IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 September 2022, accepted 2 November 2022, date of publication 7 November 2022,
date of current version 10 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3220369

RESEARCH ARTICLE

COVID-19 Rumor Detection Using Psycho-Linguistic Features

SYED MAHBUB, ERIC PARDEDE, (Senior Member, IEEE), AND A. S. M. KAYES, (Member, IEEE)
Department of Computer Science and Information Technology, La Trobe University, Bundoora, VIC 3086, Australia
Corresponding author: A. S. M. Kayes (a.kayes@latrobe.edu.au)

ABSTRACT During the onset of COVID-19 pandemic, the social media was flooded with misinformation. Irrespective of the type of the misinformation, such contents played a significant role in increasing confusion among people in the middle of an ongoing crisis. The purpose of the study is to investigate the nature of a specific type of misinformation, i.e., rumors, surrounding COVID-19. The study utilizes a publicly available and labelled Twitter dataset and proposes a novel feature space, which can detect rumor instances with high accuracy. The proposed feature space not only includes content-based features, but also includes psycho-linguistic features to further study the characteristics of the content from the perspectives of linguistics and psychology. The use of psycho-linguistic features has been utilised to understand certain dramatisation of text in the domain of conspiracy propagation and fake news detection. However, the use of such dramatisation detection approach has never been used for the purposes of rumor detection. Our study first outlines the differences between these categories of misinformation propagation and clarifies where rumor fits-in under the broader umbrella of misinformation. It further outlines how the use of psycho-linguistic features can also improve the detection accuracy of rumors on social media. The study demonstrates through multiple experimental setups that psycho-linguistic features improves the detection accuracy and associated performance measures, such as precision and recall, for COVID-19 rumors on Twitter. The observed improvements are consistent across multiple machine learning models.

INDEX TERMS Misinformation detection, rumor detection, COVID-19 rumor detection, feature engineering, psycho-linguistic features.

I. INTRODUCTION

COVID-19 changed the world as we knew it. It is not only responsible for taking millions of human lives, but also for reshaping the world's political, economic, and social landscape to a significant degree. The impact of the pandemic on social lives world-wide is significant as it forced online social networks (OSN) to be the main platform of social interactions for a substantial period of time. People around the world were already relying on OSNs for maintaining a social life as well as getting their information in general. However, COVID-19 has significantly raised this dependency to a whole new level.

According to Google News [1], the word-wide COVID-19 related death toll is 6,434,754 as of August 2022, with countries such as United States, Brazil, India, Russia, and Mexico topping the list in terms of number of deaths. During the early stages of the pandemic, people around the world started to feel anxious and panicked. On one hand, people were desperate for more information about the disease and its symptoms, and on the other hand, it was not possible to socialise physically, or reach out to a non-virtual news outlets. Therefore, people turned to online platforms to get information, as under the circumstances, it was the most accessible source. Among other platforms, Twitter has seen a significant increase in activity around the world during the period of the pandemic. According to their blog [2], between March 2020 and August 2021, there have been more than 22.2 million COVID-19 related Tweets in Australia, where the tweets initiated conversations related to health statistics, lockdown restrictions, government initiatives, and so on.

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

Word-wide, the number of Twitter users has increased by approximately 34% during the pandemic [3].

Although, OSN platforms such as Twitter provided a large group of scared and confused people the means to stay connected, the increased activity also presented fraudsters with the opportunity for spreading misinformation using these platforms. The purposes of these fraudsters can vary from personal gains to simply misleading people to promote chaos. Irrespective of the purpose, the misinformation themselves can be damaging and obstructive for healthcare workers and governments around the world. Therefore, it is absolutely critical to have implementable solutions that can automatically detect specific categories of misinformation on online platforms, such as, Twitter, which will drastically reduce the negative impacts these misinformation can cause. Consequently, it is imperative to study the nature of these online misinformation and propose a set of features that can contribute to higher detection accuracy.

A. THE CONTRIBUTIONS

In this paper we investigate the nature of rumors surrounding COVID-19 on Twitter. Our aim was to investigate whether certain linguistic markers that can identify dramatisation of texts, can improve the detection accuracy of these rumors, when included in the feature space. Our investigation demonstrates how these features indeed improve the detection accuracy of rumor tweets. The contributions of the study are briefly summarised below:

- We propose a hybrid feature space for the automated detection of COVID-19 rumors on Twitter, which includes tweet content-specific features, twitter-specific contextual features, and psycho-linguistic features.
- We include psycho-linguistic features in the detection of rumor type of misinformation, to capture text dramatisation.
- We demonstrate how the psycho-linguistic features can improve the detection accuracy of COVID-19 rumors posted on Twitter using a publicly available dataset.

B. THE OUTLINE OF PAPER

The rest of the paper is organized as follows. The next section (section II) will lay down the background of misinformation in general, including the definition of specific types of misinformation. The chapter will also outline where rumor fits-in within the wider context. The next section (section III) will discuss related works in the domain and draw a comparative analysis between them. The subsequent sections (section IV, V, and VI) will present our research methodology including the details of the dataset, our research findings, and draw a conclusive discussion, respectively.

II. BACKGROUND AND MOTIVATION

In this section, we outline the background of misinformation and explain different terms that are associated with misinformation. We then attempt to highlight the devastating role a rumor can play during a world-wide health crisis. Finally, we present the scope of our research and list down the research questions.

A. DEFINITION OF MISINFORMATION

Within the context of OSN, the term *'Misinformation'* refers to false or inaccurate information, which is disseminated through any OSN platform. However, there is more to it when it comes to specific types of inaccurate information, as a range of terms are often associated with misinformation on OSN.

According to the information disorder framework [4], inaccuracy or disorder in information can be classified into three broad categories, misinformation, disinformation, and malinformation. The framework defined them as follows:

1) ***Misinformation*** is information that is false, but not created with the intention of causing harm.
2) ***Disinformation*** is information that is false and deliberately created to harm a person, social group, organization, or country.
3) ***Malinformation*** is information that is based on reality, used to inflict harm on a person, organization or country. This category of information disorder is not within our interest, as far as COVID-19 rumor detection is concerned.

Under these broad categories, there are several types of information disorder, which are generally referred to as misinformation on OSN. Some of the types, which are often studied and addressed in research in the domains of social network analysis and data science, are defined below. Parts of the classification of these misinformation types and their definition has been addressed in previous research works [5], [6], [7]. The list below synthesizes and summarises them and also aligns them with the broader categories of the information disorder framework. However, the list does not include information disorders such as hate speech, cyberbullying, defamation, etc., since they fall under malinformation within the broader category, which is outside the scope of this research.

- *Fake news:* A fake news is a piece of news article that is false and spread intentionally. It is a means of spreading a specific propaganda. The intention behind the spread of a Fake news may include the intention of causing harm. Therefore, an instance of fake news can be a misinformation or a disinformation.
- *Urban legend:* An urban legend is an intentionally-spread fictional story. It also falls in the category of misinformation since they are created for entertainment purposes and usually do not include an intention of causing harm.
- *Rumor:* A rumor is a piece of information, where the truthfulness of the information is doubtful or uncertain. Often, rumors are spread on OSNs with intention of causing deliberate chaos, which results in harm. Similar to fake news, a rumor can also be a misinformation or a disinformation.
- *Conspiracy:* Conspiracy theories are alternative and often dramatic explanations of events, as opposed to the

actual explanation. People who spread conspiracy theories often do so with the intention of causing harm, and thus, they are regarded as disinformation.

- *Astroturfing:* Astroturfing is the process of marking or faking grass-root opinion. Astroturfers mask the sponsors or organisations while spreading an opinion about an entity (product, person, service, etc.) to portray as though the opinion is coming from grass-root participants of the society. Astroturfing is regarded as disinformation, since there is an intention of causing financial, political, or social harm.
- *Crowdturfing:* Crowdturfing is also a popular term, which refers to astroturfing that is crowd-sourced. Similar to astroturfing, crowdturfing can be categorised as disinformation.

Figure 1 illustrates the different types of information disorder and where the fall within the context of misinformation and disinformation. The purpose of outlining the definition and categories of some of the misinformation terms in this sub-section is twofold. Firstly, the definitions help us establish a scope of our research, and secondly, they help us isolate some of the specific terms, which are relevant for inaccurate information surrounding COVID-19. For an emerging health related crisis, such as, COVID-19, the specific type of inaccurate or false information which is the most relevant is rumor. The closest type of false information that has a grey boundary with rumors are conspiracy theories. However, conspiracy theories are more relevant for political and social events. For the purposes of this study, we will not differentiate between these two types (rumor and conspiracy theory) and refer to all information of doubtful veracity related to COVID-19, as rumors.

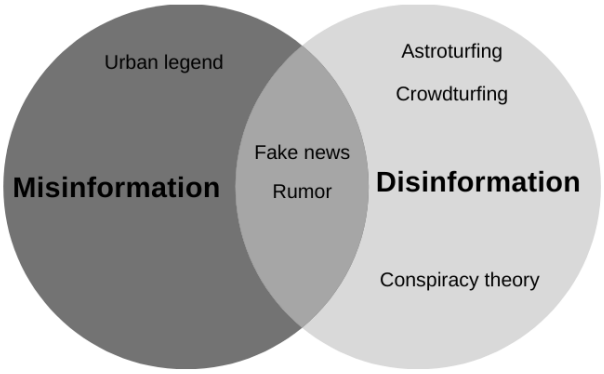

FIGURE 1. Information disorder.

B. ROLE OF RUMORS DURING A HEALTH CRISIS

During a health crisis, the importance of accurate information is critical. The policy makers and healthcare professionals are presented with the challenge of containing and managing the health crisis. In addition to that, people become anxious about the consequence of a disease, and worried about their friends and family. Under these circumstances, inaccurate information can have severe negative impact on the stability of the society. The impact becomes more severe when the crisis is as big as a pandemic.

As described in a Time news article [8], the rumors on social media during the outbreak of Ebola virus in 2014, created unnecessary panic, causing enormous waste of resources in the United States. The rumors included inaccurate information on how the disease spreads and how prevalent the disease was at that stage in a particular geographical area.

The impacts of rumors were far more severe at the early stages of world-wide COVID-19 outbreak. Studies [9], [10] reported how the use of cow dung and urine spread across India based on rumors on social media as a treatment for COVID-19.

An article [11] published in the Australian edition of 'The Conversation' outlines how based on internet rumors, people were calling the New South Wales Poison Information Centre to enquire about the benefits of inhaling hydrogen peroxide, gargling or swallowing antiseptics, bathing in bleach or disinfectant, and spraying face masks with disinfectants, in order to fight COVID-19. All of the above actions can have severe consequences on human health, including death.

There are several research works [12], [13], [14], [15], [16] that also reports rumors and their impacts, surrounding the treatment, spread patterns, do's and don'ts, potential home remedies, medicines, etc., related to COVID-19, where the rumors are essentially spread through online platforms and can cause severe damage to human health, or hinder the efforts of healthcare professionals to tackle the pandemic around the world.

More general forms of rumors and conspiracy theories are also reported to have destructive and disruptive impacts on the community. Research work surrounding the detection of COVID-19 conspiracy theory by Shahsavari, et al [17], reports the initiation of incidents including destruction of cell phone towers, racially fuelled attacks against Asians, demonstrations to defy public health orders, etc., based on rumors and conspiracy theories propagated through OSN platforms.

Therefore, it is absolutely critical to be able to develop a mechanism that can detect and prevent such rumors from spreading. To that end, in this study, we propose a model with hybrid feature space that makes use of psycho-linguistic features, in addition to content-based and contextual features, to improve the detection accuracy of COVID-19 rumors.

C. OUR SCOPE AND RESEARCH QUESTIONS

As discussed in the previous sub-sections, we focus on automated detection of COVID-19 rumors on Twitter. We treat all information where the truthfulness or veracity of the information is doubtful, as rumors. We also identify whether certain psycho-linguistic features, which may indicate dramatisation of written text, can improve the detection accuracy of such rumors.

In general, the purpose of this paper is to answer the following research questions:

- ***RQ1***. Is there a pattern of text dramatisation present in the COVID-19 rumor tweets?
- ***RQ2***. What are the psycho-linguistic features that can identify such dramatisation of text?
- ***RQ3***. Can these psycho-linguistic features, in addition to the content-based and contextual features, improve the detection accuracy of COVID-19 rumors on Twitter?

In this paper, we address these research questions and make novel contributions. Firstly, we propose a hybrid feature space, which includes tweet content-specific features, twitter-specific contextual features, and psycho-linguistic features. We also demonstrate how the psycho-linguistic features can improve the detection accuracy of COVID-19 rumors posted on Twitter using a publicly available dataset. The next section outlines the related work in the relevant domains of misinformation detection and draws a comparative analysis among works that are highly relevant to our research.

III. RELATED WORK

Since there are several areas of information disorder, as outlined in the previous section, we organise the discussion on related work based on these areas and focus on similar approaches of feature engineering. More specifically, we want to highlight what types of features have different research approaches used in the area of inaccurate information detection. We first present related approaches in the domains of fake news detection, followed by astroturfing and crowdturfing detection techniques. We then discuss works on general rumor detection. Finally, the section outlines related works on misinformation/disinformation detection specific for COVID-19 with a comparative analysis.

A. RELATED WORK IN THE DOMAIN OF FAKE NEWS DETECTION

The detection of fake news has been a trending research topic in recent years. There are several studies, who utilised data mining and feature engineering techniques to tackle the spread of online fake news. As outlined in several studies [18], [19], [20], [21], the features analysed by research in the area includes content-specific features, linguistic features, visual features, user-specific features, network-based features, source-based features, environmental features, etc.

The research work by Chen et al. [22], utilizes user sentiment and spread patterns of tweets on Twitter to detect fake news. Network-based spread patterns were also studies in several other research works [23], [24], [25] concentrated on fake news detection. The research work by Rashkin et al. [26], includes linguistic features to identify dramatisation of texts to detect fake news. Similar research works [27], [28], [29] have also investigated textual or linguistic features of the news content itself for fake news detection. The research work by Shu et al. [30] demonstrated how user profile specific features including their political bias, personality, etc., can play an important role in the detection of fake news. Additional user profile information such as user activity and social graph, user's content creation pattern, etc., have been utilised by similar research works [31], [32] in the domain of fake news detection.

B. RELATED WORK IN THE DOMAIN OF ASTROTURFING AND CROWDTURFING DETECTION

As outlined in previous study [33], approaches in the domain of astroturfing and crowdturfing detection can be very diverse. While approaches including authorship attribution [34], [35], [36] and analysis of network flow features [37], [38], [39] have been utilised in the area of astroturfing and crowdturfing detection, majority of automated detection approaches focus on identifying content-based or user-based features.

The research work by Cheng et al. [40] investigated comments on news portals to detect corporate astroturfing, where the authors considered similarity measures among comments, including features related to user activity and interaction time with contents. Analysis of content and user-based features in microblog environments, has also been carried out by other research works [41], [42], [43], [44], [45].

Features surrounding individuals or features applicable for a group of people have been studies for astroturfing and crowdturfing detection as well. Research [46], [47] demonstrates that group level features, such as, group time window, group deviation, group consent similarity, tweeting habits of a group, such as, postings of original tweets, retweeting someone's tweets who is not a friend, etc., reveal interesting information about crowdturfing groups in microblog environments, such as Twitter. Twitter-specific features such as, user profile and activity features (e.g., longevity of account, tweet steadiness, sparseness), network features (e.g., number of friends and followers), and personality features (e.g., tweet emotion), were also demonstrates to be very effective for crowdturfing detection [48].

C. RELATED WORK IN THE DOMAIN OF RUMOR DETECTION

In the domain of rumor detection, in general also includes heterogeneous approaches. In their research work, the authors Ma, et al. [49] developed two recursive neural networks. The top-down and the bottom-up tree-structured neural networks were proposed to track the propagation of general rumors in a microblog environment, such as, Twitter. The authors demonstrated that their tree-structured neural network is able to detect rumors at an early stage of propagation. Similar research works [50], [51] also propose neural network models that utilise multiple neural networks and investigate temporal, content, and propagation features to detect rumors on microblog platforms. The research work by Yang et al. [52] proposes a graph adversarial learning method to detect various strategies taken by perpetrators to camouflage rumors to bypass propagation-based detection methodologies. The work by Liu et al. [53] implements a structure-aware retweeting graph neural network, where the authors propose re-structuring of the retweet graph to align with binary tree

structure, without losing any propagation information. The authors also propose integration of content-based, user-based, and pattern-based features for improved rumor detection. The usefulness of time-series features, in addition to content-based, user-based, and lexical feature was demonstrated in similar research work by Shelke and Attar [54].

D. RELATED WORK IN THE DOMAIN OF COVID-19 MISINFORMATION OR DISINFORMATION DETECTION

Since the outbreak of COVID-19, several studies concentrated their efforts in automated detection of COVID-19 related information disorder on OSNs. Majority of the work focused on investigating contents on microblog environments, such as, Twitter, since the propagation speed and reach of such environment is higher as opposed to other classes of OSN platforms.

Research work by Moffitt et al. [55] addresses COVID-19 related conspiracy theory detection on Twitter, based on analysis of user-identities, their countries of origin, patterns of bot activities, and content of the tweet, such as, hashtag and URL analysis. The research work by Al-Rakhami and Al-Amri [56] utilises tweet content-based, tweet-based and user-based features and proposes an ensemble-learning-based framework for detection of non-credible tweets on Twitter.

The research work by Elhadad et al. [57] investigates the contents from reputable news sources and fact-checking authorities to generate ground truth data. The authors then investigate content-based features and apply multiple machine learning models for the detection of COVID-19 misleading information. The authors implement TF-IDF as the feature extraction technique on a bag-of-word model (BOW), that includes words of certain parts-of-speech tags, metadata including location-based, user-based, and time-based features, etc. A Similar research work by Al-Ahmad et al. [58], also uses TF, TF-IDF, and BOW models for feature extraction. The authors also proposed to reduce the number of symmetrical features by implementing wrapper feature selections for evolutionary classifications using particle swarm optimization (PSO), the genetic algorithm (GA), and the salp swarm algorithm (SSA). The authors utilised these features for detection of misleading information about COVID-19 using publicly available dataset of consist of 6,000 news articles, generated by a thesis [59].

The work by Hossain et al. [60] uses a corpus of Wikipedia misconceptions related to COVID-19 and classifies Twitter Tweets based on the support, deny, or neutral stance the tweet expresses in relation to the misconceptions. The research includes identification of the misconception instance that is related to a tweet, and then identification of the specific stance of the tweet towards the misconception. The research includes analysis of the contents of the tweets and misconception using several NLP techniques. A Similar research work by Vijjali et al. [61], analyses the contents of claims related to COVID-19 and computes the textual entailment between the claim and the true facts retrieved from a manually labelled COVID-19 dataset.

The work by Li et al. [62] proposes a multi-lingual and multi-dimensional COVID-19 fake news data repository. The authors collected 3981 pieces of fake news content and 7192 trustworthy information from 6 different languages, i.e., English, Spanish, Portuguese, Hindi, French, and Italian. The authors also demonstrate the reliability and robustness of the dataset by analyzing several features including social-interaction-based, tweet-based, user-based features.

Another research work by Memon and Carley [63], offers yet another COVID-19 misinformation dataset, including 4,573 annotated tweets. The authors also offer interesting insights into the characteristics of informed and misinformed groups of people. The authors performed network analysis, analysis of bots, and analysis of socio-linguistic features to differentiate between the two groups. Similar research work by Heidari et al. [64] applied the Bidirectional Encoder Representations from Transformers (BERT) on publicly available dataset. The authors used content-based, tweet-based, and user-based features for the classification of whether a tweet is generated by a bot or not. Their research concludes that the COVID-19 fake news is usually generated by human accounts, not bot accounts.

The research work by Cui and Lee [65], proposes a COVID-19 healthcare misinformation dataset that contains 4,251 news articles and 296,000 related user engagements. The authors also apply several machine learning models using different features including, content-based (text and image), sentiment-based, user-based, etc., to provide further insight into the classification of misinformation using the dataset. Related research work by Zhou at al. [66], offers another repository of COVID-19 misinformation, including fake news and conspiracies. The authors collected 2,029 news articles and 140,820 tweets, that circulated these news articles on Twitter. The authors also performed data analysis using textual, visual, temporal, and network features to provide a baseline model for future research. The dataset offered by the research work by Shahi and Nandini [67] includes 5,182 fact-checked news articles for COVID-19, across multiple language and countries. The news instances where crawled from fact-checking websites, i.e., Snopes and Poynter. The authors also applied NLP techniques for brief analysis of the data repository.

The research work by Paka, et al. [68], offers another public dataset for COVID-19 fake news detection. The authors propose a semi-supervised model, i.e., Cross-SEAN (cross-stitch based semi-supervised end-to-end neural attention model), and also an extension for the Chrome browser, i.e., Chrome-SEAN, which can automatically flag COVID-19 related fake news on Twitter. The authors utilised several textual and linguistic features (e.g., number of hashtags, number of user mentions, media count in the tweet, sentiment of the tweet text, counts of various part-of-speech tags, etc.), tweet-specific (e.g., number of hashtags, number of favourites, number of retweets, retweet status, etc.), and user-specific features (e.g., verified status, follower count,

TABLE 1. A comparative summary of COVID-19 mis/disinformation detection approaches with the proposed approach.

Research Work	Data Source	Classes of Features Used	Focused Mis/Disinformation	Publication Year
Moffitt, et al. [55]	Twitter	Content-based + Contextual	Conspiracy Theory	2021
Al-Rakhami and Al-Amri [56]	Twitter	Content-based + Contextual	Unspecified	2020
Elhadad, et al. [57]	WHO + UNICEF + UN + Multiple Fact Checking Website	Content-based + Contextual	Unspecified	2020
Al-Ahmad, et al. [58]	News Articles Across Multiple Media	Content-based	Fake News	2021
Hossain, et al. [60]	Wikipedia + Twitter	Content-based	Unspecified	2020
Vijjali, et al. [61]	Poynter	Content-based	Fake News	2020
Li, et al. [62]	Snopes + Poynter + Multiple Official Health Website + Twitter	Content-based + Contextual	Fake News	2020
Memon and Carley [63]	Twitter	Content-based + Contextual + Socio-linguistic	Unspecified	2020
Heidari, et al. [64]	Twitter	Content-based + Contextual	Fake News	2021
Cui and Lee [65]	Multiple News Media + Multiple Social Media	Content-based + Contextual	Unspecified	2020
Zhou, at al. [66]	NewsGuard and Media Bias/Fact Check + Twitter	Content-based + Contextual	Conspiracy Theory/Fake News	2020
Shahi and Nandini [67]	Snopes + Poynter	Content-based	Fake News	2020
Paka, et al. [68]	Twitter	Content-based + Contextual	Fake News	2021
Cheng, et al. [12]	Google Search Engine + Twitter	Content-based + Contextual	Rumor	2021
Proposed Work	Twitter	Content-based + Contextual + Psycho-linguistic	Rumor	-

favourites count, number of tweets, recent tweets per week, etc.).

The research work by Cheng et al. [12] also offers a comprehensive dataset for COVID-19 rumor detection. The dataset contains 4,129 news records and 2,705 tweets. The dataset is manually labelled and contains information about the veracity of the content, including stance and sentiment. The authors also performed deep learning based rumor classification on both the news and twitter dataset. The Twitter dataset contains contextual features related to the tweet, including Reply/Retweet/Like (RRL) numbers. Our research on COVID-19 rumor detection is based on the Twitter dataset offered by this particular research work.

Besides conventional feature engineering and machine learning approaches, there are other approach as well for detection of COVID-19 mis/disinformation. For example, the work by Shahsavari et al. [17], which was inspired by narrative theory, was also demonstrated to be effective in understanding the nature of COVID-19 rumors and conspiracy theories, where the authors attempt to understand the narration patterns of these conspiracy theories and identifies different clusters of communities, responsible for the propagation of these theories.

Finally, to put things into perspective, we present some of the highly relevant works in the domain of COVID-19 mis/disinformation detection in Table 1. Among other things, the table highlights the differences in the classes of features that have been used in these research works in comparison to the classes of features included in the features space of our proposed model. The table refers all features that are not content-specific, for example, tweet-specific features of retweet numbers, like numbers, etc., network-based features, user-based features, etc., as 'contextual features', since they add additional context to the content of the mis/disinformation. In addition, the table lists down 'Unspecified' under the focused area for the research works that do not specify a type and refer to mis/disinformation related to COVID-19 as, 'misinformation', 'misleading content', 'non-credible content', etc. As Table 1 outlines, our proposed feature space includes analysis of psycho-linguistic features, which makes the feature space a novel one. In the next section, we outline our research methodology, including the description of the dataset used, the feature space design, rationale behind including psycho-linguistic features, and experimental design.

IV. PROPOSED METHODOLOGY

In this section, we first outline the details of the COVID-19 rumor dataset. We then present the methodology of our feature space design. Finally, we conclude the section with the details of our experimental setup.

A. THE COVID-19 RUMOR DATASET

The COVID-19 rumor dataset is publicly available on GitHub.[1] The dataset is populated and released by research work by Cheng et al. [12]. The dataset contains two corpus of data, specific for news based rumors and tweet based rumors. The first corpus contains 4,129 instances of news records published in several news outlets. The corpus was generated using the Google Search Engine. The second corpus of data contains 2,705 tweets from Twitter. The Twitter data was crawled using COVID-19 related tags, for example, 'COVID-19', 'coronavirus', 'COVID', etc.

[1] https://github.com/MickeysClubhouse/COVID-19-Rumor-dataset

TABLE 2. Description of tweet related attributes in the twitter dataset.

Attribute	Example Content	Description
No.	1	The field is the serial number for unique identification of each tweet.
Label	T	This is the veracity label of the tweet and can have the following possible values: True (T) - The content is logical and describes a fact, e.g., 'The Trump administration fired the U.S. pandemic response team in 2018 to cut costs' False (F) - The content is made up, or contains false information, e.g., 'A homemade hand sanitizer made with Tito's Vodka can be used to fight the new coronavirus' Unverified (U) - The truthfulness of the content is unverified at the time of labelling.
Content	A major disease outbreak occurred in "every election year" since 2004	The actual content of the tweet.
Source	850196512341262000	The source of the tweet. The number is also used to identify the related comment file, which contains comments related to the particular tweet. The attributes of the comments files are further described in a separate table.
Sentiment	3	This is the sentiment label of the tweet and can have the following possible values: Very Negative (0) - The content has a strong pessimism. Negative (1) - The sentiment is pessimistic but weaker than "very strong". Neutral (2) - The content has a plain and narrative tone. Positive (3) - The content reflects positive sentiment, such as news providing tips to fight the virus. Very Positive (4) - The content contains cheerful news, such as, progress in research, or breakthrough in the vaccine research and production, etc.
Reply numbers	10	The number of replies of the tweet.
Retweet numbers	50	The number of retweets of the original tweet.
Like numbers	100	The number of likes for the tweet.

For the twitter dataset, the authors [12] not only collected the tweets themselves but also collected the associated comments and their metadata. The entire dataset was labelled by multiple human annotators, and each instance of news or tweet was labelled, True(T), False(F), and Unverified(U), based on their veracity. In addition to veracity, the dataset was also labelled based on the stance and sentiment of the tweet. For the purposes of our research, we only focused on the corpus generated from Twitter, since our focused area is microblog environments. Table 2 outlines all attributes of the main twitter dataset, including example cell value, possible values for discrete features, and description for each of them. For the sentiment label, the authors identified the sentiment of a tweet, manually. The authors also cross-checked the labelled sentiment with online sentiment analysis tool, MonkeyLearn,[2] and reported that the manual sentiment labels are more accurate, given the context of the tweets.

All tweets in the Twitter dataset with these attributes are listed in a CSV file named 'twitter.csv'. Furthermore, the dataset also contains all comments and associated information of comments, using the stance label, in a separate CSV file, for each instance of tweet in the 'twitter.csv' file. The specific CSV file containing the comments of a tweet can be traced using the source attribute value of a specific tweet. Table 3 lists down the attributes in a comment file including the description for each of the attributes. For each of the comments, the authors The stance value of the comments were labelled manually. The authors used the classical rumor stance classification, where the stance can be support, deny, comment, and query [69].

Given our task in this paper is to classify an instance of tweet from this rumor dataset into either a true or a false piece of information, we removed the tweets with label 'U', which are unverified rumors, where the veracity is unknown.

B. FEATURE SPACE DESIGN

In this sub-section, we focus on our feature selection approach and the process of generation of feature vectors from the COVID-19 rumor dataset. The feature vectors are then fed to multiple machine learning algorithms for the training of rumor classification task.

1) RATIONALE FOR SELECTING CONTENT-BASED AND CONTEXTUAL FEATURES

The first two categories of features that are crucial for identification of rumors on a microblog environment are the content-based and contextual features.

The content-based features are the features that are directly generated from the contents of the tweet themselves, and they have been demonstrated to possess characteristics, that can differentiate between a false or authentic piece of information. The use of content-based features are not only prominent in rumor detection [12], [55], [56], [57], [58], [60], [61], [62], [63], [64], [65], [66], [67], [68], but also in the domains of fake news [22], [27], [28], [29], astroturfing [41], [42], [43], [44], [45], or misleading information detection [70] in general. As demonstrated in these research works, the content-based features of a tweet, such as, the number of hashtags in a tweet, the number of smileys in a tweet, the

[2]https://monkeylearn.com/

TABLE 3. Description of comment related attributes in the twitter dataset.

Attribute	Example Content	Description
Twitter ID	9207569286939500000	The field refers to the Twitter ID of the original tweet.
Release date	Thu Jan 23 2020	This is the date on which the tweet was released on Twitter.
Comment	Take Note everyone! https://t.co/gwWGZuoBg0	An instance of comment associated with the tweet.
Time	Thu Jan 23 13:41:17 +0000 2020	The actual time the comment was made on the original tweet.
Replies	3	The number of replies to the actual comment.
Retweets	12	The number of retweets of the comment.
Likes	12	The number of likes for the comment.
Stance	Comment	This is the stance label of the comment and can have the following values: Support - Positive attitude about the content, e.g., 'I think the statement is right.' Deny - Denying attitude towards the content, e.g., 'Are you Kidding? This is wrong!' Comment - No obvious stance towards the content, e.g., 'This message is interesting' Query - Doubting the validity of the tweet, e.g., 'Is this true?'

length of the tweet itself, and so on, can have significant impact on the identification of false information.

Contextual features on the other hand are features that add more context information to the content in these microblog environments. For example, the tweet-specific features, such as, the popularity of a tweet, including number of likes, retweets, and replies, the acceptance, rejection, or indifference of an opinion expressed in a comment of a tweet, etc. These tweet-specific contextual features have been used consistently in several past research works [31], [32], [62], [63], [64], [65], [66] for identification of mis/disinformation on microblog environments. The list of all content-based and contextual features that have been used in our research are listed in section IV-B4.

2) RATIONALE FOR SELECTING PSYCHO-LINGUISTIC FEATURES

Psycho-linguistic features refer to features that are generated by analysing the language used in the content, with the goal of identifying the psychological and emotional profile of the user, who is responsible for creating the content. Using a set of psycho-linguistic features, we can understand the mindset of a certain group of people, which can be critical for identifying groups that are responsible for the propagation of misleading information on OSN platforms.

For example, previous research work by Ott, et al. [71] demonstrates that the use of first-person and second-person pronouns are indicative features for imaginative writing, where a person writes something with or without harmful intentions, which is far from the fact. The authors [71] also reported the use of superlative and comparative words in imaginative writing. These characteristics of imaginative writing have been utilised and corroborated in research work by Rashkin, et al. [26], where the authors used several linguistic features for the detection of fake news. The authors demonstrated that subjective words are used often to dramatize or sensationalize a news story. The authors also associate the use of action adverbs and manner adverbs with the dramatisation of written text for the purposes of attracting readers. The research [26] also reports that fake news articles use more swear words, subjective words, superlatives, and modal adverbs, which are used to exaggerate a piece of news. Several research works [72], [73], [74], reports the usage of hedge words for identification of vagueness and uncertainty in written language, which are also found to be relevant in detection of fake news [26].

The use of psycho-linguistic features was also demonstrated to be effective for detection of conspiracy propagators [6], where the authors investigates personality, sentiment, emotions, and linguistic patterns of the users to identify propagation of conspiracy theories. The authors reports that anti-conspiracy propagators express more emotions in their tweets compared with conspiracy propagators. Similar research [75] also demonstrates how the analysis of emotions can help with the identification of fake news in social media. Moreover, dramatic events, such as, natural disasters, mass outbreaks of diseases, etc., tends to start waves of online discussions, which have certain characteristics. Research [76] demonstrates that discussions happening in an online platform following a dramatic event exhibit sign of emotional shock, increased language complexity, and simultaneous expressions of certainty and doubtfulness, which provides insight into how spread of conspiracy theories can be identified during the escalation of an event.

Based on these research work, it is coherent that a certain degree of text dramatization and exaggeration is directly associated with the spread of mis/disinformation on OSN. Furthermore, the study of sentiments and emotions that are expressed in these mis/disinformation can also add value to the detection strategies. Therefore, our research aims to address whether these text dramatization, sentiment, and emotion markers are also relevant for COVID-19 rumor detection and hence, studies a set of psycho-linguistic features.

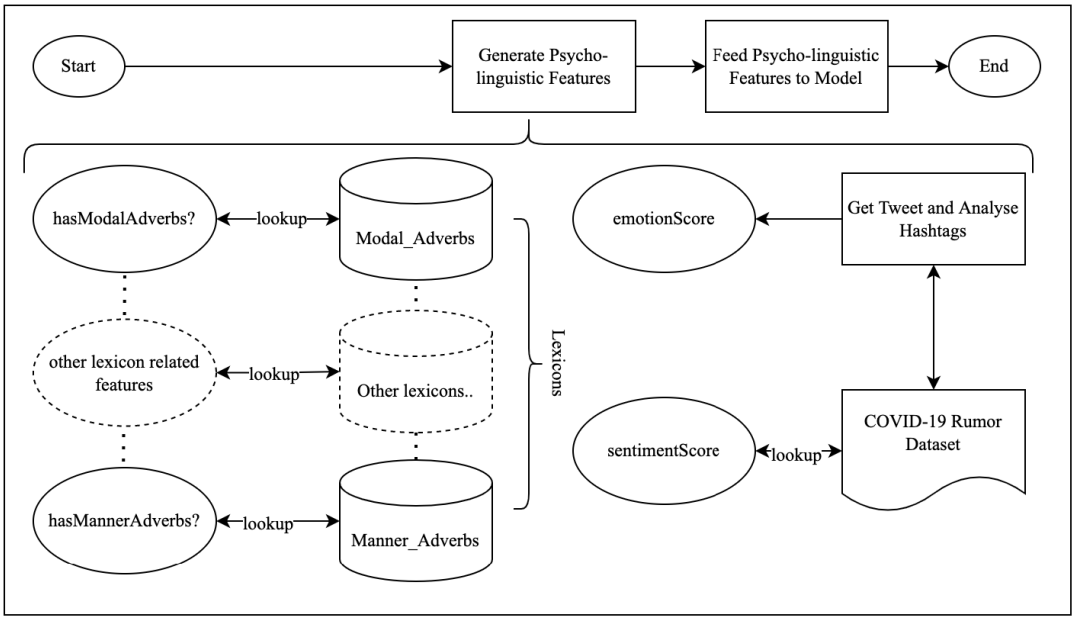

FIGURE 2. Psycho-linguistic feature generation process.

3) THE GENERATION OF PSYCHO-LINGUISTIC FEATURES

For the generation of the psycho-linguistic features, we have used several lexicons. The swear words dictionary was generated from the Noswearing[3] website. The dictionaries of modal adverbs, action adverbs, manner adverbs, comparatives, and superlatives, were downloaded from public repository, made available by Rashkin, et al. [26]. The authors compiled these dictionaries from Wiktionary[4] word lists. The dictionary of hedge words was compiled manually from previous research work by Yüksel and Kavanoz [77]. Using these dictionaries, we generate the a list of binary psycho-linguistic features based on the contents of the tweets from the Twitter corpus of the COVID-19 rumor dataset.

We also include expressed sentiment and emotion as part of the psycho-linguistic feature set. The sentiment information was extracted directly from the Twitter corpus of the COVID-19 rumor dataset. For analysis of emotion, we used the emotion detection model, EMOTEX, proposed by Hasan, et al. [78]. According to this model, emotions can be categorised into four general classes. The classes and some of the emotions that falls under each of these classes are listed below:

- Happy-Active: Happy, Excited, Delighted, Astonished, Aroused
- Happy-Inactive: Serene, Contented, Satisfied, Relaxed, Calm
- Unhappy-Active: Tense, Angry, Afraid, Annoyed, Distressed
- Unhappy-Inactive: Miserable, Depressed, Sad, Gloomy

[3]https://www.noswearing.com/dictionary
[4]https://en.wiktionary.org/

We analyse the hashtags in the tweets themselves in order to associate an instance of tweet with any one of these four classes. Figure 2 illustrates our process of psycho-linguistic feature generation, graphically. A complete list of psycho-linguistic features used as part of our research has been given in the next section IV-B4.

4) SUMMARY OF FEATURE SPACE

As discussed in the previous sections, our feature space contains three categories of features:

- Content-based, which are based on the content of the tweets
- Contextual, which are tweet-specific features and provides additional contextual information to the tweets, and,
- Psycho-linguistic features, which are linguistic features to better understand the psychology and emotions of a person responsible for creating a microblog content

Table 4 lists down the categories of features, including individual features for each category, in a summarised manner.

C. THE EXPERIMENTAL SETUP

Our experimental setup is focused surrounding identifying the answers to our research questions. We want to investigate how efficiently a false rumor can be detected using:

- Only psycho-linguistic features
- Only content-based and contextual features, and
- The combination of psycho-linguistic, content-based, and contextual features

Our aim is to see the differences in performance measures of different machine learning algorithm, for different feature space setup. Table 5 lists down the three separate

TABLE 4. Summary of feature space.

Feature Class	Features
Content-based Features	noOfSmiliesInTweet
	noOfQuestionMarksInTweet
	lengthOfTweet
	noOfSmiliesInComments
	noOfMentionsInComments
	noOfHashTagsInComments
	presenceOfURLsInComments
	noOfQuestionMarksInComments
	lengthOfComments
Contextual Features	noOfLikesTweet
	noOfCommentsTweet
	noOfRetweets
	noOfRepliesTweet
	noOfLikesComments
	noOfRetweetsOfComments
	noOfRepliesOfComments
	noOfSupportStanceOfTweet
	noOfDenyStanceOfTweet
	noOfCommentStanceOfTweet
	noOfQueryStanceOfTweet
Psycho-linguistic Features	hasSwearWords
	hasModalAdverb
	hasActionAdverb
	has1stPersonPronoun
	has2ndPersonPronoun
	hasMannerAdverb
	hasHedgeWords
	hasSuperlatives
	hasComparatives
	sentimentScore
	emotionScore

TABLE 5. Experimentation setups based on feature space.

Experiment Label	Feature Space Formation
A	Psycho-linguistic Features Only
B	Content-based + Contextual Features
C	Content-based + Contextual Features + Psycho-linguistic Features

experimentation labels and the corresponding feature space design. We assign these labels to be able to differentiate, refer to, and discuss the different feature classes more clearly.

For machine learning purposes, we use WEKA [79], a software suite, that supports data analysis and mining using multiple machine learning algorithms. For our research, we have used J48 Decision Tree, Random Forest, Naive Bayes, and JRip Rule-based classifiers. The purpose of using multiple machine learning algorithm was to check the consistency of acquired knowledge.

We conducted all three of the above experiments (A, B, and C), using each of the 4 classifiers. In the next section, we present our experimental findings, for all combinations of experimental label and classifier, and draw a conclusive discussion on our findings.

V. RESULTS AND DISCUSSION

In this section, we outline our experimental findings and discuss a few aspects of the observed results. The analysis of the findings is divided into several subsections to facilitate the discussion. First, we present our experimental findings for experiments A, B, and C, including a discussion on the observed performance measures. We then analyze the psycho-linguistic features further and offer additional insights into their distributions and prevalence in the dataset.

A. EXPERIMENTAL RESULTS ACROSS DIFFERENT SETUPS

Table 6 outlines our experimental results. For each of the combination of a machine learning algorithm and a labelled experimental setup, we recorded the accuracy, precision, recall, and f-measure as the performance matrices for the model.

As outlined in Table 6, the trend of differences in the observed performance measures for the three different experiment labels, is quite consistent across all four machine learning algorithms. The Naive Bayes classifier demonstrates the least performance, whereas the Random Forest classifier performs the best. However, irrespective of their differences in performance, all measures across different classifiers consistently support the trend of improved performance when psycho-linguistic features are included in the features space, in addition to content-based and contextual features. The measures also demonstrate that the psycho-linguistic features alone cannot provide the same level of performance as the feature set of content-based and contextual features. This is an expected behavior, since, the set of psycho-linguistic features focus only on one aspect of the content, i.e., the linguistic features that reveals information about the psychology or emotion of the content creator. However, as demonstrated in the related work section, all previous works in the domain (summarised in Table 1) have used the content-based features as the base feature set, since they convey the most information in terms of how a false information is different from an accurate information. Therefore, without considering the content-based or contextual features, it is not feasible to accurately classify an instance of information as false or true.

Nevertheless, for the purposes of our discussion, the area of the observed results that we would like to focus on is the difference in performance measures between experiment B and experiment C. The only difference between experiments B and C is the psycho-linguistic feature set. Experiment B was conducted by using the content-based and contextual features, whereas, experiment C was conducted using content-based, contextual, and psycho-linguistic features. As we can see from the observed results in Table 6, the performance measures for all four classifiers show consistent improvement in experiment C, compared to the measures in experiment B. For example, the accuracy of the J48 decision tree classifier, is increased to 77.80% in experiment C, from 76.34% in experiment B. The values of precision, recall, and f-measures are also increased to 0.768, 0.800, and 0.784, respectively, in experiment C, from the observed values of 0.767, 0.760, and 0.763, respectively, in experiment B. Even for the classifier with the least performance (Naive Bayes), in experiment C, the accuracy, precision, recall, and f-measure values

TABLE 6. Summary of experimental results.

Algorithms	Experiment A (Psycho-linguistic Features Only)				Experiment B (Content-based and Contextual Features)				Experiment C (Content-based + Contextual + Psycho-linguistic features)			
	Accuracy(%)	Precision	Recall	F-Measure	Accuracy(%)	Precision	Recall	F-Measure	Accuracy(%)	Precision	Recall	F-Measure
J48	63.77	0.692	0.502	0.582	76.34	0.767	0.760	0.763	77.80	0.768	0.800	0.784
JRip	62.18	0.664	0.499	0.570	73.98	0.731	0.761	0.746	75.50	0.740	0.787	0.763
Random Forest	66.65	0.738	0.520	0.610	80.86	0.775	0.871	0.820	81.00	0.775	0.871	0.820
Naive Bayes	62.41	0.674	0.487	0.565	61.30	0.676	0.441	0.534	66.59	0.705	0.575	0.634

are observed to have a higher value of 66.59%, 0.705, 0.575, and 0.634, respectively, in comparison to the respective values of 61.30%, 0.676, 0.441, and 0.534, in experiment B.

B. THE DISTRIBUTIONS OF PSYCHO-LINGUISTIC FEATURES

In order to gain more insight into the distributions and prevalence of the psycho-linguistic features within the dataset, we conducted further analysis of the feature space. We illustrate a summary of our findings in Figure 3. The first sub-illustration(a) shows that among the false rumors, majority express a positive sentiment, and the expressed sentiments among false rumors are neutral, positive, and very positive. This distribution of sentiments among false rumors aligns with our initial rationale of selecting sentiment as part of the psycho-linguistic feature set, where we discussed the patterns of content creators to sensationalize or enliven a story, during the spread of false rumors. An ecstatic expression of sentiment (positive and very positive) in majority of false rumors is an affirmation of our initial understanding.

The second sub-illustration(b) in Figure 3, shows a different side of rumors. The distribution of emotions among false rumor instances, shows that the majority of emotions expressed through hashtags in the false rumor tweets represents the emotion class of unhappy-active. This particular emotion class is an overarching representation of granular emotions, such as, 'Tense', 'Angry', 'Afraid', 'Annoyed', 'Distressed'. This distribution is logical on its own, since the most prominent emotions expressed in the tweets related to COVID-19 should reflect fear, anger, distress, etc. The other types of emotions that are often observed in false rumors equally are happy-active and happy-inactive, which represents similar tones of positive sentiments, as observed in the sentiment distribution. It is worth noting here that the EMOTEX [78] model for emotion analysis relies on the existence of hashtags associated with a tweet. The model clusters a group of hashtags under one specific emotion class and does not perform analysis of the tweet content, whereas the sentiment analysis in the COVID-19 rumor dataset has been performed by analysing the contents of the tweets. Therefore, the distribution of sentiments and the distribution of the emotions may not align categorically. In future extension, we plan to look into more advanced approaches of emotion analysis, such as the NRC emotion lexicon [80].

The third sub-illustration(c) in Figure 3, demonstrates the percentages of false rumors containing different categories of words. From the bar-chart, we can see that 1 in every 4 false rumor instances contains a hedge word, 1 in every 10 rumor instances contains a superlative, and 1 in every 10 rumor instances contains a comparative word. The other linguistic markers observed among false rumors include 1st person pronouns, 2nd person pronouns, manner adverbs, and modal adverbs. The observations confirm the presence of linguistic markers that can represent dramatisation or vagueness of text among COVID-19 false rumors, and further justifies the design of our feature space.

C. A SUMMARY OF OBSERVATIONS

Based on the discussion in the previous two subsections, we now draw a conclusive summary of observations and associate the observations with our framed research questions.

- The observed results across all experimental setups clearly demonstrate how inclusion of the psycho-linguistic features can improve the detection performance of COVID-19 rumors in real-life dataset. The observed measures in experiment C, which included the psycho-linguistic features in the feature space, are consistently higher compared to the observed measures in experiment B, which did not include the psycho-linguistic features. This inference answers our third research question (RQ3).
- As outlined in our rationale for selecting psycho-linguistic features in section (IV-B2), there are certain linguistic markers that are often observed in dramatized texts. Our analysis of the psycho-linguistic features reveals that these linguistic markers are indeed present in the COVID-19 rumor tweets. The presence of these markers corroborates the fact that the false rumor instances in the COVID-19 rumor dataset exhibit a certain degree of text dramatisation for enlivening the content or attracting attention of the readers. This deduction answers our first research question (RQ1).
- Our analysis of the psycho-linguistic features further offers insight into the distributions and prevalence of specific features within the feature space. From the analysis, we can observe that certain features (such as, the presence of hedge words) are more prevalent in rumors, among the set of linguistic markers. We can also observe clear patterns of sentiment and emotion distributions among rumors. These observations provide insight into the features that can identify the dramatisation in texts more clearly and answers our second research question (RQ2).

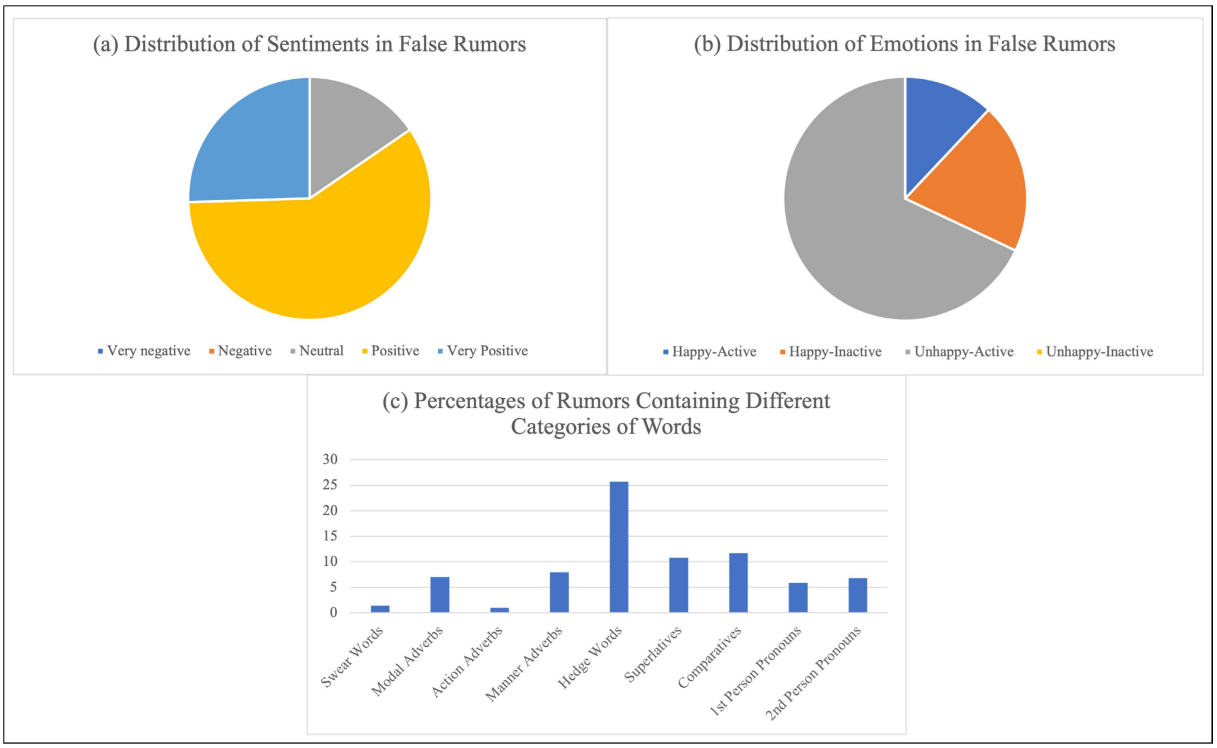

FIGURE 3. Analysis of the psycho-linguistic features. (a) Distribution of sentiments in false rumors (b) Distribution of emotions in false rumors (c) Percentages of rumors containing different categories of words.

VI. CONCLUSION AND FUTURE DIRECTIONS

The aftermath of COVID-19 will continue to have an impact on the world for many years to come. The success of ongoing management and recovery efforts heavily rely on the progress of research in the domain of health science. However, part of the success, also rely on the accuracy of information that is available online. An online mis/disinformation, which is related to the pandemic can have devastating effects on the health of the patients, the well-being of vulnerable groups in the society, and the efforts of the healthcare and government professionals. In this paper, we address this concerning issue surrounding the spread of COVID-19 related mis/disinformation on OSN. We propose a novel feature space for the detection of COVID-19 rumors and signify the effectiveness of the proposed feature space using real-life dataset. Our research demonstrates that a set of psycho-linguistic features, that reveals interesting information about the psychology and emotion of the content creator, can also provide insight into the veracity of the content. Our proposed model can be used in the back-end of an OSN platforms for the detection and prevention of the spread of such mis/disinformation. The model can also be used for detection of rumors related to any social topic, as the features themselves are not specific to COVID-19 pandemic.

The future extension of this work may include further study into the behavioural and psychological characteristics of the OSN users, which may help with the expansion of psycho-linguistic feature set, consequently, increase the performance measures of the detection model. Additional linguistic features that can identify text dramatization or vagueness, more accurately can be included in the future model, which requires further study of the linguistics in general. Future studies can apply more NLP techniques to further investigate the patterns of linguistic markers in rumors. Future studies may also include detection strategies suitable for other languages, since a large number of OSN users do not use English as the language of written communication. There is also scope for improvement in the model, by including a medical knowledge base as part of the model, built around medical facts, which may help debunk some of the false rumors straight away. However, such directions must include collaboration across multiple disciplines.

REFERENCES

[1] Google News. (2022). *Coronavirus (COVID-19)*. Accessed: Aug. 15, 2022. [Online]. Available: shorturl.at/cgikx

[2] Twitter Australia. (2021). *Australians Have Turned to Twitter to Stay Connected During the Pandemic*. Accessed: Aug. 15, 2022. [Online]. Available: shorturl.at/aJNRT

[3] John-Paul Ford Rojas. (2020). *Coronavirus: Lockdowns Drive Record Growth in Twitter Usage*. Accessed: Aug. 15, 2022. [Online]. Available: shorturl.at/ptY12

[4] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policymaking," Council Eur., Strasbourg, France, Tech. Rep. DGI(2017)09, 2017.

[5] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, "Misinformation in social media: Definition, manipulation, and detection," *ACM SIGKDD Explor. Newslett.*, vol. 21, no. 2, pp. 80–90, Nov. 2019.

[6] A. Giachanou, B. Ghanem, and P. Rosso, "Detection of conspiracy propagators using psycho-linguistic characteristics," *J. Inf. Sci.*, Jan. 2021, Art. no. 0165551520985486, doi: 10.1177/0165551520985486.

[7] B. Collins, D. T. Hoang, N. T. Nguyen, and D. Hwang, "Fake news types and detection models on social media a state-of-the-art survey," in *Proc. Asian Conf. Intell. Inf. Database Syst.* Cham, Switzerland: Springer, 2020, pp. 562–573.
[8] V. Luckerson. (2014). Accessed: Jul. 20, 2022. *Fear, Misinformation, and Social Media Complicate EBOLA Fight*. [Online]. Available: https://time.com/3479254/ebola-social-media/,
[9] S. Daria and M. R. Islam, "The use of cow dung and urine to cure COVID-19 in India: A public health concern," *Int. J. Health Planning Manag.*, vol. 36, no. 5, pp. 1950–1952, Sep. 2021.
[10] M. Y. Essar, S. K. Kazmi, M. M. Hasan, A. C. D. S. Costa, and S. Ahmad, "The rampant use of cow dung to treat COVID-19: Is India at the brink of a zoonotic disease outbreak?" *J. Med. Virology*, vol. 93, no. 12, pp. 6471–6473, Dec. 2021.
[11] (2021). *Darren Roberts and Nicole Wright, People Want to Use Bleach and Antiseptic for COVID and are Calling us For Advice*. Accessed: Jul. 25, 2022. [Online]. Available: https://theconversation.com/people-want-to-use-bleach-and-antiseptic-for-covid-and-are-calling-us-for-advice-168660
[12] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, and P. Bogdan, "A COVID-19 rumor dataset," *Frontiers Psychol.*, vol. 12, May 2021, Art. no. 644801.
[13] S. Tasnim, M. M. Hossain, and H. Mazumder, "Impact of rumors and misinformation on COVID-19 in social media," *J. Preventive Med. Public Health*, vol. 53, no. 3, pp. 171–174, May 2020.
[14] W. Zou and L. Tang, "What do we believe in? Rumors and processing strategies during the COVID-19 outbreak in China," *Public Understand. Sci.*, vol. 30, no. 2, pp. 153–168, Feb. 2021.
[15] Z. Hou, F. Du, X. Zhou, H. Jiang, S. Martin, H. Larson, and L. Lin, "Cross-country comparison of public awareness, rumors, and behavioral responses to the COVID-19 epidemic: Infodemiology study," *J. Med. Internet Res.*, vol. 22, no. 8, Aug. 2020, Art. no. e21143.
[16] A. Abdoli, "Gossip, rumors, and the COVID-19 crisis," *Disaster Med. Public Health Preparedness*, vol. 14, no. 4, pp. 29–30, Aug. 2020.
[17] S. Shahsavari, P. Holur, T. Wang, T. R. Tangherlini, and V. Roychowdhury, "Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news," *J. Comput. Social Sci.*, vol. 3, no. 2, pp. 279–317, Nov. 2020.
[18] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.
[19] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manag.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025.
[20] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–40, 2020.
[21] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised learning for fake news detection," *IEEE Intell. Syst.*, vol. 34, no. 2, pp. 76–81, Mar./Apr. 2019.
[22] W. Chen, C. Yang, G. Cheng, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, "Exploiting behavioral differences to detect fake ne," in *Proc. 9th IEEE Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Nov. 2018, pp. 879–884.
[23] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD Explor. Newslett.*, vol. 21, no. 2, pp. 48–60, 2019.
[24] C. Song, K. Shu, and B. Wu, "Temporally evolving graph neural network for fake news detection," *Inf. Process. Manag.*, vol. 58, no. 6, Nov. 2021, Art. no. 102712.
[25] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 14, 2020, pp. 626–637.
[26] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2931–2937.
[27] D. P. Kasseropoulos and C. Tjortjis, "An approach utilizing linguistic features for fake news detection," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.* Cham, Switzerland: Springer, 2021, pp. 646–658.
[28] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification," *Exp. Syst. Appl.*, vol. 169, May 2021, Art. no. 114171.
[29] M. S. Espinosa, R. Centeno, and R. Odrigo, "Analyzing user profiles for detection of fake news spreaders on Twitter," in *Proc. CLEF*, 2020, pp. 1–12.
[30] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, "The role of user profiles for fake news detection," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2019, pp. 436–439.
[31] T. Hamdi, H. Slimi, I. Bounhas, and Y. Slimani, "A hybrid approach for fake news detection in Twitter based on user features and graph embedding," in *Proc. Int. Conf. Distrib. Comput. Internet Technol.* Cham, Switzerland: Springer, 2020, pp. 266–280.
[32] K. Shu, S. Wang, and H. Liu, "Understanding user profiles on social media for fake news detection," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 430–435.
[33] S. Mahbub, E. Pardede, A. Kayes, and W. Rahayu, "Controlling astroturfing on the internet: A survey on detection techniques and research challenges," *Int. J. Web Grid Services*, vol. 15, no. 2, pp. 139–158, 2019.
[34] J. Peng, R. K.-K. Choo, and H. Ashman, "Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, Aug. 2016, pp. 121–128.
[35] J. Peng, K.-K.-R. Choo, and H. Ashman, "Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles," *J. Netw. Comput. Appl.*, vol. 70, pp. 171–182, Jul. 2016.
[36] J. Peng, S. Detchon, K.-K.-R. Choo, and H. Ashman, "Astroturfing detection in social media: A binary n-gram-based approach," *Concurrency Comput., Pract. Exper.*, vol. 29, no. 17, p. e4013, Sep. 2017.
[37] J. Ratkiewicz, M. Conover, M. Meiss, B. Goncalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in *Proc. Int. AAAI Conf. Web Social Media*, 2011, vol. 5, no. 1, pp. 297–304.
[38] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Truthy: Mapping the spread of astroturf in microblog streams," in *Proc. 20th Int. Conf. Companion World Wide Web*, 2011, pp. 249–252.
[39] Y. Liu, Y. Liu, M. Zhang, and S. Ma, "Pay me and i'll follow you: Detection of crowdturfing following activities in microblog environment," in *Proc. IJCAI*, vol. 16, 2016, pp. 3789–3796.
[40] C. Chen, K. Wu, V. Srinivasan, and X. Zhang, "Battling the internet water army: Detection of hidden paid posters," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2013, pp. 116–120.
[41] K. Lee, P. Tamilarasan, and J. Caverlee, "Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 7, pp. 331–340, 2013.
[42] B. Miller, "Automated detection of Chinese government astroturfers using network and social metadata," 2016, doi: 10.2139/ssrn.2738325.
[43] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated crowdturfing attacks and defenses in online review systems," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 1143–1158.
[44] J. Song, S. Lee, and J. Kim, "CrowdTarget: Target-based detection of crowdturfing in online social networks," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 793–804.
[45] G. Wang, T. Wang, H. Zheng, and B. Y. Zhao, "Man vs. Machine: Practical adversarial detection of malicious crowdsourcing workers," in *Proc. 23rd USENIX Secur. Symp.*, 2014, pp. 239–254.
[46] C. Xu, J. Zhang, K. Chang, and C. Long, "Uncovering collusive spammers in Chinese review websites," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manag. (CIKM)*, 2013, pp. 979–988.
[47] X. Wang, B. Zhou, Y. Jia, and S. Li, "Detecting internet hidden paid posters based on group and individual characteristics," in *Proc. Int. Conf. Web Inf. Syst. Eng.* Cham, Switzerland: Springer, 2015, pp. 109–123.
[48] K. Lee, S. Webb, and H. Ge, "Characterizing and automatically detecting crowdturfing in Fiverr and Twitter," *Social Netw. Anal. Mining*, vol. 5, no. 1, pp. 1–16, Dec. 2015.
[49] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on Twitter with tree-structured recursive neural networks," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1–10.
[50] S. Srinivasan and D. Babu, "A parallel neural network approach for faster rumor identification in online social networks," *Int. J. Semantic Web Inf. Syst.*, vol. 15, no. 4, pp. 69–89, Oct. 2019.
[51] S. Santhoshkumar and L. D. D. Babu, "Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks," *Social Netw. Anal. Mining*, vol. 10, no. 1, pp. 1–17, Dec. 2020.
[52] X. Yang, Y. Lyu, T. Tian, Y. Liu, Y. Liu, and X. Zhang, "Rumor detection on social media with graph structured adversarial learning," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1417–1423.

[53] B. Liu, X. Sun, Q. Meng, X. Yang, Y. Lee, J. Cao, J. Luo, and R. K.-W. Lee, "Nowhere to hide: Online rumor detection based on retweeting graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 6, 2022, doi: 10.1109/TNNLS.2022.3161697.
[54] S. Shelke and V. Attar, "Rumor detection in social network based on user, content and lexical features," *Multimedia Tools Appl.*, vol. 81, no. 12, pp. 17347–17368, May 2022.
[55] J. Moffitt, C. King, and K. M. Carley, "Hunting conspiracy theories during the COVID-19 pandemic," *Social Media Soc.*, vol. 7, no. 3, 2021, Art. no. 20563051211043212.
[56] M. S. Al-Rakhami and A. M. Al-Amri, "Lies kill, facts save: Detecting COVID-19 misinformation in Twitter," *IEEE Access*, vol. 8, pp. 155961–155970, 2020.
[57] M. K. Elhadad, K. F. Li, and F. Gebali, "Detecting misleading information on COVID-19," *IEEE Access*, vol. 8, pp. 165201–165215, 2020.
[58] B. Al-Ahmad, A. M. Al-Zoubi, R. A. Khurma, and I. Aljarah, "An evolutionary fake news detection method for COVID-19 pandemic information," *Symmetry*, vol. 13, no. 6, p. 1091, Jun. 2021.
[59] A. Koirala, "COVID-19 fake news classification with deep learning," M.S. thesis, Asian Inst. Technol., Bangkok, Thailand, 2020.
[60] T. Hossain, R. L. Logan, A. Ugarte, Y. Matsubara, S. Young, and S. Singh, "COVIDLies: Detecting COVID-19 misinformation on social media," in *Proc. 1st Workshop NLP*, 2020, pp. 1–11.
[61] R. Vijjali, P. Potluri, S. Kumar, and S. Teki, "Two stage transformer model for COVID-19 fake news detection and fact checking," in *Proc. 3rd NLP IF Workshop NLP Internet Freedom, Censorship, Disinformation, Propaganda*, 2020, pp. 1–10.
[62] Y. Li, B. Jiang, K. Shu, and H. Liu, "MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation," 2020, *arXiv:2011.04088*.
[63] S. Ali Memon and K. M. Carley, "Characterizing COVID-19 misinformation communities using a novel Twitter dataset," 2020, *arXiv:2008.00791*.
[64] M. Heidari, S. Zad, P. Hajibabaee, M. Malekzadeh, S. HekmatiAthar, O. Uzuner, and J. H. Jones, "BERT model for fake news detection based on social bot activities in the COVID-19 pandemic," in *Proc. IEEE 12th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Dec. 2021, pp. 0103–0109.
[65] L. Cui and D. Lee, "CoAID: COVID-19 healthcare misinformation dataset," 2020, *arXiv:2006.00885*.
[66] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "ReCOVery: A multimodal repository for COVID-19 news credibility research," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2020, pp. 3205–3212.
[67] G. K. Shahi and D. Nandini, "FakeCOVID—A multilingual cross-domain fact check news dataset for COVID-19," 2020, *arXiv:2006.11343*.
[68] W. S. Paka, R. Bansal, A. Kaushik, S. Sengupta, and T. Chakraborty, "Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection," *Appl. Soft Comput.*, vol. 107, Aug. 2021, Art. no. 107393.
[69] M. Cheng, S. Nazarian, and P. Bogdan, "VRoC: Variational autoencoder-aided multi-task rumor classifier based on text," in *Proc. Web Conf.*, Apr. 2020, pp. 2892–2898.
[70] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris, "Detection and visualization of misleading content on Twitter," *Int. J. Multimedia Inf. Retr.*, vol. 7, no. 1, pp. 71–86, Mar. 2018.
[71] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," 2011, *arXiv:1107.4557*.
[72] D. B. Buller and J. K. Burgoon, "Interpersonal deception theory," *Commun. Theory*, vol. 6, no. 3, pp. 203–242, Aug. 1996.
[73] E. Choi, C. Tan, L. Lee, C. Danescu-Niculescu-Mizil, and J. Spindel, "Hedge detection as a lens on framing in the GMO debates: A position paper," 2012, *arXiv:1206.1066*.
[74] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2013, pp. 1650–1659.
[75] B. Ghanem, P. Rosso, and F. Rangel, "An emotional analysis of false information in social media and news articles," *ACM Trans. Internet Technol.*, vol. 20, no. 2, pp. 1–18, May 2020.
[76] M. Samory and T. Mitra, "Conspiracies online: User discussions in a conspiracy community following dramatic events," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 12, 2018, pp. 1–10.
[77] H. G. Yüksel and S. Kavanoz, "Expressing claim: Hedges in English language learners' writing," *J. Teach. Educ.*, vol. 4, no. 1, pp. 263–269, 2015.
[78] M. Hasan, E. Rundensteiner, and E. Agu, "Emotex: Detecting emotions in Twitter messages," in *Proc. Socialcom/Cybersecurity Conf.*, 2014, pp. 1–10.
[79] F. Eibe, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA, USA: Morgan Kaufmann, 2016.
[80] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using mechanical Turk to create an emotion lexicon," in *Proc. NAACL HLT Workshop Comput. Approaches Anal. Gener. Emotion Text*, 2010, pp. 26–34.

SYED MAHBUB received the bachelor's degree in computer science and engineering from the Bangladesh University of Engineering and Technology, in 2012, and the master's degree in information technology from La Trobe University, in 2016. He worked in the industry as a Software Engineer for about three years after completing his bachelor's degree. He is currently finalizing his Ph.D. thesis on social network analysis and feature engineering. He has been working with the Department of CS & IT as an Associate Lecturer for the past three years. His research interests include natural language processing, social network analysis, and feature engineering.

ERIC PARDEDE (Senior Member, IEEE) received the master's degree in information technology and the Ph.D. degree in computer science from La Trobe University, Melbourne, Australia. He is currently an Associate Professor with La Trobe University. His research work has been published in more than 100 publications in international journals and conference proceedings. His research interests include data analytics, IT education, and entrepreneurship.

A. S. M. KAYES (Member, IEEE) received the Ph.D. degree from the Swinburne University of Technology, Australia. He is currently a Senior Lecturer in cybersecurity with the Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia. His research interests include information modeling, data privacy and security, context-aware access control, the Internet of Things, and cloud and fog security.