigating people's sentiments and emotions from texts. For example, three studies [76], [77], [78] solely focused on extracting the users' polarity from COVID-19 datasets with the aim of building more accurate classification models. All three studies used Twitter datasets extracted in 2020, the year of the worldwide outbreak of COVID-19. Alhashmi et al. [79] investigate COVID-19 as a current example of a critical event. A prevailing negative sentiment can cause the information to be more likely perceived as unfavorable and prompt negative emotions. Thus, people make decisions based on a rather negative information base [1]. A prime field of application for sentiment analysis is emotion detection, which was often applied during the pandemic [80], [81], [82], [83]. Moreover, Chourdrie et al. [84] conducted a study analyzing emotions during various points in time in different countries. Besides, sentiment analysis can be used to predict the number of infections, recoveries, and the death toll, as shown by a study by Mittal and Aggarwal [85]. Likewise, Sing et al. [86] found a correlation between negative tweets and, respectively, the number of global cumulative infections, global cumulative deaths, and cumulative recoveries (in China).

During the pandemic, panic buying has become an apparent coping mechanism of people when exposed to a threat [87]. When people face danger, a common response is to react with panic and act egocentrically to protect themselves - which can be to the detriment of other members of society [1]. Prentice et al. [88] investigated the connection between government measures and panic buying. The study's focus was to learn more about the timing effect between the imposed measure and the corresponding reaction. Moreover, sentiment analysis can be applied for disaster relief, as proposed by Behl et al. [70]. By analyzing tweets posted during a disaster, authorities can assess the needs and the availability of emergency supplies. They pre-trained their model on Twitter datasets from Nepalese and Italian earthquakes and tested it on 70,000 COVID-19-related tweets.

As can be seen in the examples above, sentiment analysis can be used to extract sentiment polarity and emotions from tweets to better understand the underlying feelings and emotions of the population. Moreover, sentiment analysis can be used to become aware of prevailing fears in the public that can

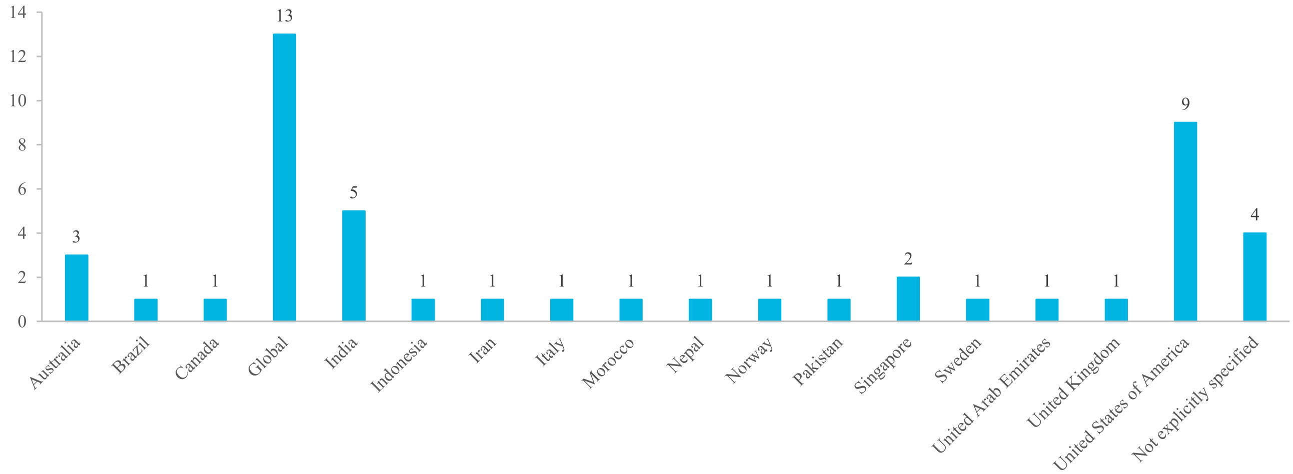


FIGURE 5. **Number of the included studies' Twitter datasets sorted by geographical location.**

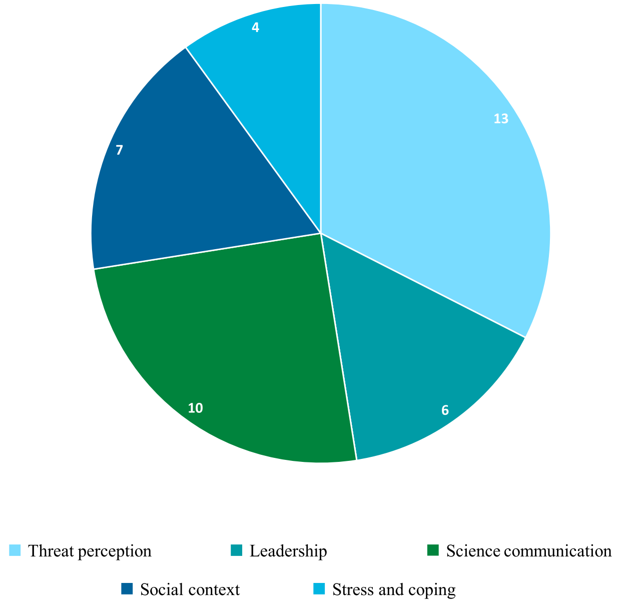


FIGURE 6. **Distribution of papers in terms of the categories by van Bavel [1].**

be mitigated through adjusted communication. Apart from that, specific public reactions like "panic buying" can also be identified with sentiment analysis. This can help public health authorities analyze triggers and enable them to take necessary measures to prevent over-reactions.

2) LEADERSHIP

Van Bavel et al. [1] distinguish three different types of leadership during pandemics: trust and compliance, identity leadership, and elevating the in-group without demeaning others. The subcategory trust and compliance entails all actions of people helping to build trust in health officials, health organizations, or governmental institutions. High levels of trust in institutions and government may lead to positive effects on the utilization of health services [89]. For instance, Gupta et al. [90] researched the Indian public's sentiments regarding the imposed lockdown in April 2020. They conclude that with sentiment analysis, they could examine the common public's reaction towards the imposition of lockdowns by the Indian government. Based on this examination, they argue that most Indians support and agree with the government's decision to introduce lockdowns to reduce new infections. Trust in institutions can be measured, for example, by comparing the public's sentiment regarding homegrown or imported vaccines. Nezhad and Deihimi [91] compared the perception of foreign vaccines, such as Pfizer/BioNTech, AstraZeneca/Oxford, Moderna, and Sinopharm to the Iranian vaccine COVIran Barekat by using sentiment analysis. They conclude that trust in foreign vaccines is higher than in the homegrown Iranian vaccine COVIran Barekat, which mirrors the trust in the country's institutions and government. Public opinion towards preventive measures or government can change in time. Therefore, studies such as the one by Miao et al. [92] that analyze tweets on Twitter in a certain city (New York) provide a status picture of public opinion. In the paper, they tried to process tweets in real-time to monitor public opinions and trust regarding intervention measures that the government implemented. Other authors tried to examine the public's sentiment when announcements about the development of vaccines took place [93]. Yu et al. [93] as well as Rahmanti et al. [94] highlight that social media data can be used to monitor the public's reaction toward COVID-19 events such as restrictions and other government measures. Goel and Sharma [95] specifically looked at highly influential people ("leaders") and used text analysis to cluster the tweets of this influential group. The social leaders mostly discussed popular public concerns such as disease symptoms, vaccination, disease countermeasures or hygiene, and disease transmission channels. The social leaders were clustered in four categories, namely people from academics, news, health, and politics. Interestingly, a lot of widespread concerns that the influential people were discussing were examples that van Bavel et al. [1] expected in their paper.

Against the backdrop of COVID-19, sentiment analysis can be used to analyze the public's perception of leaders, governments, or public health agencies. It allows real-time assessments of specific topics, such as the public's opinion of vaccines or containment measures. This knowledge can be used as input to evaluate possible consequences, and looming resistance [96].

3) SCIENCE COMMUNICATION

A study published two years before the outbreak of COVID-19 already pointed out that misinformation on social media is one of the biggest public-health threats in future pandemics [97]. Writing the year 2022, current research reports that worldwide, a certain percentage of the population does not trust scientific-proven information about COVID-19 - due to "fake news" on social media platforms. This development is perilous since there is a correlation between susceptibility to misinformation, vaccine hesitation, and the willingness to comply with preventive measures [98]. In this context, van Bavel [1] reveals strategies to distinguish scientific from misleading information and how to counter them effectively. Hence, in this subsection, we present papers that offer solutions to deal with misinformation. The categories "conspiracy theories" and "fake news and misinformation" are grouped due to their similarity in the investigated papers. Misinformation needs to be identified in the first place before counteractions can be taken. Madani et al. [99] provide a sentiment analysis approach to filter out COVID-19 epidemic fake news in Moroccan Corona-related tweets. Another study proposes a solution that enables the detection of informative tweets. In turn, the method would allow the restriction of irrelevant information and avoid the spread of negative sentiments [100].

Effective science communication is especially vital to stimulate the uptake of COVID-19 vaccines. The papers included in the following "persuasion" section reveal how sentiment analysis could help extract citizens' concerns regarding COVID-19 vaccines that might support health- authorities or decision-makers to design an effective communication strategy. Sv et al. [101], for example, used topic modeling to identify the concerns Indian citizens raise about the side effects of COVID-19 vaccines. Among the most reported fears were a lack of efficiency in the workplace, fear of death, fear of long-term effects, fear of blood clot, and the efficiency of the vaccine. The same authors conducted another research study in which they examined the general attitude towards COVID-19 vaccines in India. Similar research was conducted by Liu and Liu [14], who studied as well the public attitude towards COVID-19 vaccines. The study shows that the public sentiment and the number of tweets significantly increased after BioNTech/Pfizer announced that the first COVID-19 vaccine had reached an effectiveness of 90%, and then slowly declined until the end of December 2020. Cotfas et al. [102] monitored the public's stance on COVID-19 vaccines in the UK during the first month after the announcement of the first effective vaccine. In particular, they matched the respective tweets with major events reported in the media. They found out that the majority of tweets have a neutral stance, and the number of tweets in favor of vaccines exceeds the ones against. An interesting finding of the study was that the peak of against tweets was recorded on the day of the BioNTech/Pfizer COVID-19 vaccine authorization. Yousefinaghani et al. [12] included additional dimensions such as their progression by time, geographical distribution, main themes, keywords, posts engagement metrics, and account characteristics in their analysis. In total, 4,522,652 English tweets from 7 January 2020 until 3 January 2021 were collected to identify and compare vaccine sentiments. The negative sentiments focused mainly on concerns regarding vaccine development, doubts about vaccine safety, or reactions to governments, political figures, and manufacturers. Two authors closely examined popular tweets that were retweeted. One study provides further insights into the design of a system that determines the popularity of tweets and could be used as a tool to increase the amount of retweets [103], and the other one examined the sentiment polarity of popular tweets (tweets that have been retweeted at least 1000 times) [83]. Finding relevant posts by health practitioners depends to a large extent on the chosen hashtags. By analyzing more than 6,9 million tweets that contained at least one hashtag, Petersen and Gerken [104] noticed that only 1,192 hashtags were used more than 1,000 times, resulting in 13 different themes. Apart from the message and its source, the timing of the message could be essential, for example, to determine when to launch a campaign. In this vein, Satu et al. aimed at identifying frequently occurring topics and analyzed their sentiment. The authors conclude that these findings could be helpful in developing strategies that integrate human behavior. In addition, a study by Lyu et al. [105] explored the different characteristics of people that tweet about COVID-19 vaccines in the USA. The authors collected 1,874,468 English tweets from 28 September 2020 until 4 November 2020. The study shows that women are more likely to hold hesitant opinions on vaccination than men. Moreover, older people tend to be pro-vaccine. The lower-income group and religious people are more likely to hold polarized opinions on the vaccine debate. The political diversion (measured by followers of political party accounts or politicians) indicates a divided opinion about the potential COVID-19 vaccines in the USA.

In summary, sentiment analysis provides an effective tool to identify issues of public concern. In science communication, it can be used to detect misinformation on social media platforms. Furthermore, understanding the public's concerns can serve as a basis for designing target-oriented communication strategies.

4) SOCIAL CONTEXT

The social and cultural environment around us influences the process and velocity of changing and adapting our behavior.

This category can be subdivided into four smaller subsections, namely social norms, social inequality, culture, and polarization [1]. Whereas the paper by van Bavel et al. [1] paper gives further advice on how to identify risk factors and plan intervention measures, we point out the sentiment analysis tools that focus on the extraction of social context issues. Social norms and the striving to comply and adhere to these norms influences human behavior [1]. Nonetheless, people's perceptions are not universal and can be corrected by positive, reinforcing messages that promote information about what the majority of people are doing [1]. Imran et al. [106] conducted a study at the beginning of the corona pandemic, in which they examined the cultural differences between neighboring countries. They extracted 560,286 English tweets from February 2020 to April 2020 from Pakistan, India, Norway, Sweden, the USA, and Canada. Likewise, Garcia and Berton [107] examined the differences between countries using topic identification and sentiment analysis on Portuguese and English datasets from Twitter.

Sentiments can vary not only on a country level but also on a city level. This can be illustrated by a study by Yao et al. [108]. They investigated how public sentiments evolved in New York, London, Los Angeles, Chicago, Washington, Seattle, Boston, Singapore, and Rome and found out that even in the same country, the sentiment in big cities is not identical.

The impact of COVID-19 on individuals depends on a variety of socioeconomic factors. Members of marginalized communities or ethnic minorities show a higher vulnerability to being negatively impacted by the virus and might be more susceptible to public health information [1]. On a country-based level, Rahman et al. [109] explored the socioeconomic factors associated with positive and negative sentiments of US-American citizens about reopening the economy in the USA in the midst of the COVID-19 pandemic. The results show that factors such as "living in the western regions of the US", "working-class", "gross household rents", and "low-income" are positively associated with reopening the economy. Whereas factors such as "average family size" and "household income" are negatively associated with the reopening sentiment. These findings that across the US, reactions are not uniformly and influenced by the socio-economical situation are supported by a study by Surano et al. [110]. Political factors and the adherence to a certain party can impact the individuals' perception of COVID-19. A study conducted by Caliskan [111], who investigated the awareness of COVID-19 in Ohio, founds staggering differences between supporters of the democrats and supporters of republicans.

The results presented in this section indicate that sentiment analysis can be applied to screen cross-cultural differences and to extract information on individual societal groups in terms of interests, race, political orientation, or other factors. This allows health-authorities and decision-makers to tailor their communication strategy, taking into account the social context.

5) STRESS AND COPING

The psychological consequences of lockdown and isolation can affect households that have been spared so far by the virus, namely in terms of stress, anxiety, and economic difficulties.

We group the subcategories of social isolation and relationships as well as healthy mind-sets for our analysis. It is a human instinct to connect with others for emotion regulation, stress management, and getting through difficult times unscathed. Isolation can have a negative impact not only on mental health but also on cardiovascular and immunological health, especially for extroverted people [112]. Isolation is not the same as loneliness, but it can exacerbate and promote it. Through media communication, it must be made clear that social connection is possible even when physically separated, for example, through online interactions. Especially for the older generation, who have a higher risk for loneliness, especially through isolation [113]. Also to be considered here is that attitudes and situational assessments influence stress outcomes. Thus, stress reactions can lead to positive emotions through positive thinking and thus prevent psychological as well as physiological problems [114].

Kabir and Madria [115] used one of the largest self-extracted Twitter datasets that contained a total of 56,014,158 English tweets. They find that negative emotions increased during the course of the pandemic. Besides, the amount of COVID-19 cases in a specific state correlates with the detected negative sentiment. Matching these findings, Kaur et al. [116] found out that COVID-19 communication over social media has both a positive and negative impact on the lives of people. A study focusing on India studied the general causes of stress, anxiety, and trauma during COVID-19 [117]. Their findings indicate that death and lockdown have the strongest impact on Indian people's mental health. Another study investigated the sentiment and mental health of people living in Australia. Zhou et al. [118] analyzed Australian tweets during the pandemic in terms of their sentiment. The authors collected the largest dataset with 94,707,726 English tweets from 183,104 Twitter users that live according to their Twitter location in New South Wales, Australia. The results show that policies and epidemic events affect people's emotions differently during various stages. For example, in times of increasing COVID-19 infections and introduced lockdowns, the general sentiment shifted significantly towards the negative.

All four studies extract emotional health in a specific situation and the impact of government communications on the public. Sentiment analysis of Twitter can provide important insights into the current mental health status of the population. This information can be used to react in real-time to looming problems that the public health care system might

TABLE 3. Categorization of the included studies according to the relevant social and behavioral science topics according to van Bavel et al. [1].

Ref.	Author	Subcategory
Threat perception		
[79]	Alhashmi et al.	Emotion
[76]	Alsayat	
[83]	Chakraborty et al.	
[84]	Choudri et al.	
[82]	Kaur et al.	
[85]	Mittal and Aggarwal	
[80]	Morshed et al.	
[77]	Ramya et al.	
[81]	Ridwhan and Hargreaves	
[86]	Singh et al.	
[78]	Yigitcanlar et al.	
[70]	Behl et al.	Disaster and panic
[88]	Prentice et al.	
Leadership		
[95]	Goel and Sharma	Identity leadership
[90]	Gupta et al.	Trust and Compliance
[92]	Miao et al.	
[91]	Nezhad and Deihimi	
[94]	Rahmanti et al.	
[93]	Yu et al.	
Science Communication		
[102]	Cotfas	Misinformation
[14]	Liu and Liu	
[105]	Lyu et al.	
[99]	Madani et al.	
[103]	Mahdikhani	
[100]	Malla and P.J.A.	
[104]	Petersen and Gerken	
[120]	Satu et al.	
[121]	Sv et al.	
[101]	Sv et al.	
[12]	Yousefinaghani et al.	
Social Context		
[107]	Garcia and Berton	Culture
[106]	Imran et al.	
[108]	Yao et al.	
[111]	Caliskan	Polarization
[110]	Surano et al.	
[109]	Rahman et al.	Social norms
Stress and coping		
[115]	Kabir and Madria	Mental health
[116]	Kaur et al.	
[117]	Sv et al.	
[118]	Zhou et al.	

have to deal with in the upcoming future [119]. We summarize all papers according to their category and subcategory in Table 3.

B. TECHNIQUES FOR SENTIMENT ANALYSIS

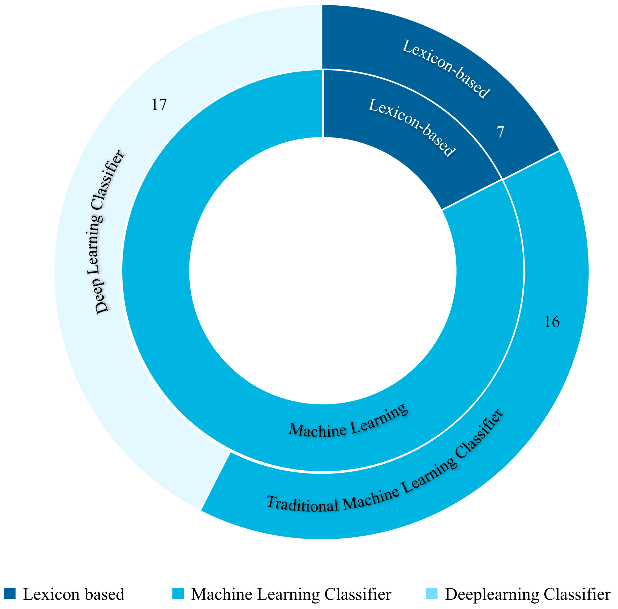


FIGURE 7. Categories of classification algorithms used in the literature review.

The previous chapter shed light on sentiment analysis from a social and behavioral science lens. Sentiment analysis tackles the problem of extracting and analyzing large quantities of information that provide insights into the public's behavior. We illustrated, by means of reviewing practical examples from literature, how sentiment analysis can support the decision-making in the social and behavioral categories identified by van Bavel et al. [1]. Decision-making in the context of COVID-19 can be improved by having a sound understanding of the publics' feelings and emotions. However, the classification algorithms proposed for sentiment analysis vary across the included papers in terms of their performance. Hence, it has to be ensured that sentiment classifiers provide reliable and accurate results. Therefore, the reported accuracy of a classifier further indicates its overall performance and reliability. In the following three sub-chapters, we review the included papers from a technical perspective. In particular, the studies are categorized according to the classification technique they use.

1) OVERVIEW OF CLASSIFICATION TECHNIQUES

In the reviewed literature, mainly three classification approaches have been used: lexicon-based, ML-based and DL-based methods. An overview of the classification algorithms used in the included studies and their respectively assigned categories in the framework by van Bavel et al. [1] can be found in Table 4. Out of 40 papers included in this literature review, 7 papers used a lexicon-based, 16 a traditional ML-based, and 17 a DL-based classification method. In case various algorithms were analyzed in one study, only the best-performing algorithm was selected (see Fig. 7).

A common characteristic of the lexicon-based papers was that all used the open-source Valence Aware Dictionary and sEntiment Reasoner (VADER), a rule-based lexicon sentiment analysis tool for analyzing sentiments. These studies have in common that mostly polarity detection was

TABLE 4. Overview of sentiment classification algorithms used in included studies.

Ref.	General approach	Specific classification algorithm used
Threat perception		
[85]	Lexicon based	VADER
[88]	Lexicon based	VADER
[79]	Traditional ML	MultinominalNB, RF, SVM, Bi-LSTM, CNN, Bayes Factor Tree Augmented Naïve Bayes technique (BFTAN)
[83]	Traditional ML	32 different ML-algorithms (MultinominalNB, LinearSVC, RF, . . .)
[82]	Traditional ML	IBM Watson Tone Analyzer
[77]	Traditional ML	NB
[81]	Traditional ML	Sentiment analysis: VADER, Emotion detection: Pre-trained RNN
[76]	DL	Ensemble Method (LSTM with FastText + BERT + IBM Watson + Microsoft Sentiment Analysis API)
[70]	DL	MLP-W
[84]	DL	RoBERTa, BiLSTM, BERT
[80]	DL	DL model; not explicitly specified
[86]	DL	BERT
[78]	DL	WEKA
Leadership		
[93]	Lexicon based	VADER
[90]	Traditional ML	MultinomialNB, BernoulliNB, LR, LinearSVC, AdaBoostClassifier, RidgeClassifier, PassiveAggressiveClassifier, Perceptron
[95]	DL	SVC, RF, NN, BERT
[92]	DL	LSTM + GloVe Distillation
[91]	DL	CNN-LSTM model
[94]	Traditional ML	NB, Maximum Entropy
Science communication		
[14]	Lexicon based	VADER
[12]	Lexicon based	VADER
[99]	Traditional ML	Sentiment analysis: TextBlob2, Fake detection: LR, DT, RF, NB, Gradient Boosting, SVM, MLP
[103]	Traditional ML	RF, SGD, LR, EVC
[121]	Traditional ML	TextBlob
[101]	Traditional ML	TextBlob
[102]	DL	MNB, RF, SVM, Bi-LSTM, CNN, BERT
[105]	DL	XLNet, VADER, LDA
[100]	DL	MVEDL (Mayority Voting technique-based Ensemble Deep Learning model) based on RoBERTa, BERTweet, CT-BERT
[104]	DL	SpaCy library
[120]	DL	TClustVID (Ensemble method with clustering)
Social context		
[110]	Lexicon based	VADER
[107]	Traditional ML	LR, RF, Linear SVM
[109]	Traditional ML	LR
[108]	Traditional ML	NB
[111]	DL	RNN, CNN
[106]	DL	DNN (Baseline), LSTM + FastText, LSTM + GloVe, LSTM + GloVe Twitter, LSTM without pre-trained embedding, BERT, BiLSTM, GRU
Stress and coping		
[118]	Lexicon based	VADER
[116]	Traditional ML	Hybrid Heterogeneous SVM, RNN, Linear SVM
[117]	Traditional ML	TextBlob
[115]	DL	SVM, NTUA-SLP, BilSTM

Abbreviations included in the table:

Support Vector Classifier (SVC), Random Forest (RF), Logistic Regression (LR), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Random Neural Network (NN), Random Neural Network and Bidirectional Encoder Representations from Transformers (BERT), Deep Neural Network (DNN), Long-short term memory (LSTM), Multinominal Naïve Bayes (MNB), Multi-Layer Perceptron (MLP), Bidirectional Long Short-Term Memory (Bi-LSTM), Bayes Factor Tree Augmented Naïve Bayes technique (BFTAN), Tree Augmented Naïve Bayes (TAN), MVEDL based on RoBERTa, BERTweet and CT-BERT, Decision Tree (DT), Waikato Environment for Knowledge Analysis (WEKA), Global Vectors for Word Representation (GloVe), Ensemble Voting Classifier (EVC), NTUA-SLP is a DL model and the winner of the SemEval-2018 Task1 competition.

performed. Polarity detection is often equated with sentiment analysis [11]; however, it is only one specific aspect.

Polarity detection deals with classifying a text's sentiment into positive, negative, or neutral [23]. VADER can be used to explore the sentiment of a variety of topics [88], [93]. Lexicon-based approaches provide an uncomplicated way of studying novel topics since they do not require labeled data. Liu and Liu [14], for example, evaluated COVID-19 vaccine-related tweets regarding their polarity score. Across various countries, different trends were observed: The polarity score was the lowest for Brazil and the highest for the United Arab Emirates. A similar study by Yousefinaghani et al. [12] used VADER and a keyword function from the Gensim library to categorize tweets regarding their sentiment and vaccination-specific topic. Interestingly, they found out that negative tweets are mainly published by bots or political activists. This could help authorities take necessary countermeasures to fight vaccine hesitancy and design a target-oriented communication strategy. Hence, Zhou et al. [118] studied the sentiment in tweets in the Australian region of New South Wales during the pandemic. Mittal et al. [85] analyzed the correlation between public sentiment and factors such as global infections, global deaths or recoveries, and Surano et al. [110] used sentiment polarity as a predictor for other socioeconomic factors.

For sentiment analysis, ML classification techniques are prominently used [23]. This is because lexicon-based methods come with a significant drawback: they rely exclusively on annotated lexicons of words and do not consider context-specific information or domain-specific information. For sentiment analysis on Twitter, where topics emerge and change frequently, pre-recorded dictionaries must be updated constantly. Moreover, ML-based approaches can include context-specific information for classification [23]. Due to this reason, two sub-chapters will explicitly deal with traditional ML and DL classification methods used in the light of sentiment analyses of COVID-19 Twitter data. Besides, four studies in the category of traditional ML used the Python library TextBlob [99], [101], [117], [121]. This library is frequently used for NLP tasks, such as phrase extraction, part-of-speech tagging, tokenization, sentiment analysis, classification, and frequency of words and phrases. However, the focus of this research was mainly on the use case of sentiment analysis, and no performance metrics, i.e., accuracy or F1 score, were reported in detail. This phenomenon could be observed in another study that employed the IBM WatsonAnalyzer, a NLP tool that can detect emotions and language tones [82]. Since we compare the performance of ML classifier to showcase best practices, studies without or with insufficiently reported accuracy data were excluded from our detailed analysis in the following two sub-chapters. Therefore, in addition to the previously mentioned papers, we excluded the papers by [80], [81], [91], [104], [108], [116], and [107] on the grounds of missing performance metrics and inaccessibility of supplementary performance data [108].

In the next sub-chapter, ML classifiers with reported performance indicators are studied in greater detail. As stated before, we only included the accuracy of the best-performing classification algorithm. Nonetheless, comparing the accuracy of sentiment analysis papers with each other poses some difficulties. Each study has its unique characteristics in terms of its dataset, which includes variations in, among others, the extraction time and duration, the hashtags and keywords used for data extraction, in the selected location, the number of observations, or preprocessing steps. Besides, the different classification algorithms, feature extraction methods, and fine-tuning of various parameters, such as the number of folds in cross-validations, make comparisons between individual studies challenging. A small change in these fundamental parameters can significantly impact the overall performance of a classifier. In our literature review, out of 33 studies that used ML classifiers, 15 applied traditional ML classifiers, and 18 applied DL classifiers. Since we only include studies with reported performance metrics, the remaining 6 studies using traditional ML classifiers and 13 studies using DL classifiers are examined in detail.

To recap, lexicon-based methods are beneficial when no labeled data are available and if the dictionary is adjusted to the specific domain. ML approaches require a significant amount of labeled data to train its classification algorithm but can include context-specific information and language particularities, which are commonplace in Twitter messages. Therefore, for sentiment analysis on Twitter, ML classifiers are mostly used [23].

2) TRADITIONAL MACHINE LEARNING CLASSIFIERS

This sub-chapter reviews ML classifiers used in the included papers for conducting sentiment analyses. In total, six papers have reported the accuracy of their traditional ML classifiers. The majority of studies evaluated various ML algorithms. Only two studies used a single classifier [77], [109]. Most frequently used were RF [95], probabilistic classifiers such as NB [77], or other supervised models such as SVC [90] or logistic regression (LR) [83], [109]. Two studies employed an ensemble classifier approach (in this paper, we consider RFs as not being an ensemble classifier). The advantage of combining different classifiers is that the generalization errors of each individual classifier are less likely to produce an error in the collective decision [122]. Mahdikhani et al. [103] created an Ensemble Voting Classifier (EVC) that consists of a RF, Stochastic Gradient Descent (SGD), and a LR [103]. Alhashmi et al. combined a RF with a NB classifier and created a so-called Bayes Factor Tree Augmented Naïve Bayes (BFTAN) classifier [79]. The ML classifiers have been evaluated in several studies in combination with different feature extraction methods. Most popular were the TF-IDF [95], [123], n-grams [77], [90], [107], [123] and Word2Vec [79].

Goel and Sharma [95] investigated world leaders and their expressed concerns during the pandemic. Their model predicts a tweets' probability to fall into one of the four categories: "politics", "health", "research", or "news". Their evaluations were based on a dataset of 42,468 tweets. Besides, they preprocessed more than 30 million tweets with TF-IDF embedding to extract the public concerns and with a LDA and Gibbs sampling method to detect emotions. They clustered the emotions into anticipation, anger, trust, surprise, sadness, joy, fear, and disgust and used them as standardized input features for the classification algorithm. The authors compared different classification models regarding the Area Under the Receiver Operating Characteristic Curve (AUC ROC). They tested SVCs, RFs, a Random Neural Network (a combination of a CNN and LSTM), and BERT. The most effective model was the RF classifier with an AUC ROC of 96% to select the correct cluster. The classifier with the best accuracy achieves a model from Mahdikhani [103] that develops a model to predict the retweetability of tweets. The classification procedure is set up similarly to the one by Goel and Sharma [95]. In a first step, frequent topics in the Tweet dataset were identified using the LDA algorithm. In a second step, their emotional intensity was measured and categorized into fear, anger, joy, and sadness using the CrystalFeel algorithm. Topic modeling was performed and evaluated using the feature extraction methods LDA, LDA plus TF-IDF, BOW by TF-IDF, Doc2Vec, and Doc2Vec plus TF-IDF. These were used as additional content features. An ensemble model was used for the classification task that integrates a RF, SGD, and LR algorithm. The study shows that the LDA plus TF-IDF vectorizer combined with the Ensemble Voting Classifier can reach the highest accuracy of 95.04% and a F1-score of 95%. The third-best classifier with an accuracy of more 92.49% was the NB algorithm employed by Ramya et al. [77]. The authors analyzed the sentiment polarity of tweets using the NB method combined with n-grams as features. Depending on the length of a tweet, the model's accuracy varies between 92.49% for short messages with less than 70 characters and 60.56% for messages having between 70 and 150 characters. Interesting to see is that the accuracy of the same classifier varies with the Tweet length. Apart from the NB classifier, they could find out that a LR method performs the classification task with an accuracy of 74%.

The study by Gupta et al. [90] differs from the previous studies already in the data annotation process. The authors use TextBlob and VADER in combination to label their extracted dataset. Only when both algorithms assign the same polarity to a tweet it is included in the sample. Out of 12,741 extracted tweets with the hashtag #Indianlockdown, after the consolidation, only about half of the initial dataset (7,284 tweets) remained. Eight different ML classifiers were trained and tested, namely Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes, LR, LinearSVC, AdaBoostClassifier, RidgeClassifier, PassiveAggressiveClassifier, and Perceptron. 80% of the data was used for training and 20% for testing. They additionally performed a ten-fold cross-validation for each variation and showed that the LinearSVC with unigram features achieves the highest accuracy of 84.4%. The same model achieves a lower accuracy when features are extracted as bigrams (81.2%) or trigrams (78.3%). Interestingly, in the labeling step, TextBlob and VADER assigned different labels to nearly half of the sample – which were in turn discarded and not included in the sample. Similar to Mahdikhani [103], Alhashmi et al. [79] used an ensemble model to evaluate the sentiment during critical events such as the current pandemic. Their proposed model, a combination of SVM with Naïve Bayes, a so-called Bayes Tree Augmented Naïve Bayes technique (BFTAN), outperforms all other benchmark models. The author evaluated their model against standard models such as Tree Augmented Naïve Bayes, NB, SVM, and RF. With Word2Vec as a feature extraction method, the model achieves an accuracy of 82.8%.

Chakraborty et al. [83] have evaluated more than 32 traditional ML classifiers on two different samples. Among the evaluated classifiers were MNB, LinearSVC, AdaBoostClassifier, and LR. In contrast to the study of Gupta et al. [90] in which the LinearSVC achieved the highest accuracy, Chakraborty et al. found out that the LR classifiers outperformed all other 32 evaluated ML classifiers on both datasets, despite the difference in sample size (dataset one contains approximately 23,000, and dataset two contains 226,000 retweets) and despite their different observation periods during the pandemic. The accuracy of the LR classifier on the smaller dataset is 81% (trigrams under TI-IDF) and on the larger dataset 75% (with bigrams under TI-IDF).

Another study that used a LR for predicting class labels is the one of Rahman et al. [109]. It undertakes an investigation into the driving factors that US inhabitants associate with positive or negative sentiments in the context of the reopening of the economy. The authors used the R packages Syuzhet and sentimentr on a dataset of 2,507 tweets that have been downloaded from Twitter. After determining the sentiment polarity, a LR was used to find underlying factors influencing the associated sentiment. They used BoW, Document term matrix (DTM), POS, Dependency parsing (DP), and N-gram as exploration techniques. The model has an accuracy of 56.18%. A summary of the traditional ML classification algorithms is provided in Table 5.

ML-based classifiers for performing sentiment analyses show in the context of COVID-19 a considerably high accuracy of more than 80% - with one exception [109]. Since ML-based classification models require labeled data, various methods for acquiring data can be observed: either a pre-labeled, open-source dataset for training the classifier was used [86], [106], the dataset was annotated by making use of a machine-driven annotation process, e.g., using TextBlob or VADER to annotate data [83], [90], [109], or parts of the dataset were manually labeled [70], [77], [124]. Since tweets contain unstructured data, preprocessing steps are necessary. Particularly for real-time sentiment analyses, DL models - which are analyzed in the following chapter - provide a certain edge over ML-based classification methods.

3) DEEP LEARNING CLASSIFIERS

Twelve in twenty studies that reported performance indicators for their classifiers used DL models. Six studies used a transformer-based model, such as BERT, RoBERTA or XLNet [76], [84], [86], [100], [102], [105]. Four studies used a LSTM model, either as a single classifier or as part of an ensemble classifier [76], [92], [106], [115]. The remaining studies have adopted a model that is either a combination of RNN and CNN [111], a Multi-Layer

TABLE 5. Summary of traditional machine learning classification approaches that have reported performance metrics.

Ref.	Traditional Machine Learning	Ensemble method	Best-performing feature extraction and classification method	Accuracy
[95]	x		TF-IDF + Random Forest	96% (AUC ROC)
[103]	x	x	Ensemble Voting Classifier (RF, SGD, LR)	95.04%
[77]	x		N-grams + Naïve Bayes	92.49% Short tweets (<70 characters), 60.56% Long tweets (<150 characters)
[90]	x		Unigram + LinearSVC	84.4%
[79]	x	x	Word2Vec + Bayes Factor Tree Augmented Naïve Bayes technique (BFTAN)	82.8%
[83]	x		Trigrams and TF-IDF + Logistic Regression	81.4%
[109]	x		BOW, DTM, POS, DP and n-grams + Logistic Regression	56.18%

Abbreviations included in the table:

Term Frequency-Inverse Document Frequency (TF-IDF), Support Vector Classifier (SVC), Random Forest (RF), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Bag-Of-Words (BOW), Document Term Matrix (DTM), Part-of-Speech (POS), and Dependency Parsing (DP)

Perceptron (MLP) [70], or combine different DL models in an ensemble classifier [120]. As illustrated in Table 6, the three best-performing classifiers have all used BERT. The second best to the fourth-best classifiers are all ensemble methods that can reach an overall accuracy of more than 90%.

The model with the highest accuracy was observed in the study by Singh et al. [86]. The BERT model was used for emotion detection. They used a self-selected dataset split into a global and an Indian dataset. Moreover, they split the data into test and training sets and determined the polarity of each Tweet using VADER and TextBlob - similar to the study by Gupta et al. [90]. BERT was fine-tuned with the test dataset and an open-source "Emotion dataset". The customized BERT model achieves an accuracy of 93.89% in extracting the tweets' emotions. Besides, it combines a LSTM model with FastText as a word embedding technique and words as trigrams. Moreover, it uses pre-built models such as BERT, IBM Watson, BERT, and Microsoft Sentiment Analysis API to evaluate the results of the LSTM model. The accuracy of the ensemble method is higher than the accuracy of the individual models. The overall accuracy of the ensemble model is 92.65%. The customized DL classifier achieves an accuracy of 90.25% on the Twitter dataset – higher than the one achieved with BERT alone. The authors found out that a neural network with 200 hidden layers produces the best results on the test dataset. In a similar setting, Malla et al. [100] applied as well an ensemble DL model on Twitter data to filter informative tweets based on their content. They propose an ensemble algorithm, a "Majority Voting technique-based Ensemble Deep Learning (MVEDL)" model that uses three DL algorithms, RoBERTa, BERTweet, and CT-BERT. The classifier was trained on a dataset comprising more than 10,000 labeled tweets. The ML algorithm reached an accuracy of 91.75% and F1-score of 91.14% in classifying data into informative and uninformative categories. Satu et al. [120] employed an ensemble method as well; however, they introduce a further clustering step prior to the classification. They propose a novel approach that includes, after pre-processing with GloVe, subdividing the labeled dataset into different clusters using a k-means algorithm. The breakdown into groups helps to improve the accuracy of the ML classifier models by more than five percentage points. The authors compared various ML and DL classifiers: Decision Trees, Gradient Boosting, K-Nearest Neighbor, LR, MLP, NB, RF, SVM, XGBoost, and LSTM. In the majority of clusters, the best performance was achieved with the LSTM model. The proposed algorithm TClustVID chooses the best-performing classifier of each cluster that possesses the highest accuracy, and thus the overall accuracy is optimized.

Kabir and Madria [115] propose a DL model, which results indicate solid robustness in terms of accounting for content-specific details of tweets. The authors employ a DL classifier to automatically predict ten specific emotions in a COVID-19 dataset. The authors evaluate different ML models: SVM-Unigrams, NTUA-SLP, and BiLSTM on a self-selected and manually labeled dataset, the SemEval2018-Task 1 dataset, and an emotion dataset. Their proposed multi-layered neuronal network is based on a bidirectional LSTM model that uses a pre-trained RoBERTa model. The model reaches an accuracy of 89.51%. One of the few studies that tested their classifiers on an unseen dataset was the one of Behl et al. [70]. They explored the usability of Twitter information in a disaster relief operation. Therefore, the authors trained five different DL classifiers on a publicly available dataset that comprises recorded tweets during an earthquake in Nepal and Italy. They evaluated the re-usability of their model on a COVID-19 dataset. The authors evaluated the different models' performance for correctly classifying the tweets into the categories "resource needs", "resource availability", and "others", which could support the allocation of aid during a natural disaster. The authors evaluated cutting-edge ML-classifiers such as LT with TF-IDF features (LR-TF), CNN without fine-tuning, CNN with fine-tuning, MLP with TF-IDF features, and MLP without TF-IDF features (MLP-W) but instead a Word2Vec embedding. The models were pre-trained and tested on two individual datasets as well as on a combined dataset. The comparative analysis of the model trained on the combined dataset showed that MLP-W outperformed all other classifiers in all metrics. Besides, the study was one of the few that tested both, ML and DL classifiers. Their analysis shows that DL classifiers predict correct labels with a higher accuracy than traditional ML classifiers. The model achieves an accuracy on the manually-labeled COVID-19 dataset of 83% and a F1-Score of 85%. A striking observation is that although all other evaluated models achieve a higher accuracy on the combined test training dataset, the MLP-model performed better on unseen data.

The study by Imran et al. [106] illustrates well that the selected features can have an impact on the performance of the classifier. The authors compared six different DL models in terms of their capability to predict polarity using the publicly accessible Sentiment140 dataset. The evaluated DL models were the following: deep neural network (DNN), LSTM with FastText, LSTM with GloVe, LSTM with GloVe with and without pre-trained embeddings on Twitter data. The LSTM model based on FastText achieves the highest accuracy (82.4%) and a F1-score of 82.4% for the task of detecting polarities. When tested on COVID-19 data, the model's accuracy decreases to 76%. Although, it has to be emphasized that the authors considered the tweets' emoticons as their actual labels – which might distort their findings. Besides polarity detection, the authors tested the performance of DL classifiers in predicting positive (joy and surprise) and negative emotions (sadness, fear, and anger). For emotion detection, the authors trained their algorithms on the publicly available Emotional tweets dataset. As seen in the previous evaluation, the LSTM model outperforms all other classifiers again. With an accuracy of 81.9% and respectively 69.9% positive and negative emotions were correctly predicted. However, GloVe as a feature extraction method shows the best performance this time. In both cases, the models are pre-trained on Twitter data. With the same analytical focus as Imran et al. [106], Choudrie et al. [84] create a DL model for text-based emotion analysis. RoBERTa, BERT, BiLSTM, and LSTM are tested as classifier models. The model is based on RoBERTa and fine-tuned with transfer-learning with the open-source ''Emotion in Text'' dataset by CrowdFlower. The accuracy of the model is 80.33%. Instead of using an open-source dataset, Coftas et al. [102] manually labeled 7,530 tweets, which equals approximately 1% of the whole dataset, to examine UK citizen's stance toward COVID-19 vaccination. They used popular ML and DL classification algorithms such as MNB, RF, DVM, Bidirectional-LSTM, and CNN. The best performance delivered the classifier BERT with an accuracy of 78.94%, followed by SVM (76.23%) and Bidirectional-LSTM (74.7%) with GloVe embedding.

GloVe embedding is frequently used with DL models. Also, in the following two studies, GloVe is pre-trained on Twitter data and used as a feature extraction method. In the study by Caliskan [111] DL algorithms are employed to extract awareness and emotions. For the classification of emotions, pre-trained GloVe vectors on Twitter data were used. Their DL model, which is based on a series of RNN and CNN algorithms, achieves an average accuracy of 71%. Miao et al. [92] analyzed tweets on Twitter from New York in terms of their public opinion. Due to the limited amount of labeled data, a data augmentation of training data was applied. The authors tested three representation models and three classification algorithms. They trained their classifier model with the existing datasets StanceData and Sentiment140, as well as with manually labeled data. A Synthetic Minority Oversampling Technique (SMOTE) was tested for SVM (with BOW) to account for the imbalances. The technique did not lead to improvements and was not used for the other models. Also, the labeled-train dataset resulted in the best results in combination with SVM. As a result, only the labeled train dataset was used to test the models. They could find out that a Deep-Learning Model, LSTM with GloVe – using 50 dimensions and distillation – significantly outperforms other tested classification models. LSTM-GloVe reached an accuracy of 66%. Striking findings can be observed in the study by Lyu et al. [124]. They collected a Twitter dataset that consists of more than 1,850,000 unique tweets linked to the COVID-19 vaccine and vaccine-related keywords. The focus of the study is on the stance detection of Twitter users with regard to the COVID-19 vaccination. The three researchers sampled 2,000 unique tweets from the dataset and manually, independently annotated the label for each Tweet. The respective four categories were ''irrelevant'', ''pro-vaccine'', ''vaccine-hesitant'', or ''anti-vaccine''. This data is used as training data for a XLNet model. However, when the model was validated on a novel, external dataset with 400 labeled tweets, it only had an accuracy of 63%.

In brief, ensemble methods that combine several DL classifiers have shown the most promising results. In contrast to ML models, the datasets used to train DL models are, on average larger. The included samples above have shown that DL classifiers can process context-specific information, which is exemplified by the study by Kabir and Madria [115]. However, in most studies, the training and performance evaluation has taken place on the same dataset. The study by [124] impressively demonstrated that also DL models face difficulties when exposed to unseen data.

V. DISCUSSION

Previous research has investigated sentiment analysis in fighting multiple infectious diseases from an application's perspective [24], from a health and well-being perspective [125], in the context of vaccine-hesitancy [26], or in terms of the approaches and ML techniques used for predicting disease outbreaks [86]. In this literature review, we build on the framework of van Bavel et al. [1] and combine the previously unrelated streams of social and behavioral science research with sentiment analysis. Experts highlight that the pandemic can only be contained by making substantial changes to our regular, everyday human behavior [73]. Therefore, we have examined how sentiment analysis can provide valuable insights and sound data basis for making decisions in human behavior-related areas that are vital to

TABLE 6. Summary of deep learning classification approaches that have reported performance metrics.

Ref.	Deep Learning	Ensemble method	Best-performing feature extraction and classification method	Accuracy
[86]	x		BERT	93.89%
[76]	x	x	Ensemble Classifier (LSTM + FastText, BERT, ...)	92.65%
[100]	x	x	MVEDL (RoBERTa, BERTweet, CT-BERT)	91.75%
[120]	x	x	TClustVID (Ensemble method w. DL-Models)	>90%
[115]	x		Bidirectional LSTM	89.51%
[70]	x		Word2Vec + Multi-Layer Perceptron (MLP) + Twitter	83%
[106]	x		Polarity detection: LSTM + FastText, Emotion detection: LSTM + GloVe (Twitter)	Polarity detection: 82.4% Emotion detection: 81.9% (positive emotions); 69.9% (negative emotions)
[84]	x		RoBERTa	80.33%
[102]	x		BERT	78.94%
[111]	x		GloVe + DL model (RNN, CNN)	71%
[92]	x		GloVe + LSTM	66%
[105]	x		XLNet	63%

Abbreviations included in the table:

Bidirectional Encoder Representations from Transformers (BERT), Long-short term memory (LSTM), MVEDL is based on RoBERTa, BERTweet and CT-BERT, TClustVID is an ensemble method with Deep Learning models, Global Vectors for Word Representation (GloVe), Deep Learning (DL), Recurrent Neural Network (RNN), and Convolutional Neural Networks (CNN)

containing the pandemic. Sentiment analysis is frequently applied in the medical setting and offers a rich opportunity for research [126]. In this section, we aim to highlight and discuss the best-performing ML classification algorithms for COVID-19-related Twitter data.

After carefully analyzing existing research in the domain of sentiment analysis of COVID-19-related Twitter data, our findings reveal promising prospects for sentiment analysis in the domain of COVID-19 Twitter data. From a technological point of view, our results indicate that ensemble classifier work particularly well for COVID-19 Twitter data. Previous research has already shown that an ensemble classifier, which combines the individual classifiers' predictions in a unique way, achieves higher accuracy than each individual base classifier [127]. Other scholars have found that ideally the classifier should be combined in a more heterogeneous fashion by including a more diverse set of classifiers, i.e. from ML to lexicon-based classifier, which can lead to performance improvements of more than 5% [127]. In our study, we were not able to approve this finding due to a lack of comparative data. Especially among the studies that used DL classifiers, three out of the four best-performing classifiers were ensemble classifiers. This finding exactly matches the observations by Zimbra et al. [33]. Other scholars conducting research in other, related fields of sentiment analysis also highlight the superior performance of ensemble classifiers [58], [128], [129]. This could be reasoned, among others, in their better handling of class imbalances [33], [130].

In general, among the classifiers with reported performance metrics, state-of-the-art DL classifiers were widely applied in the included studies and matched the current practice of the sentiment analysis research [59]. State-of-the-art DL techniques have shown high performance on COVID-19-related Twitter data [76], [86], [100]. DL methods have several advantages over more traditional approaches: In contrast to lexicon-based methods, which require updating their dictionary constantly to include new words or abbreviations, or traditional ML approaches, which require a time-consuming feature design, DL automates the feature learning process [131]. Our results indicate that classifiers such as BERT or ROBERTa, which include contextual embeddings, have shown to work well for COVID-19 Twitter data. This finding has been shown as well in related studies which applied sentiment analysis in disaster prediction. Moreover, classifiers were pre-trained on domain-specific lexica for fine-tuning purposes [86], [100]. We see a possibility for further studies to shed more light on the characteristics of the pre-training part. Lexicon-based and linguistic resources can be useful to overcome the challenges of correctly processing Twitter data [33].

However, the reported accuracies of the classification models must be questioned critically. Overall, high accuracy was reported and in some studies, it exceeded even an accuracy of 90%. This finding is in stark contrast to, for instance, a study by Zunic et al. [125] who investigated health and well-being with sentiment analysis and reported overall lower accuracies. Even among the included studies which reported performance metrics, not all studies compared various classifiers with each other to come up with the best model. Besides, each study uses a different dataset, different pre-processing steps, and evaluation procedures. Therefore, direct comparisons between the mentioned studies and respective best-performing classifiers are limited due to the different design approaches. Hence, there is no ''one-size-fits-all'' classifier solution that outperforms all other classifiers. Nonetheless, based on our findings, we can provide concrete recommendations for the technological implementation of sentiment analysis of COVID-19-related Twitter data. Besides, we raise awareness of common challenges that have to be considered when using sentiment analysis as an instrument for governmental policies.

A major drawback that we identified across the large majority of included studies is that a single dataset was used for training, testing, and performance assessment. Just a few studies evaluated their models on unseen data or on data that was extracted during different points in time. When tested on new, unseen data, the included studies showed a significant decrease in performance [70], [124]. There is a general tendency of studies that aimed at maximizing a classifier's prediction accuracy on training data — a common problem in supervised machine learning [132]. When ML models are trained to fit the training data in the best way possible, there is a significant risk of over-fitting, e.g., the model integrates ''noise'' instead of searching for a general predictive rule [132]. We highly encourage future studies to focus on maximizing a classifier's prediction accuracy on novel data, not the accuracy on training data. Further studies should investigate the performance of their classifier during various points in time and evaluate them on unseen data.

Another challenge arises due to the nature of ML methods: they require vast amounts of labeled data to achieve desired results [63]. The best-performing DL classification models were trained on datasets that comprised at least 5,000 labeled Tweets. DL models have an edge over traditional ML models when trained on large datasets as they can learn more representations and capture non-linear and complex patterns [32]. Labeled datasets are often not available and labeling data is a tedious process. In our literature review, we observed that there is no standard labeling process and studies resorted to labeling the data themselves or used models for example a lexicon-based approach to label the data. Hence, there is no standard labeling process in place. Sticking to labeling guidelines could assure that the data is correctly classified in the first place [133]. Otherwise, the trained model will have a bias due to incorrectly labeled data. State-of-the-art DL models are often considered by many researchers as a ''black box''. It is difficult to understand the dynamics of neural networks, i.e. their feature selection and prediction process [32].

When developing sentiment analysis tools for COVID-19-related data, we have observed a strong limitation in terms of the transferability of classification models. The majority of studies rely on English Tweet datasets. Due to linguistic and semantic differences between languages, it cannot be inferred that what works well on an English dataset, works to the same degree on a non-English dataset [25], [134]. This is particularly relevant in the context of COVID-19. To capture the mood of the public, it is essential for a government to include Tweets in various languages. Countries and societies are not homogeneous in their language. If only Tweets from one language are selected, a representation bias might arise [135]. Besides, access to Twitter is restricted in some countries and, thus, a true representation of the population can not be guaranteed. Sentiment analysis could result in misleading conclusions if particular user groups are over- or underrepresented in the Twitter dataset.

Particularly for classification models that rely on pre-training or use general-purpose lexica, using non-English Tweets is particularly challenging since the standard dictionaries are nearly exclusively in English [136]. Further research should explore sentiment analysis of Twitter data in languages other than English to expand the research field to governments and health organizations from other countries.

VI. CONCLUSION

In the light of COVID-19, there is a growing demand for governments and health organizations to analyze the public's sentiment on social media. By applying a lens from social and behavioral science research, we explored how sentiment analysis can provide relevant information for managing the pandemic. The literature review was conducted following the PRISMA guidelines to search and categorize existing literature on sentiment analysis on COVID-19 Twitter data with a particular focus on ML techniques. Out of 425 initial studies published in renowned journals, 40 papers were selected through a multistage screening process. Our study aims to provide governments and health authorities with guidance on the possible areas of sentiment analysis application for COVID-19 Twitter data and spark the development of innovative applications and algorithms. Since containing the pandemic requires a significant change in human behavior, we adopted the structure of van Bavel et al. [1] from social and behavioral science. Hence, decision-makers can extract, analyze and measure the public's thinking, feeling, and acting and make use of this ''sentiment dashboard'' to design target-oriented measures and communication campaigns.

In greater depth, we have compared the included studies' ML classification approaches in terms of their performance and pointed out shortcomings as well as areas for further research. Our findings show that sentiment analysis in the context of COVID-19 is mostly domain and application-specific. This means the classification techniques performing well on one dataset must not necessarily achieve satisfying results on a different dataset. In general, most studies relied on English datasets; thus, the transferability of their results might be limited. In terms of performance, we found out that ensemble models that comprise various ML classifiers commonly outperform a single classifier model. Besides, DL classifiers show a high accuracy given the availability of sufficiently labeled data. In particular, BERT or RoBERTa models provide promising results when pre-trained on Twitter data. In the daily fight against the coronavirus and future variants, sentiment analysis of COVID-19 Twitter data could provide governments and health-authorities with a tool to

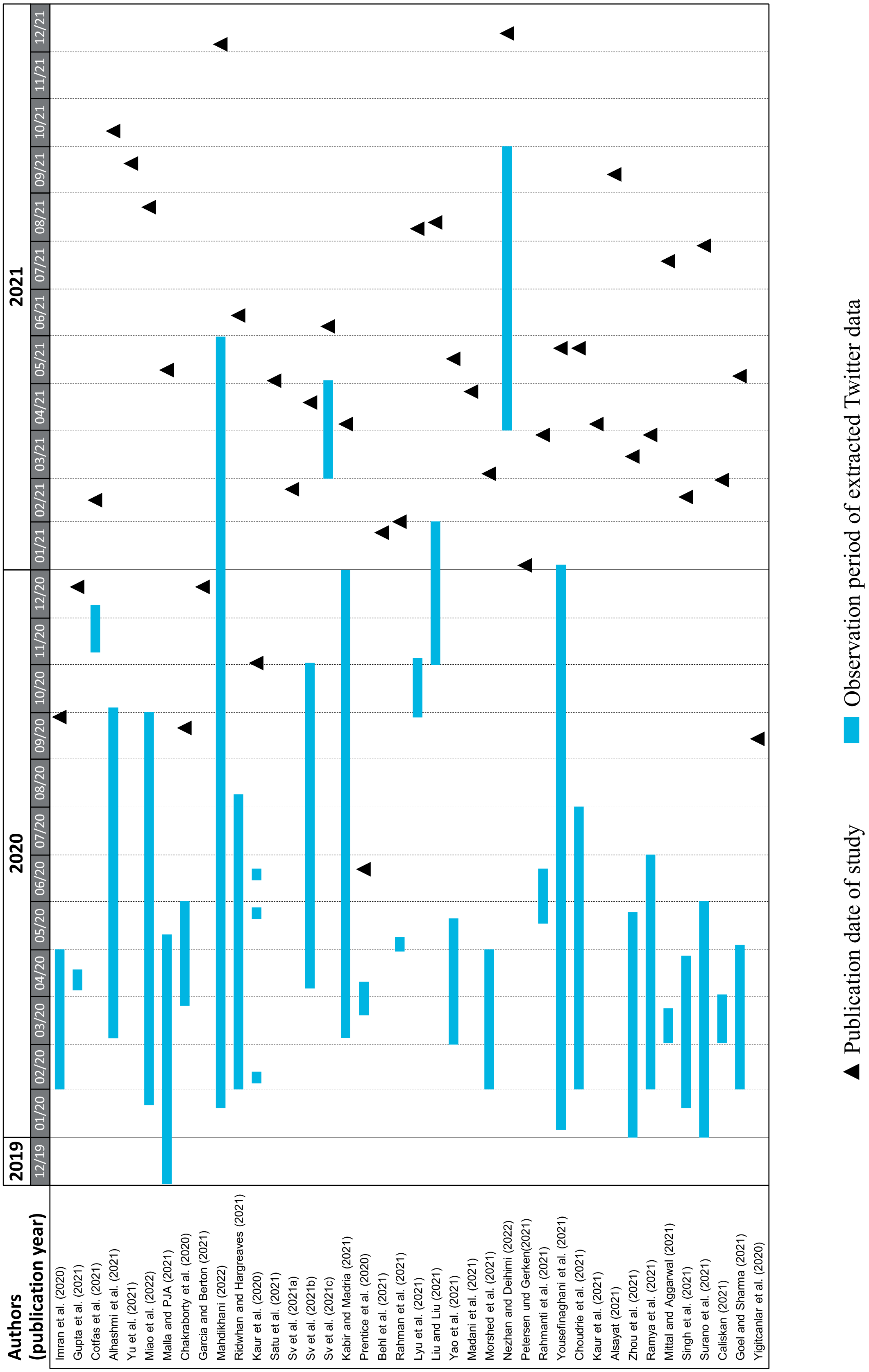


FIGURE 8. Overview of the observation period of the extracted Twitter data and the date of publication of the study.

TABLE 7. Detailed categorization of ML approaches according to the dataset, number of tweets, observation period, features, classification algorithm, and accuracy.

Ref.	Dataset	Number of tweets	Observation period	Features and classification algorithm	Accuracy
[106]	1) Self-collected COVID-19 dataset for testing 2) Emotional tweets dataset for training	1) 27,357 2) 21,051	1) Early Feb. 2020 until the end of Apr. 2020 2) N/A	1) GloVe (Twitter) + LSTM 2) FastText + LSTM	1) 81.9% (pos. emotions); 69.9% (neg. emotions) 2) 82.4%
[90]	Self-collected COVID-19 dataset that is split in 80% data for training and 20% for testing	12,741	5 Apr. 2020 to 17 Apr. 2020	Unigram + LinearSVC	84.4%
[102]	Self-collected COVID-19 dataset of which 1% is manually annotated for training and 99% for testing	2,349,659	9 Nov. 2020 and 8 Dec. 2020	BERT	78.94%
[79]	1) COVID-19 dataset 2) Expo2020 dataset The authors merged both datasets and used 70% for training and 30% for testing	1) 120,000 2) 5,000	1) 11 May 2020 to 15 May 2020 and 27 Sep. 2020 to 3 Oct. 2020 2) 14 May 2020 and 16 May 2020	Word2Vec + Bayes Factor Tree Augmented Naïve Bayes technique (BFTAN)	82.8%
[92]	LockdownTweets dataset (by Chen et al. 2020) where 1,098 randomly selected tweets got manually labeled and split into 733 labeled-train tweets and 365 labeled-test tweets	1,098	22 Jan. 2020 and 30 Sep. 2020	GloVe + LSTM	66%
[100]	WNUT 2020 Shared Task2 (Nguyen et al. 2020) contains around 10,000 tweets of which 7,000 tweets were used for training, 1,000 tweets were used for validation, and 2,000 tweets were used for testing	10,000	N/A	MVEDL Ensemble model consisting of RoBERTa, BERTweet, and CT-BERT	91.75%
[83]	Self-collected COVID-19 dataset of which 90% of the data is for training, 5% for validation, and 5% for training	22,985	1 Jan. 2019 to 23 Mar. 2020	Trigrams and TF-IDF score + Logistic Regression	81.4%
[103]	Self-collected COVID-19 dataset of which the tweets were classified using five-fold cross-validation with a split ratio of 75% to train the classifiers	1,251,216	1. 20 Jan. 2020 to 29 May 2021	CrystalFeel Ensemble Voting Classifier (RF, SGD, LG)	95.04%
[120]	COVID-19 dataset, which is a subset of IEEE Data portal developed by Rabindra Lamsal	16,000,000	1 Jan. 2020 to 20 Mar. 2020	TClustVID (Ensemble method with DL-Models)	>90%
[115]	Self-collected COVID-19 dataset of which 10,000 tweets were manually labeled into 10 different emotions. Kabir and Madria applied a 5-fold cross-validation with that self-extracted dataset, using 80% of the tweets as training data and 20% as testing data	56,014,158	5 Mar. 2020 to 31 Dec. 2021	Bidirectional LSTM	89.51%
[70]	1) Nepal earthquake (2015) by Basu et al. (2019) for training 2) Italian earthquake (2016) by Basu et al. (2019) for training 3) Self-collected COVID-19 dataset for testing	1) 70,897 2) 51,846 3) 2,274	1) 2015 2) 2019 3) N/A	Word2Vec + MLP	83%
[109]	Self-collected COVID-19 dataset	293,597	30 Apr. 2020 to 8 May 2020	Various exploration techniques used: BOW, DTM, POS, DP and n-grams, Logit model	56.18%
[105]	Self-collected COVID-19 dataset	6,314,327	28 Sep. 2020 to 4 Nov. 2020	XLNet	63%
[84]	1) Self-collected COVID-19 dataset for testing 2) "Emotion in Text" dataset by Crowdflower for training	1) 2,000,000 2) 39,740	1) Feb. 2020 to Jun. 2020 2) N/A	RoBERTa	80.33%
[77]	Self-collected COVID-19 dataset with a total of 11,000 tweets, 10,000 were used as training data and 1,000 as testing data	11,000	Feb. 2020 to Jun. 2020	Naïve Bayes	92.49% short tweets (<70 characters), 60.56% long tweets (<150 characters)
[76]	1) Self-collected COVID-19 dataset for testing 2) Crowdflower dataset for training and validation 3) Yelp dataset Challenge Repository 2015 for training and validation 4) Amazon dataset for training and validation	1) 4,242 2) 18,000 3) 299,000 4) 2,000,000	1) N/A 2) N/A 3) N/A 4) N/A	Ensemble Classifier (LSTM + FastText, BERT, ...)	92.65%
[86]	1) Self-collected COVID-19 dataset from various countries for training and testing 2) Self-collected COVID-19 dataset from India for training and testing 3) Github-data for testing	1) 417,023 2) 189,761 3) N/A	1) 20 Jan. 2020 to 25 Apr. 2020 2) 20 Jan. 2020 to 25 Apr. 2020 3) N/A	BERT	93.89%

TABLE 7. ***(Continued.)*** **Detailed categorization of ML approaches according to the dataset, number of tweets, observation period, features, classification algorithm, and accuracy.**

[111]	Self-collected COVID-19 dataset	46,078,750	1 Jan. 2020 to 30 Apr. 2020	Pre-trained GloVe + Deep Learning Model (RNN, CNN)	71%
[95]	Self-collected COVID-19 dataset. A 5-fold cross-validation has been applied.	29,469,349	1 Feb. 2020 to 2 May 2020	TF-IDF Random Forest	96% (AUC ROC)

Abbreviations included in the table:

Global Vectors for Word Representation (GloVe), Long-short term memory (LSTM), Linear Support Vector Classifier (LinearSVC), Valence Aware Dictionary for Sentiment Reasoning (VADER), Bidirectional Encoder Representations from Transformers (BERT), MVEDL is based on RoBERTa, BERTweet and CT-BERT (all differentiations of BERT), Term Frequency-Inverse Document Frequency (TF-IDF), Afinn, Random Forest (RF), Stochastic Gradient Descent (SGD), Logistic Regression (LR), Iterative deepening A* (IDA), Latent Dirichlet Allocation (LDA), TClustVID is an ensemble method with Deep Learning models, Multi-Layer Perceptron (MLP), Bag-Of-Words (BOW), Document term matrix (DTM), Part-of-Speech (POS), Dependency Parsing (DP), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN).

change the trajectory of the pandemic by making timely and well-informed decisions.

A. LIMITATIONS

Our research comes not without limitations. Due to the nature of a systematic literature review, there is the possibility that our literature search string was too narrowly defined and did not capture all relevant studies. By design, we included only peer-reviewed journal papers from renowned databases and did not include conference papers. Moreover, further limitations were identified: Firstly, our study has not found a clear pattern indicating which category or situation a certain kind of classifier performs best. Secondly, the majority of studies are based on English datasets. Thus, the transferability of the results to other non-English speaking countries is limited. Thirdly, the robustness of ML classification models in the context of COVID-19 has not been exhaustively examined and might not justify the application of sentiment analysis over conducting a less technical survey study. Fourthly, the opinion expressed on Twitter might not mirror the public's opinion due to representation errors.

B. FUTURE WORK

Our literature review has identified several opportunities for future research. Following the categories of van Bavel et al. [1], future work could focus on the unresearched topic "individual and collective interest" to further explore sentiment analysis from a behavioral and social science lens. From a technical point of view, we see a need for future studies to stimulate the development and implementation of real-world applications of sentiment analysis in the context of COVID-19. Hence, the performance of classification algorithms should be evaluated on new, unseen data at various time points. Moreover, a stronger focus on cross-validated results instead of optimizing a classifier's accuracy on training data might be beneficial in developing new sentiment analysis applications for COVID-19 Twitter data. Our study reveals that most of the included papers used datasets that date back to the first two quarters of 2020. It would be interesting to conduct follow-up studies to evaluate the usability of the proposed classification models on more recent data. This could provide more information on the robustness of the classification models. Since most of the included studies are based on English datasets, we identified a need to use datasets of another language for sentiment analysis. In closing, we hope to inspire additional interdisciplinary research that supports the fight against the COVID-19 pandemic by combining sentiment analysis with behavioral and social science research.

APPENDIX A
COPYRIGHT NOTICE

"TWITTER, TWEET, RETWEET and the Twitter Bird logo are trademarks of Twitter Inc. or its affiliates." - https://about.twitter.com/en/who-we-are/brand-toolkit.

All icons from UXWing are free to download and can be used for personal and commercial projects - https://uxwing.com/license.

All emojis designed by OpenMoji – the open-source emoji and icon project. License: CC BY-SA 4.0 - https://openmoji.org.

APPENDIX B
OBSERVATION PERIODS AND PUBLICATION DATES

(see Figure 8)

APPENDIX C
DETAILED TECHNICAL CATEGORIZATION OF MACHINE LEARNING APPROACHES

(see Table 7)

APPENDIX D
SEARCH STRINGS FOR SCIENCE DIRECT

Here we present in detail the search string for ScienceDirect. In the ScienceDirect database, the limit for boolean operators is eight. Also, wildcards ("*") are not supported in the database search. Our original search string has 10 boolean operators and also a wildcard. For that reason, the search string is split into the following three search strings:

1) (("Sentiment Analysis") AND ("Twitter" OR "Tweet") AND ("Machine Learning" OR "Deep Learning" OR "Natural Language Processing" OR "NLP") AND ("COVID-19" OR "Pandemic"))

2) ((“Opinion Analysis”) AND (“Twitter” OR “Tweet”) AND (“Machine Learning” OR “Deep Learning” OR “Natural Language Processing” OR “NLP”) AND (“COVID-19” OR “Pandemic”))
3) ((“Opinion Mining”) AND (“Twitter” OR “Tweet”) AND (“Machine Learning” OR “Deep Learning” OR “Natural Language Processing” OR “NLP”) AND (“COVID-19” OR “Pandemic”))

Search string 1) results in 220 studies, search string 2) results in 7 studies, and search string 3) results in 46 studies. After eliminating the duplicates, the search in the ScienceDirect database results in a total of 234 studies.

ACKNOWLEDGMENT

(Niklas Braig and Alina Benz contributed equally to this work.)

REFERENCES

[1] J. J. Van Bavel et al., “Using social and behavioural science to support COVID-19 pandemic response,” *Nature Human Behaviour*, vol. 4, no. 5, pp. 460–471, Apr. 2020.

[2] D. Adam, “The pandemic’s true death toll: Millions more than official counts,” *Nature*, vol. 601, no. 7893, pp. 312–315, Jan. 2022.

[3] S.-F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, and Z. A. Butt, “What social media told us in the time of COVID-19: A scoping review,” *Lancet Digit. Health*, vol. 3, no. 3, pp. e175–e194, Mar. 2021.

[4] A. Górska, D. Dobija, G. Grossi, and Z. Staniszewska, “Getting through COVID-19 together: Understanding local governments’ social media communication,” *Cities*, vol. 121, Feb. 2022, Art. no. 103453.

[5] M. Perez-Cepeda and L. G. Arias-Bolzmann, “Sociocultural factors during COVID-19 pandemic: Information consumption on Twitter,” *J. Bus. Res.*, vol. 140, pp. 384–393, Feb. 2022.

[6] E. Chen, K. Lerman, and E. Ferrara, “Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set,” *JMIR Public Health Surveill.*, vol. 6, no. 2, May 2020, Art. no. e19273.

[7] M. Paul and M. Dredze, “You are what you tweet: Analyzing Twitter for public health,” in *Proc. Int. AAAI Conf. Web Social Media*, vol. 5, no. 1, 2011, pp. 265–272.

[8] A. Amara, M. A. H. Taieb, and M. B. Aouicha, “Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis,” *Appl. Intell.*, vol. 51, no. 5, pp. 3052–3073, Feb. 2021.

[9] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, “A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks,” *Exp. Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 114006.

[10] R. Buettner and K. Buettner, “A systematic literature review of Twitter research from a socio-political revolution perspective,” in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 2206–2215.

[11] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, Jan. 2008.

[12] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, “An analysis of COVID-19 vaccine sentiments and opinions on Twitter,” *Int. J. Infectious Diseases*, vol. 108, pp. 256–262, Jul. 2021.

[13] H. Piedrahita-Valdés, D. Piedrahita-Castillo, J. Bermejo-Higuera, P. Guillem-Saiz, J. R. Bermejo-Higuera, J. Guillem-Saiz, J. A. Sicilia-Montalvo, and F. Machío-Regidor, “Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019,” *Vaccines*, vol. 9, no. 1, p. 28, Jan. 2021.

[14] S. Liu and J. Liu, “Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis,” *Vaccine*, vol. 39, no. 39, pp. 5499–5505, Sep. 2021.

[15] S. Sommariva, J. Mote, H. B. Bon, H. Razafindraibe, D. Ratovozanany, V. Rasoamanana, S. Abeyesekera, P. Muhamedkhojaeva, T. Bashar, J. James, and M. Sani, “Social listening in eastern and Southern Africa, a UNICEF risk communication and community engagement strategy to address the COVID-19 infodemic,” *Health Secur.*, vol. 19, no. 1, pp. 57–64, Feb. 2021.

[16] A. Bandura, “Self-efficacy: Toward a unifying theory of behavioral change,” *Psychol. Rev.*, vol. 84, no. 2, pp. 191–215, Feb. 1977.

[17] I. Ajzen, “The theory of planned behavior,” *Org. Behav. Hum. Decis. Process.*, vol. 50, no. 2, pp. 179–211, Dec. 1991.

[18] J. R. Martin and P. R. R. White, *The Language of Evaluation*. Basingstoke, U.K.: Palgrave Macmillan, 2005.

[19] P. Korenek and M. Šimko, “Sentiment analysis on microblog utilizing appraisal theory,” *World Wide Web*, vol. 17, no. 4, pp. 847–867, Aug. 2014.

[20] C. Whitelaw, N. Garg, and S. Argamon, “Using appraisal groups for sentiment analysis,” in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2005, pp. 625–631.

[21] C. Soo-Guan Khoo, A. Nourbakhsh, and J. Na, “Sentiment analysis of online news text: A case study of appraisal theory,” *Online Inf. Rev.*, vol. 36, no. 6, pp. 858–878, Nov. 2012.

[22] A. Kumar and A. Jaiswal, “Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on Twitter,” *Multimedia Tools Appl.*, vol. 78, no. 20, pp. 29529–29553, Oct. 2019.

[23] A. Giachanou and F. Crestani, “Like it or not: A survey of Twitter sentiment analysis methods,” *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–41, Jun. 2016.

[24] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, H. Hameed, and M. Alaa, “Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review,” *Exp. Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114155.

[25] K. K. Agustiningsih, E. Utami, and H. Al Fatta, “Sentiment analysis of COVID-19 vaccine on Twitter social media: Systematic literature review,” in *Proc. IEEE 5th Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Nov. 2021, pp. 121–126.

[26] A. H. Alamoodi, B. B. Zaidan, M. Al-Masawa, S. M. Taresh, S. Noman, I. Y. Y. Ahmaro, S. Garfan, J. Chen, M. A. Ahmed, A. A. Zaidan, O. S. Albahri, U. Aickelin, N. N. Thamir, J. A. Fadhil, and A. Salahaldin, “Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy,” *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 104957.

[27] M. J. Page et al., “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” *Systematic Rev.*, vol. 10, no. 1, Dec. 2021, Art. no. 105906.

[28] J. Bandy, “Problematic machine behavior,” *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, pp. 1–34, Apr. 2021.

[29] M. V. Mäntylä, D. Graziotin, and M. Kuutila, “The evolution of sentiment analysis—A review of research topics, venues, and top cited papers,” *Comput. Sci. Rev.*, vol. 27, pp. 16–32, Feb. 2018.

[30] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New avenues in opinion mining and sentiment analysis,” *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, Mar. 2013.

[31] N. V. Babu and E. G. M. Kanaga, “Sentiment analysis in social media data for depression detection using artificial intelligence: A review,” *Social Netw. Comput. Sci.*, vol. 3, no. 1, p. 74, Jan. 2021.

[32] A. Yadav and D. K. Vishwakarma, “Sentiment analysis using deep learning architectures: A review,” *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020.

[33] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, “The state-of-the-art in Twitter sentiment analysis,” *ACM Trans. Manage. Inf. Syst.*, vol. 9, no. 2, pp. 1–29, Jun. 2018.

[34] S. R. Rufai and C. Bunce, “World leaders’ usage of Twitter in response to the COVID-19 pandemic: A content analysis,” *J. Public Health*, vol. 42, no. 3, pp. 510–516, 2020.

[35] U. Qazi, M. Imran, and F. Ofli, “GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information,” *SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, Jun. 2020.

[36] O. Edo-Osagie, B. De La Iglesia, I. Lake, and O. Edeghere, “A scoping review of the use of Twitter for public health research,” *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103770.

[37] Z. Nanli, Z. Ping, L. Weiguo, and C. Meng, “Sentiment analysis: A literature review,” in *Proc. Int. Symp. Manage. Technol. (ISMOT)*, Nov. 2012, pp. 572–576.

[38] K. Sunil and S. Beniwal, “Sentiment analysis: A tool for mining opinions and emotions,” in *Proc. Int. Conf. Innov. Comput. Commun.*, Dec. 2020, pp. 1–8, doi: 10.2139/ssrn.3746951.

[39] R. Feldman, “Techniques and applications for sentiment analysis,” *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[40] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

[41] G. A. Miller, "WordNet," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[42] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," *Proc. Comput. Sci.*, vol. 161, pp. 707–714, Jan. 2019.

[43] C. Bhadane, H. Dalal, and H. Doshi, "Sentiment analysis: Measuring opinions," *Proc. Comput. Sci.*, vol. 45, pp. 808–814, Jan. 2015.

[44] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Jul. 2016, pp. 628–632.

[45] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proc. 1st ACM Conf. Online Social Netw.*, Oct. 2013, pp. 27–38.

[46] K. Ayyub, S. Iqbal, E. U. Munir, M. W. Nisar, and M. Abbasi, "Exploring diverse features for sentiment quantification using machine learning algorithms," *IEEE Access*, vol. 8, pp. 142819–142831, 2020.

[47] D. Effrosynidis, S. Symeonidis, and A. Arampatzis, "A comparison of pre-processing techniques for Twitter sentiment analysis," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*, Sep. 2017, pp. 394–406.

[48] Z. Jianqiang and G. Xiaolin, "Comparison research on text preprocessing methods on Twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017.

[49] P. Han, S. Shen, D. Wang, and Y. Liu, "The influence of word normalization in English document clustering," in *Proc. IEEE Int. Conf. Comput. Sci. Autom. Eng. (CSAE)*, May 2012, pp. 116–120.

[50] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, "Monitoring the public opinion about the vaccination topic from tweets analysis," *Exp. Syst. Appl.*, vol. 116, pp. 209–226, Feb. 2019.

[51] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for Twitter sentiment analysis," *Exp. Syst. Appl.*, vol. 110, pp. 298–310, Nov. 2018.

[52] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *Proc. AAAI Workshop Learn. Text Categorization*, vol. 752, no. 1, 1998, pp. 41–48.

[53] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[54] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop (COLT)*, Jul. 1992, pp. 144–152.

[55] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[56] A. Qazi, R. G. Raj, G. Hardaker, and C. Standing, "A systematic literature review on opinion types and sentiment analysis techniques," *Internet Res.*, vol. 27, no. 3, pp. 608–630, Jun. 2017.

[57] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[58] Ankit and N. Saleena, "An ensemble classification system for Twitter sentiment analysis," *Proc. Comput. Sci.*, vol. 132, pp. 937–946, Jan. 2018.

[59] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, p. e1253, Mar. 2018.

[60] N. Jones, "Computer science: The learning machines," *Nature*, vol. 505, no. 7482, pp. 146–148, Jan. 2014.

[61] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.

[62] Y. Bengio, Y. LeCun, and G. E. Hinton, "Deep learning for AI," *Commun. ACM*, vol. 64, pp. 58–65, Jun. 2021.

[63] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, Dec. 2015.

[64] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep machine learning—A new frontier in artificial intelligence research [research frontier]," *IEEE Comput. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, Nov. 2010.

[65] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[66] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020.

[67] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[68] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Oct. 2018, pp. 4171–4186.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[70] S. Behl, A. Rao, S. Aggarwal, S. Chadha, and H. S. Pannu, "Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises," *Int. J. Disaster Risk Reduction*, vol. 55, Mar. 2021, Art. no. 102101.

[71] Z. Taheri Zadeh, S. Rahmani, F. Alidadi, S. Joushi, and K. Esmaeilpour, "Depresssion, anxiety and other cognitive consequences of social isolation: Drug and non-drug treatments," *Int. J. Clin. Pract.*, vol. 75, no. 12, Dec. 2021, e14949.

[72] (Mar. 1, 2022). *FACT SHEET: President Biden to Announce Strategy to Address Our National Mental Health Crisis, As Part of Unity Agenda in His First State of the Union*. The White House. [Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2022/03/01/fact-sheet-president-biden-to-announce-strategy-to-address-our-national-mental-health-crisis-as-part-of-unity-agenda-in-his-first-state-of-the-union/

[73] C. Betsch, "How behavioural science data helps mitigate the COVID-19 crisis," *Nature Human Behaviour*, vol. 4, no. 5, p. 438, Mar. 2020.

[74] N. Salari, A. Hosseinian-Far, R. Jalali, A. Vaisi-Raygani, S. Rasoulpoor, M. Mohammadi, S. Rasoulpoor, and B. Khaledi-Paveh, "Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: A systematic review and meta-analysis," *Globalization Health*, vol. 16, no. 1, p. 57, Dec. 2020.

[75] T. Sharot, "The optimism bias," *Current Biol.*, vol. 21, no. 23, pp. R941–R945, Dec. 2011.

[76] A. Alsayat, "Improving sentiment analysis for social media applications using an ensemble deep learning language model," *Arabian J. Sci. Eng.*, vol. 47, no. 2, pp. 2499–2511, Oct. 2022.

[77] B. N. Ramya, S. M. Shetty, A. M. Amaresh, and R. Rakshitha, "Smart Simon Bot with public sentiment analysis for novel COVID-19 tweets stratification," *Social Netw. Comput. Sci.*, vol. 2, no. 3, p. 227, May 2021.

[78] T. Yigitcanlar, N. Kankanamge, A. Preston, P. S. Gill, M. Rezayee, M. Ostadnia, B. Xia, and G. Ioppolo, "How can social media analytics assist authorities in pandemic-related policy decisions? Insights from Australian states and territories," *Health Inf. Sci. Syst.*, vol. 8, no. 1, p. 37, Oct. 2020.

[79] S. M. Alhashmi, A. M. Khedr, I. Arif, and M. El Bannany, "Using a hybrid-classification method to analyze Twitter data during critical events," *IEEE Access*, vol. 9, pp. 141023–141035, 2021.

[80] S. A. Morshed, S. S. Khan, R. B. Tanvir, and S. Nur, "Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis," *J. Urban Manage.*, vol. 10, no. 2, pp. 155–165, Jun. 2021.

[81] K. Mohamed Ridhwan and C. A. Hargreaves, "Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore," *Int. J. Inf. Manage. Data Insights*, vol. 1, no. 2, Nov. 2021, Art. no. 100021.

[82] S. Kaur, P. Kaul, and P. M. Zadeh, "Monitoring the dynamics of emotions during COVID-19 using Twitter data," *Proc. Comput. Sci.*, vol. 177, pp. 423–430, Jan. 2020.

[83] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment analysis of COVID-19 tweets by deep learning classifiers—A study to show how popularity is affecting accuracy in social media," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106754.

[84] J. Choudrie, S. Patil, K. Kotecha, N. Matta, and I. Pappas, "Applying and understanding an advanced, novel deep learning approach: A COVID 19, text based, emotions analysis study," *Inf. Syst. Frontiers*, vol. 23, no. 6, pp. 1431–1465, Dec. 2021.

[85] R. Mittal, A. Mittal, and I. Aggarwal, "Identification of affective valence of Twitter generated sentiments during the COVID-19 outbreak," *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 108, Dec. 2021.

[86] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 33, Dec. 2021.

[87] S. Billore and T. Anisimova, "Panic buying research: A systematic literature review and future research agenda," *Int. J. Consum. Stud.*, vol. 45, no. 4, pp. 777–804, Jul. 2021.

[88] C. Prentice, J. Chen, and B. Stantic, "Timed intervention in COVID-19 and panic buying," *J. Retailing Consum. Services*, vol. 57, Nov. 2020, Art. no. 102203.

[89] M. Alsan and M. Wanamaker, "Tuskegee and the health of black men," *Quart. J. Econ.*, vol. 133, no. 1, pp. 407–455, Feb. 2018.

[90] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, "Sentiment analysis of lockdown in India during COVID-19: A case study on Twitter," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 992–1002, Aug. 2021.

[91] Z. B. Nezhad and M. A. Deihimi, "Analyzing Iranian opinions toward COVID-19 vaccination," *IJID Regions*, vol. 3, pp. 204–210, Jun. 2022.

[92] L. Miao, M. Last, and M. Litvak, "Tracking social media during the COVID-19 pandemic: The case study of lockdown in new York state," *Exp. Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 115797.

[93] X. Yu, M. D. Ferreira, and F. V. Paulovich, "Senti-COVID19: An interactive visual analytics system for detecting public sentiment and insights regarding COVID-19 from social media," *IEEE Access*, vol. 9, pp. 126684–126697, 2021.

[94] A. R. Rahmanti, D. N. A. Ningrum, L. Lazuardi, H.-C. Yang, and Y.-C. Li, "Social media data analytics for outbreak risk communication: Public attention on the 'New Norma' during the COVID-19 pandemic in Indonesia," *Comput. Methods Programs Biomed.*, vol. 205, Jun. 2021, Art. no. 106083.

[95] R. Goel and R. Sharma, "Studying leaders & their concerns using online social media during the times of crisis–A COVID case study," *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 46, Dec. 2021.

[96] M. Siegrist, L. Luchsinger, and A. Bearth, "The impact of trust and risk perception on the acceptance of measures to reduce COVID-19 cases," *Risk Anal.*, vol. 41, no. 5, pp. 787–800, May 2021.

[97] H. J. Larson, "The biggest pandemic risk? Viral misinformation," *Nature*, vol. 562, no. 7727, p. 309, Oct. 2018.

[98] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. J. Freeman, G. Recchia, A. M. Van Der Bles, and S. Van Der Linden, "Susceptibility to misinformation about COVID-19 around the world," *Roy. Soc. Open Sci.*, vol. 7, no. 10, Oct. 2020, Art. no. 201199.

[99] Y. Madani, M. Erritali, and B. Bouikhalene, "Using artificial intelligence techniques for detecting COVID-19 epidemic fake news in Moroccan tweets," *Results Phys.*, vol. 25, Jun. 2021, Art. no. 104266.

[100] S. Malla and P. J. A. Alphonse, "COVID-19 outbreak: An ensemble pretrained deep learning model for detecting informative tweets," *Appl. Soft Comput.*, vol. 107, Aug. 2021, Art. no. 107495.

[101] P. Sv, J. Tandon, Vikas, and H. Hinduja, "Indian citizen's perspective about side effects of COVID-19 vaccine—A machine learning study," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 15, no. 4, Jul. 2021, Art. no. 102172.

[102] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanas, D. S. Gherai, and F. Tajariol, "The longest month: Analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement," *IEEE Access*, vol. 9, pp. 33203–33223, 2021.

[103] M. Mahdikhani, "Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of COVID-19 pandemic," *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 1, Apr. 2022, Art. no. 100053.

[104] K. Petersen and J. M. Gerken, "#COVID-19: An exploratory investigation of hashtag usage on Twitter," *Health Policy*, vol. 125, no. 4, pp. 541–547, Apr. 2021.

[105] H. Lyu, J. Wang, W. Wu, V. Duong, X. Zhang, T. D. Dye, and J. Luo, "Social media study of public opinions on potential COVID-19 vaccines: Informing dissent, disparities, and dissemination," *Intell. Med.*, vol. 2, no. 1, pp. 1–12, Feb. 2022.

[106] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets," *IEEE Access*, vol. 8, pp. 181074–181090, 2020.

[107] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107057.

[108] Z. Yao, J. Yang, J. Liu, M. Keith, and C. Guan, "Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19," *Cities*, vol. 116, Sep. 2021, Art. no. 103273.

[109] M. M. Rahman, G. G. M. N. Ali, X. J. Li, J. Samuel, K. C. Paul, P. H. J. Chong, and M. Yakubov, "Socioeconomic factors analysis for COVID-19 U.S. reopening sentiment with Twitter and census data," *Heliyon*, vol. 7, no. 2, Feb. 2021, Art. no. e06200.

[110] F. V. Surano, M. Porfiri, and A. Rizzo, "Analysis of lockdown perception in the United States during the COVID-19 pandemic," *Eur. Phys. J. Special Topics*, vol. 231, no. 9, pp. 1625–1633, Jul. 2022.

[111] C. Caliskan, "How does 'A bit of everything America' state feel about COVID-19? A quantitative Twitter analysis of the pandemic in Ohio," *J. Comput. Social Sci.*, vol. 5, no. 1, pp. 19–45, May 2022.

[112] D. A. Gubler, L. M. Makowski, S. J. Troche, and K. Schlegel, "Loneliness and well-being during the COVID-19 pandemic: Associations with personality and emotion regulation," *J. Happiness Stud.*, vol. 22, no. 5, pp. 2323–2342, Jun. 2021.

[113] C. L. Jarzyna, "Parasocial interaction, the COVID-19 quarantine, and digital age media," *Human Arenas*, vol. 4, no. 3, pp. 413–429, Sep. 2021.

[114] A. J. Crum, S. Peter, and A. Shawn, "Rethinking stress: The role of mindsets in determining the stress response," *J. Personality Social Psychol.*, vol. 104, no. 4, pp. 716–733, Feb. 2013.

[115] M. Y. Kabir and S. Madria, "EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets," *Online Social Netw. Media*, vol. 23, May 2021, Art. no. 100135.

[116] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, "A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets," *Inf. Syst. Frontiers*, vol. 23, no. 6, pp. 1417–1429, Dec. 2021.

[117] S. V. Praveen, R. Ittamalla, and G. Deepak, "Analyzing Indian general public's perspective on anxiety, stress and trauma during COVID-19—A machine learning study of 840,000 tweets," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 15, no. 3, pp. 667–671, May 2021.

[118] J. Zhou, S. Yang, C. Xiao, and F. Chen, "Examination of community sentiment dynamics due to COVID-19 pandemic: A case study from a state in Australia," *Social Netw. Comput. Sci.*, vol. 2, no. 3, p. 201, May 2021.

[119] C. Moreno et al., "How mental health care should change as a consequence of the COVID-19 pandemic," *Lancet Psychiatry*, vol. 7, no. 9, pp. 813–824, Sep. 2020.

[120] M. S. Satu, M. I. Khan, M. Mahmud, S. Uddin, M. A. Summers, J. M. W. Quinn, and M. A. Moni, "TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107126.

[121] S. Praveen, R. Ittamalla, and G. Deepak, "Analyzing the attitude of Indian citizens towards COVID-19 vaccine—A text analytics study," *Diabetes Metabolic Syndrome: Clin. Res. Rev.*, vol. 15, no. 2, pp. 595–599, Mar. 2021.

[122] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Oct. 1990.

[123] A. Chakraborty and S. Bose, "Around the world in 60 days: An exploratory study of impact of COVID-19 on online global news sentiment," *J. Comput. Social Sci.*, vol. 3, no. 2, pp. 367–400, Nov. 2020.

[124] J. C. Lyu, E. L. Han, and G. K. Luli, "COVID-19 vaccine related discussion on Twitter: Topic modeling and sentiment analysis," *J. Med. Internet Res.*, vol. 23, no. 6, Jun. 2021, Art. no. e24435.

[125] A. Zunic, P. Corcoran, and I. Spasic, "Sentiment analysis in health and well-being: Systematic review," *JMIR Med. Informat.*, vol. 8, no. 1, Jan. 2020, Art. no. e16023.

[126] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artif. Intell. Med.*, vol. 64, no. 1, pp. 17–27, May 2015.

[127] J. Kazmaier and J. H. van Vuuren, "The power of ensemble learning in sentiment analysis," *Exp. Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 115819.

[128] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decis. Support Syst.*, vol. 57, pp. 77–93, Jan. 2014.

[129] R. Srivastava and M. P. S. Bhatia, "Ensemble methods for sentiment analysis of on-line micro-texts," in *Proc. Int. Conf. Recent Adv. Innov. Eng. (ICRAIE)*, Dec. 2016, doi: 10.1109/ICRAIE.2016.7939525.

[130] A. Hassan, A. Abbasi, and D. Zeng, "Twitter sentiment analysis: A bootstrap ensemble framework," in *Proc. Int. Conf. Social Comput.*, Sep. 2013, pp. 357–364.

[131] Z. Pan and W. Xu, "Deep learning based sentiment analysis during public health emergency," in *Proc. 13th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, Aug. 2021, pp. 137–140.

[132] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 326–327, Sep. 1995.

[133] S. Mohammad, "A practical guide to sentiment annotation: Challenges and solutions," in *Proc. 7th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2016, pp. 174–179.

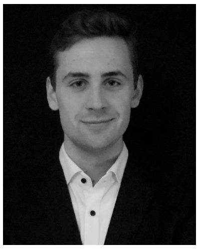
[134] H. Peng, E. Cambria, and A. Hussain, "A review of sentiment analysis research in Chinese language," *Cognit. Comput.*, vol. 9, no. 4, pp. 423–435, Aug. 2017.

[135] Y. Huang, D. Guo, A. Kasakoff, and J. Grieve, "Understanding U.S. regional linguistic variation with Twitter data analysis," *Comput., Environ. Urban Syst.*, vol. 59, pp. 244–255, Sep. 2016.

[136] M. F. R. A. Bakar, N. Idris, L. Shuib, and N. Khamis, "Sentiment analysis of noisy Malay text: State of art, challenges and future work," *IEEE Access*, vol. 8, pp. 24687–24696, 2020.

NIKLAS BRAIG received the B.A. degree in business and management from the Management Center Innsbruck (MCI), Austria, in 2019. He is currently pursuing the M.Sc. degree in business administration with the University of Bayreuth, Germany. From 2017 to 2018, he has spent one year at the Instituto Tecnológico Autónomo de México (ITAM). His current research interests include the intersection between corporate finance and machine learning.

ALINA BENZ received the B.A. degree in business administration from the Augsburg University of Applied Sciences, Germany, in 2019. She is currently pursuing the master's degree in business administration, with a specialization in information systems, with the University of Bayreuth, Germany. Her current research interests include machine learning and deep learning, in the field of supply chain management and process management.



SOEREN VOTH received the bachelor's degree in economics from the University of Bayreuth, Germany, in 2021, where he is currently pursuing the master's degree in economics, with a specialization in information systems. His current research interests include machine learning, natural language processing, and deep learning.

JOHANNES BREITENBACH received the B.Eng. degree in automotive technology from Coburg University, Germany, in 2019, and the M.Sc. degree in information systems from Aalen University, Germany, in 2021. He is a former Research Associate at the Chair of Information Systems and Data Science, University of Bayreuth, where he works in the field of data engineering. His research interests include machine learning, deep learning, image processing, computer vision at the interface to materials science, medicine, and health sciences.

RICARDO BUETTNER (Senior Member, IEEE) received the Dipl.-Inf. degree in computer science and the Dipl.-Wirtsch.-Ing. degree in industrial engineering and management from the Technical University of Ilmenau, Germany, the Dipl.-Kfm. degree in business administration from the University of Hagen, Germany, the Ph.D. degree in information systems from the University of Hohenheim, Germany, and the Habilitation (venia legendi) degree in information systems from the University of Trier, Germany.

He is currently a Chaired Professor of information systems and data science at the University of Bayreuth, Germany. He has published over 140 peer-reviewed articles, including articles in *Electronic Markets*, *AIS Transactions on Human–Computer Interaction*, *Personality and Individual Differences*, *European Journal of Psychological Assessment*, *PLOS ONE*, and IEEE Access.

Dr. Buettner has received 17 international best paper, the best reviewer, and the service awards and award nominations, including Best Paper Awards by *AIS Transactions on Human–Computer Interaction*, *Electronic Markets* journal, and HICSS, for his work.