Received: 24 October 2022 | Revised: 26 April 2023 | Accepted: 7 June 2023

DOI: 10.1111/1468-5973.12483

ORIGINAL ARTICLE

WILEY

A two-step approach to detect and understand dismisinformation events occurring in social media: A case study with critical times

Seungwon Yang[1] | Haeyong Chung[2] | Dipak Singh[3] | Shayan Shams[4]

[1]School of Information Sciences, Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana, USA

[2]Department of Computer Science, University of Alabama in Huntsville, Huntsville, Alabama, USA

[3]Department of Computer Science, Stephen F. Austin State University, Nacogdoches, Texas, USA

[4]Department of Applied Data Science, San Jose State University, San Jose, California, USA

Correspondence
Seungwon Yang, School of Information Sciences, Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana, USA.
Email: seungwonyang@lsu.edu

Funding information
Russell B. Long Professorship in the College of Human Sciences and Education, Louisiana State University; National Science Foundation, Grant/Award Number: HCC-2146523; National Institute of Justice, Grant/Award Number: 2019-75-CX-K001

Abstract

This article describes a novel two-step approach of detecting and understanding dis/misinformation events in social media that occur during disasters and crisis events. To detect false news events, we designed a deep learning-based detection algorithm and then trained it with a transfer learning scheme so that the algorithm could decide whether a given group of rumor-related tweets is a dis/misinformation event. For understanding how dis/misinformation was diffused in social networks and identifying those who are responsible for creating and consuming false information, we present *DismisInfoVis*, which consists of various visualisations, including a social network graph, a map, line charts, pie charts, and bar charts. By integrating these deep learning and multi-view visualisation techniques, we could gain a deeper insight into dis/misinformation events in social media from multiple angles. We describe in detail the implementation, training process, and performance evaluations of the detection algorithm and the design and utilization of *DismisInfoVis* for dis/misinformation data analyses. We hope that this study will contribute to improving the quality of information generated and shared on social media during critical times, eventually helping both the affected and the general public recover from the impacts of disasters and crisis events.

KEYWORDS
deep learning, dis/misinformation, disasters, fake news, information visualisation, social media

1 | INTRODUCTION

Social media is a double-edged sword. Although various platforms help us to connect and interact with family, friends, or colleagues online, the growing amount of dis- and misinformation spreading through social networks has been affecting the integrity of our society from microscopic (e.g., individuals) to macroscopic levels (e.g., large groups with political affiliations, states, countries) in a harmful way (Cook et al., 2015; Figueira & Oliveira, 2017). Disinformation is information that is intentionally misleading, possibly with malicious intent, and misinformation is false and inaccurate but may not have bad intent (Fallis, 2014; Wardle & Derakhshan, 2018; Wu et al., 2019). In this article, we refer to *dis/misinformation* as the term to cover both dis- and misinformation.

Fast-spreading dis/misinformation in social media has disrupted many areas including politics, online markets, celebrity gossip, and so on. The harmful effect of dis/misinformation can be especially severe during disasters and crisis events. Sapir and Lechat (1986) presented the special characteristics of natural disasters by placing each of the four major natural disasters (earthquakes, cyclones, floods, and

This is an extended and revised version of a preliminary conference proceeding: Reference information detail is not provided here due to double-blind review purposes.

drought-related famine) on relative scales (low to high) of predictability, lethality, scope, and onset delay. Their study shows that disasters such as cyclones (or their stronger version, hurricanes) or earthquakes have low predictability, high lethality, wider/pervasive scope (due to floods caused by cyclones), and a sudden onset delay. Considering that these types of natural disasters often make response and recovery efforts more difficult and time-consuming, it is clear that immediate and flexible reactions to fast-changing conditions during disasters are necessary. However, the prevalence of incorrect or intentionally-fabricated information spreading in social media spaces could delay or disrupt such responses (i.e., prompt rescue and recovery operations) and cost human lives, further delaying the affected communities' return to normal.

Thus, identifying dis/misinformation events spreading via social media, especially during disaster events, and understanding their patterns and mechanisms of diffusion into social networks are crucial for the resilience of the affected community and that is the reason the authors initiated this study. Identifying, understanding, and coping with dis/misinformation would also help us have a stable and sustainable society via an open and trustworthy Internet ecosystem.

There have been many prior efforts to detect dis/misinformation. Early studies relied on the analysis of texts from social media data (Rubin et al., 2015). However, more recent and successful studies have integrated the characteristics of social media users (e.g., user bio, number of followers/followees), interactions among them (e.g., retweets, mentions), engagement patterns between users and textual content, as well as the spatiotemporal aspects (e.g., timestamps) of the news being diffused through the networks (Conroy et al., 2015; Qazvinian et al., 2011; Ruchansky et al., 2017; Shu et al., 2017; Vosoughi et al., 2018). When these diverse characteristics of dis/misinformation events are integrated into a model in a holistic manner, with the help of deep learning-based algorithms we may expect more accurate detection of such questionable events from social media.

To support the multifaceted analyses of dis/misinformation events from a large number of tweets, it is necessary for the analyst to use coordinated multi-view visualisations (CMV), which enable them to connect and make sense of various related data sets on the same visual space. Specifically, CMVs are useful for the analyst to perceive and assess the comprehensive perspectives of dis/misinformation events, compare and link the different types of the main content (tweet text) and tweet metadata (time, retweet network, geolocation, user bio, hashtag, sentiment, etc.), as well as evaluate other external information including websites and news outlets spreading rumors. In this paper, we present *DismisInfoVis*, a new visual analytics tool, which is based on seven views. In *DismisInfoVis*, each of the seven views supports more detailed and independent analysis of individual tweet metadata; at the same time, these individual views are connected, and their data representations are updated according to the selection of and interaction with data items in the other views. For example, a change in the time range in the Timeline view will highlight the corresponding data items (which belong to the selected time range) in the other views.

Only when these multiple characteristics of dis/misinformation events are integrated into a holistic model, we can expect more accurate detection of such events from social media. Deep learning-based algorithms are the most suitable tools for integrating multifaceted data commonly generated from social media posted during large-scale disasters (Cao et al., 2020; Jain et al., 2018; Sun et al., 2019). This is because deep learning models typically have more than one layer with a large number of parameters, and thus they can create complex and accurate models. Another advantage of deep learning-based algorithms is that they do not need extensive feature engineering on the data set, allowing users to focus more on training the algorithm.

Upon detecting a group of tweets, each of which shares a certain dis/misinformation, by applying the deep learning-based detection algorithm, a multifaceted analysis of those tweets using multiview visualisations could be necessary to make sense of the event from different angles.

Figure 1 summarizes the phases of dis/misinformation detection and visual analysis presented in this paper. Before Phase 1, there was a data preparation step, in which we collected social media data and developed a training data set for our detection algorithm. A group of tweets sharing a specific dis/misinformation during a hurricane event were collected using automatic means. In Phase 1, our deep learning-based detection model is trained. Once the training is completed, the algorithm can identify whether or not a given group of tweets contains dis/misinformation. In Phase 2, we apply our developed multiview visualisation tool that can show the who, where, and what aspects of the dis/misinformation events based on the identified group of tweets. Our visualisation tool helps users make sense of the event through iterations of interactively selecting, zooming, panning, and linking of data points.

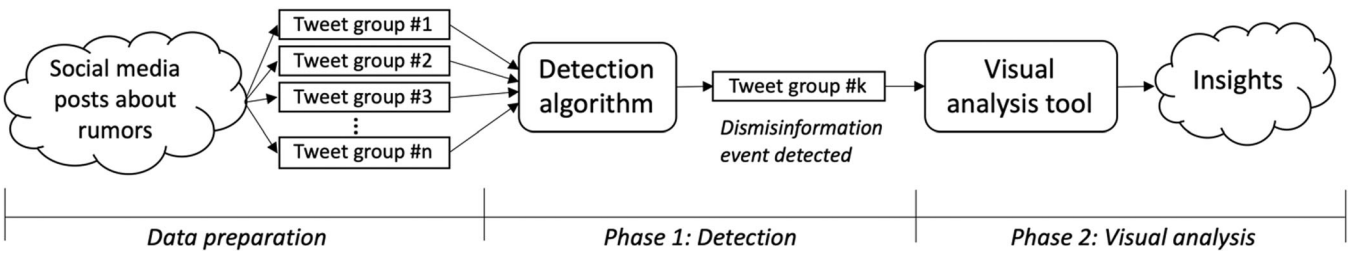

FIGURE 1 Overview: detecting dis/misinformation events in Phase 1; visual analysis in Phase 2.

In the following sections, we introduce our deep learning model for dis/misinformation detection and our design of the visualisation tool, followed by the methodology and analysis results of a use case study involving a real-world social media data set. Then, we discuss our approaches and case study results. The brief summary and our plans for future works are provided in the conclusion section.

2 | RELATED WORK

2.1 | Characteristics of dis/misinformation and crisis management

The terms, fake news, misinformation, disinformation, mal-information, or rumors, are often used interchangeably, causing some degree of confusion among users. Multiple studies have been published to 'calm the troubled water' (Egelhofer & Lecheler, 2019) caused by the widespread use of the term 'fake news' in the online information environment, providing comprehensive reviews and clarification regarding what the term actually means for various areas including politics. Wardle and Derakhshan's study (2018) focuses on the various types of dis/misinformation and where they could be placed on the spectrum of 'information disorder'. The authors compare misinformation and disinformation, both being false information, and contrast that they have the opposite intent of the people who disseminate the information: the person who disseminates misinformation believes it is true, while the person who disseminates disinformation knows it is false. Another interesting term introduced in this study is mal-information, which is information based on real and non-false information used to inflict harm on people or organisations. Examples of mal-information include revealing private information, harassment, or hate speech.

Regarding the types of fake news, Wardle and Derakhshan (2018) provide seven types of narratives (i.e., satire and parody, false connection, misleading content, false content, imposter content, manipulated content, and fabricated content) that could be placed on the information disorder spectrum. Tandoc et al. (2018) also provide the six types of fake news. Among them, categories such as news satire, news parody, and news overlap with those of Wardle and Derakhshan (2018). However, categories such as photo manipulation, propaganda, and advertising are nonoverlapping. Tandoc et al. further identify two dimensions of fake news, (1) level of facticity and (2) immediate intention (to deceive), each having continuous degrees from low to high. Then, they place the six categories of fake news within this model using the two dimensions as a map of different definitions of fake news from multiple studies for clarification purposes.

Researchers also examine the 'dimensions' of fake news when analysing dis/misinformation as public communication (Egelhofer & Lecheler, 2019). They propose two dimensions: (1) the fake news genre, which refers to intentionally creating disinformation in journalistic formats, and (2) the fake news label, which is the term used by political actors as an effective weapon to delegitimize journalism or news media. In their definition, the fake news genre is about intentionally creating disinformation in journalistic formats and the fake news label is the term used by political actors when they use the term as an effective weapon to delegitimize journalism or news media. The authors compare the fake news genre/label with other concepts such as propaganda, rumors, conspiracy theories, or media criticism, and provide a research agenda to reduce damage to journalism as a whole.

In their recent study, Yang and Luttrell (2022) provide typologies of fake news and identify factors that may potentially contribute to the wider dissemination of such news along with multiple AI-based detection methodologies to combat the spread of dis/misinformation. They also present the theory of content consistency, a framework to semantically measure news content in multiple levels including the journalism domain level, creator level, platform level consistency, display or presentation-tier level, and network level. Examining the content of viral rumors is an essential component in the identification of dis/misinformation, and our model includes that component in our detection algorithm.

Dis/misinformation can cause harm, especially during a crisis, as it can lead to confusion, panic, and incorrect decision-making for the victims and emergency managers (Hunt et al., 2020; Naeem et al., 2021). Crisis management is the process of preparing for, responding to, and recovering from a crisis. The spread of dis/misinformation during crises events or terrorist attacks can seriously undermine the effectiveness of crisis management efforts considering that such false information could disrupt emergency communications that have to occur at these critical moments (Hunt et al., 2020). Hunt et al. (2020) present several example cases that involved the dissemination of dis/misinformation during crisis events. One of them was the rumor during Hurricane Harvey in 2017 stating that undocumented immigrants could not get into Texas shelters without checking their IDs. This rumor potentially put the lives of over 500,000 undocumented immigrants in Houston area in extreme danger since many of them had been evacuated and had nowhere to seek safety.

Due to the devastating effect of dis/misinformation during crisis events, government and emergency organisations, as well as online sites such as Snopes (www.snopes.com/), FactCheck.org (www.factcheck.org/), or PolitiFact (www.politifact.com/) have been debunking viral rumors and disseminating verified information as part of their crisis management efforts. For example, the Federal Emergency Management Agency (FEMA) sets up rumor control pages for several hurricanes and Covid-19 pandemic to list multiple rumors and facts (www.fema.gov/blog/harvey-rumor-control). Fact-checking websites investigate viral rumors and attach labels (e.g., true, mostly true, neutral, false, unfounded, unproven, outdated, etc.) along with detailed information regarding sources used for the fact-checking. Pennycook et al. (2021) examine the reasons behind people's sharing of dis/misinformation and potentially effective interventions. They found that if people are exposed to fact-checking interventions and have an opportunity to focus their attention on the accuracy of the information, they are less likely to share dis/misinformation on Twitter.

Debunking rumors and tracking social media posts by humans requires a significant amount of time and effort (Hunt et al., 2020). Thus, this became one of the motivations for the authors of this article to pursue the development of a dis/misinformation detection algorithm and an interactive visualisation interface to understand the found dis/misinformation events on Twitter.

2.2 | Automatic detection of dis/misinformation events in social media

Various research attempts to address the issue of detecting false information events occurring in social media. Pierri and Ceri (2019) group the main approach of the studies into three categories that focus on: (1) the content of false news itself, (2) the social context of false news, and (3) the combination of content and context.

Content-based dis/misinformation event detection algorithms focus on the linguistic features of news content (Horne & Adali, 2017; Wang, 2017; Potthast et al. 2018; Hosseinimotlagh & Papalexakis, 2018; Popat et al., 2018; Pérez-Rosas et al. 2018). The goal of social context-based approaches is the identification of a false news cascade. A cascade in social media is a group of social media postings sharing identical or very similar false news content. The context-based approach analyses social aspects of the news being diffused on Twitter (Liu & Wu, 2018; Tacchini et al., 2017; Volkova et al., 2017; Wang et al., 2018; Wu & Liu, 2018). User profile information, interactions amongst users, and also interactions between users and dis/misinformation could become features to consider in uncovering such cascades. Finally, the combined approaches integrate content-based approaches with social context-based approaches. They are considered the most effective and the latest approach for dis/misinformation event detection from social media (Ruchansky et al., 2017; Shu et al., 2017; Volkova & Jang, 2018). For this reason, our proposed detection algorithm is also designed based on this combined approach.

Ma et al.'s study (2016) introduces four recurrent neural networks (RNN)-based models that have been trained for dis/misinformation detection: Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997; Graves 2013); tanh-RNN; Gated Recurrent Unit (GRU) (Cho et al., 2014) with 1-layer hidden units; and GRU with an extra hidden layer. Deep learning-based models, including RNN (Rumelhart et al. 1986), are known to be advantageous over traditional machine learning models. These researchers created a training data set by converting both Twitter and Weibo (Chinese version of Twitter) data into time series data having variable length. After training, the model performances were compared against those of machine learning models. The result shows that their model outperformed the traditional machine learning models for the Weibo data set. Although GRU models showed better performance (i.e., higher accuracy and F measure) with the Twitter data, the Support Vector Machine model performed slightly better in terms of precision and recall measures. We adopted the RNN-based model in our algorithm and trained it (in part) with the publicly-shared Twitter data set by Ma et al. However, our model is different from that of Ma et al.'s in that ours includes a user module that can independently address social aspects of the dis/misinformation propagation. Additionally, the training of our model includes a fine-tuning process based on the transfer learning (TL) scheme using our own dis/misinformation data set.

The study by Ruchansky et al. (2017) presents a hybrid model, which consists of three components: the Capture, the Score, and the Integrate modules. Their Capture module analyses the textual data of dis/misinformation, as well as the timestamps of user-to-news interactions utilizing *doc2vec*, which is a popular paragraph embedding approach (Le & Mikolov, 2014) and the volume of dis/misinformation posts per unit time. The Score module computes the user-to-user engagement frequencies mediated by dis/misinformation articles. The outputs from the Capture and Score modules are combined into a single Integrate module for final prediction.

One of limitations of the Ruchansky et al.'s model was possibly that the Score module was dependent on user-to-user interactions appearing only in the training data set. This approach may make the model less robust due to such dependency. Our model also has components that correspond to the Capture and Score modules in Ruchansky's study. However, the differences included that: (1) our model addresses the dependency issues regarding user-to-user interactions by using the Twitter user profiles and (2) our model adopts the TL training scheme to address the limited size of a disaster-specific training data set.

2.3 | Use of information visualisation for social network data analysis

In our data analysis, we created a customized visualisation tool, *DismisInfoVis*, for facilitating the understanding of dis/misinformation events spreading in social networks. Recently, data visualisation and visual analytics have become essential to discover and understand patterns on social media networks (Chen et al., 2017; Wu et al., 2016). *DismisInfoVis* extends following existing social network visualisations by focusing on visualising retweets and diffusion of dis/misinformation among users, simultaneously showing multiple aspects of attributes of social media data such as user locations, volume of posts over time, frequently-shared URLs, and so forth.

2.3.1 | Social network visualisations

In social networks, a large number of users make connections, and these relationships are formed with respect to following and posting messages (Heer & Boyd, 2005; Brandes & Nick, 2011), reposting, or retweeting the messages initially posted by others (Cao et al., 2016; Chen et al., 2016). Such reposting activities often lead to the construction of a diffusion network of social media posts. There is a network visualisation that illustrates the diffusion among users (Viégas et al., 2013).

2.3.2 | Spatiotemporal visualisation in social media

Social media data includes several attributes including time stamps, text, images or videos, and possibly geolocation data. Both the time stamp and geolocation data can provide spatiotemporal characteristics of dis/misinformation diffusion and user behaviors. Geolocation data provides the spatial context of users in the information diffusion process. Prior visualisation works supported analysis of the geographic information diffusion (Cao et al., 2012; Chua et al., 2014). Also, a visualisation for event detection in social media (Marcus et al., 2011) enables users to explore different events in a timeline, which integrates spatial and temporal information using unique visual representation.

2.3.3 | Visual analytics for dis/misinformation

There exist two prior visual analytics for analysis and detection of dis/misinformation diffusion patterns. In general, social media data can be quite large and embedded with complex behavioral and pattern-related data, both of which are demanding for analysts to extract and understand (Loyola-González et al., 2019). Focusing on dis/misinformation social media analysis and visual analytics approaches allow for comparison, identification, and clustering of dis/misinformation-related tweets. It also provides dynamic interaction visualising multiple attributes (Chen et al., 2017; Wu et al., 2016).

Recent efforts have employed visualisation and visual analytics systems to more effectively discover and understand patterns of dis/misinformation diffusion in social media by enabling users to gain insight into single pieces of tweet data (i.e., text and retweet data). For instance, the following visual analytics studies focus on analysing text information of social media data. XFake (Yang et al., 2019) visualizes decision trees, thereby enabling users to track the decision process about fake news tweets to determine the classification of certain news as either fake or authentic. This type of visual analytics supports the enhanced assessment of news articles from social media for explaining fake news using three frameworks (ATTN, MIMIC and PERT), which allow the users to analyse text information of news from different text analytics aspects. Another visual analytics study conducted by Seref et al. (2020) presents text analytics based on a Context Map approach, which entails the visualisation of a connected network of n-grams (a set of n consecutive words in a corpus) designed to identify similar fake news content from the perspective of text analytics.

It should also be noted that there are studies using graph visualisations to analyse structural and dynamics characteristics of the fake news diffusion network from retweeted data. For example, the visual analytics by Zhao et al. (2020) visualize how fake news spreads differently in comparison to verifiable news in terms of the relative size and complexity of the cascades, which represent a hierarchy of retweeting/reposting in a network graph. Similarly, HOAXY (a graph visualisation) developed by Shao et al. (2018) visualizes diffusion network graphs to enable the analysis of how misinformation spreads and competes for dominance over Twitter.

While these visual analytics programs can provide a more detailed assessment of fake news content and diffusion networks, fake news may emerge from a range of other data sources that cannot be easily analysed with such existing tools. Consider that a single tweet includes various types of metadata including timestamps, geolocation, sentiment (e.g., like or dislike), keywords, retweets, followers, hashtags, and so forth. Thus, using various metadata along with retweet information, we can understand how dis/misinformation is spread by people in different locations and embracing divergent political orientations. In this regard, our *DismisInfoVis* was designed to provide more diversified and comprehensive perspectives of dis/misinformation by considering the metadata of tweets to better understand the context of one or more dis/misinformation events.

Similar to our *DismisInfoVis*, FakeNewsTracker (Shu et al., 2019) was designed to detect and assist the user in understanding dis/misinformation. Specifically, the visual analytics tool employs visual representations such as networks, maps and timeline visualisations. However, in contrast to FakeNewsTracker, our *DismisInfoVis* is able to facilitate the sensemaking of synthesized information from multiple views instead of examining each single view separately. *DismisInfoVis* is particularly designed to link and combine diverse information that emerges from both available metadata and external information (such as website and news article) in a single visual analytics system. By combining and synthesising visual representations from different types of attributes extracted from tweet streams conveying dis/misinformation, we can better understand the patterns of dis/misinformation events in social networks, as well as how such diffusion may (or may not) be influenced by user behaviors within social networks.

3 | METHODOLOGY

3.1 | Data preparation for detection model development

Our training data set includes multiple groups of tweets. Each group of tweets has a label as either 'false' or 'true' depending on the veracity of the rumor to which the group of tweets were related.

3.1.1 | Data collection

Figure 2 shows the procedures for collecting rumor-related tweets and the content details of the training data set that is developed from the collected Twitter posts. First, the rumor debunking articles and their labels (false or true) are harvested from a fact-checking website (www.snopes.com). Second, representative keywords are manually selected from the titles and descriptions of each rumor. Then, using a Web browser automation system (i.e., Selenium), a database, Twitter APIs, and the keywords for rumors, a Twitter data collection is built by querying Twitter's search engine using keywords that are associated with each rumor. As the next step, returned tweets are stored in a MongoDB

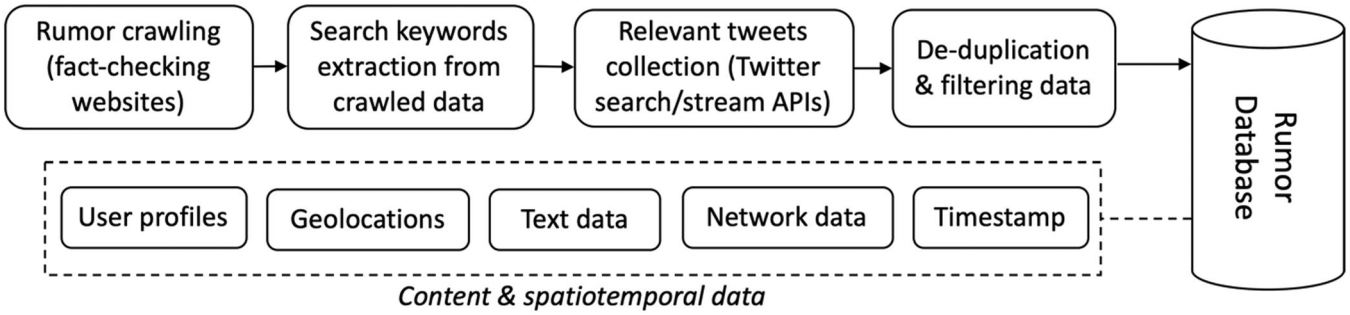

FIGURE 2 Rumor Twitter data collection overview.

database after going through the de-duplication and filtering steps. From tweets stored in MongoDB, we extract metadata including the user profiles, counts for replies, likes, or retweets, and also geolocation data if it exists to create our training data set as enclosed within the dotted box in Figure 2.

3.1.2 | Ground truth label

Snopes.com, one of the popular fact-checking websites, classifies rumors against broad categories of ratings such as false, mostly false, mixed, mostly true, unproven, or miscaptioned. For this study, the authors collected the Snope's articles that have labels such as 'false' or 'mostly false' (we merged 'false' and 'mostly false' into the 'false' category), as well as 'true'.

3.1.3 | Data preparation

The input data for our Content Analysis module (Figure is prepared based on the methodology of Ruchansky et al.'s (2017). Each Twitter event, *e*, is a temporal sequence of Twitter posts that have timestamps within a time interval *t*. The engagement between a user, *ui*, and an event, *ei*, at time *t* can be represented as a tuple $X = (\in, \Delta T, X_u, X_t)$. $\in$ denotes the tweet count within the time interval ΔT. X_u denotes a specific row of an adjacency matrix, which can be constructed by an event *ei* and a user *ui* involved in ΔT. X_t is a vector representation of tweet texts in numeric form within ΔT. Let's say that there are $e_1, e_2,\ldots, e_n$ dis/misinformation events. Each event e_i has a set of tweets $tw_1, tw_2,\ldots, tw_k$ posted by users $u_1, u_2,\ldots, u_l$. We partition these tweets into different sets $s_1, s_2,\ldots, s_m$ based on ΔT, which is the time interval. Then, X can be represented as the vector of tuples $(\in, \Delta T, X_u, X_t)_1$, $(\in, \Delta T, X_u, X_t)_2$ … $(\in, \Delta T, X_u, X_t)_m$. In this case, $\in$ denotes the total tweets in a set s_i. To compute X_u, we create an adjacency matrix that consists of all users and events. In this matrix, each row represents the number of times a user, u_i, is involved in events from e_1 to e_n. Further, the mean value of the collection of rows (i.e., the users involved in a set s_i) of the adjacency matrix is computed as X_u. We then reduce the dimension of the adjacency matrix with the application of the Principal Component Analysis (Abdi & Williams, 2010). As a result, the matrix will have a dimension of 20. In a similar manner, we process a group of tweets in a set *si* and use the document embedding technique, *doc2vec*, to produce the vector representations of tweet texts, X_t.

The Context Analysis module analyses the metadata extracted from the Twitter user profile data, and these profiles usually represent the users' involvement in an event, *ei*. For acquiring the user profile data, we relied on our data collection system, which combines a Web automation tool, Selenium, a database, and multiple Python scripts. These components work together by sending search queries to Twitter's search box and collecting the returned Twitter data. The user profile metadata includes the count of shared tweets, the numbers of followers/followees/likes/links/media shared, as well as the age of the profile, the status of the account (e.g., private or verified), and the latitude/longitude data showing where the tweet was posted.

We process Twitter user profiles so that user characteristics could be uncovered based on their activities within social networks. Thus, users having different levels of susceptibility for dis/misinformation could be discerned by our Context Analysis module.

3.2 | Detection model development

Our Content Analysis module that captures the temporal and textual features from tweets was designed following the Ruchansky et al.'s study (2017). Further, we incorporated an upgraded Context Analysis module. As mentioned earlier, one of the drawbacks of Ruchansky et al.'s model could be the model dependency on user-to-user interaction data. If the model must predict labels for unseen (i.e., not existing in the training data set) users and unseen news articles, this model may not perform robustly. Our approach was to design the Context Analysis module so that it could characterize an individual user's data based on the Twitter user profile. Another challenge that we faced in the model development is that we had the limited size of a disaster-related dis/misinformation data set. To address this challenge, we depended on the TL scheme.

3.2.1 | Content analysis module

This module (Figure 3a) analyses the temporal and textual features identified from dis/misinformation events. A time-distributed

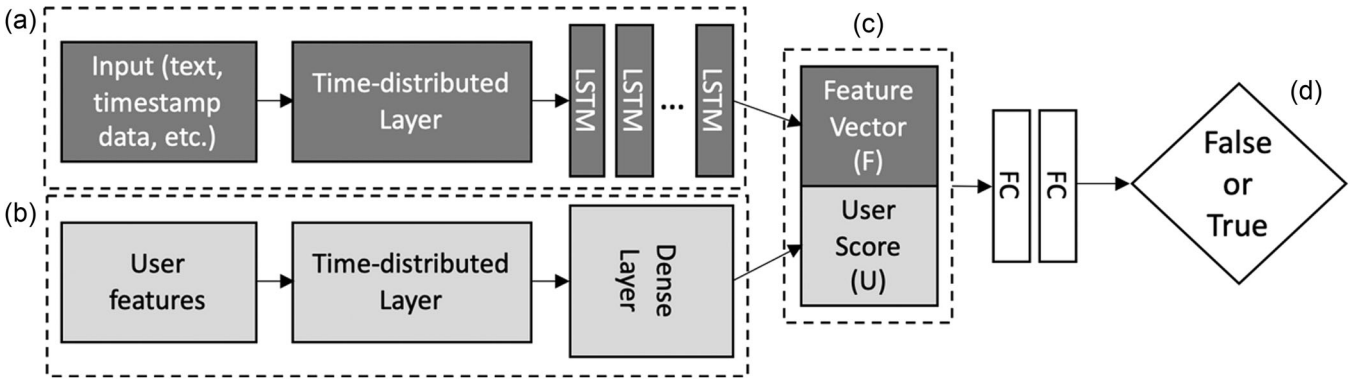

FIGURE 3 Components of the detection algorithm. (a) Content Analysis module, (b) Context Analysis module, (c) merging of the feature vector F and the user score vector U, and (d) final decision (T/F).

embedding layer was incorporated for standardising the input *X*, before feeding it into the LSTM model. LSTM models perform well in tasks such as capturing long-term dependencies or handling variable-length inputs. Then, the final (hidden) layer of the LSTM model is connected to a fully-connected layer, outputting a low-dimensional feature vector *F* as the result. This feature vector *F* contains temporal and textual characteristics.

3.2.2 | Context analysis module

This module (Figure 3b) analyses user-related features from the Twitter profile data. After going through the time-distributed layer, the dimension of the user feature matrix decreases from eight to four. Then, the dense layers are further applied to construct the user feature vector *U*. To identify users who are involved in a specific batch, masking is applied as well.

The outputs from the Content Analysis and the Context Analysis modules are then concatenated so that the entire model could be trained jointly (Figure 3c). As the last step, final decision of whether the input is true or false information is made (Figure 3d).

3.2.3 | Training with TL

We relied on TL to make our detection model perform robustly (See Disaster Data set in Table 1). First, we pre-trained our model using a large publicly-shared data set (but not specifically pertaining to false/true news during disasters), Rumdect Data set, in Table 1. Second, we trained our model further on the more specific training data set, the Disaster Data set in Table 1, which was developed from tweets about false/real news during a hurricane event. The generic set of features were learned in the first step, followed by the fine-tuning of the detection model in the second step.

TL is applied only to the Content Analysis module. Then, this module is jointly trained with the Context Analysis module. We have both the cross-entropy loss (L_{ce}) and the *L*2 regularisation in the loss function:

$$L = L_{ce} + \lambda \|W\|^2$$

To address the overfitting issue, we randomly dropped the weights in the Dense and Time-distributed Layers. Tensorflow 1.8 and Nvidia GeForce GTX 1080Ti were used for the training of our model. When using the Rumdect Data set, we divided it into ratios: 80% for training, 5% for validation, and 15% for testing. The accuracy of the model is measured with a five-fold cross-validation with the Disaster Data set that we developed.

3.2.4 | Training data set preparation

For our Rumdect Data set, we only used the tweet data part, excluding the Weibo data. The Rumdect data set provides only the tweet IDs (Twitter IDs [snowflakes] 2010; Ma et al., 2016). Thus, we had to hydrate those IDs so that we could access the full tweet texts and metadata. In the process, we found that some of the tweets were not accessible, possibly because those accounts had been suspended or deleted at the time of this study. Thus, our 'hydrated' Rumdect Data set included 991 events and about 570,000 tweets associated with those events. Each event consists of a news story (both true and false news), tweet IDs linked to the news story, and the label for the news (false or true). The tweet IDs may represent the user (*ui*) engagements with each event at time *t*. The Disaster Data set contains 91 instances of false or true news events, which are associated with relevant tweets and user profiles.

3.2.5 | Parameter setting

Each engagement in the data set is segmented with ΔT, which denotes the time interval. All the engagements in ΔT are considered as a single input to LSTM. For the hyperparameters, the

Data sets used for TL	Rumdect Data set (pretraining)	Disaster Data set (fine-tuning)
No. of events	991	91
No. of false events	497	46
No. of real events	494	45
No. of users	226,791	31,305
Total no. of tweets	569,912	37,975
Avg. event period (hours)	1961	3924
Avg. no. of tweets/event	575	417
Max no. of tweets	37,475	4041
Min no. of tweets	4	6
Avg. no. of tweets/user	2	1

TABLE 1 Comparison of the two data sets used in TL.

regularization loss parameter was set to $\lambda = 0.01$, the dropout probability to be 0.4, and the learning rate to be 0.001. We used the Adam optimizer. As the partition unit ΔT, we used a 1-hour granularity. Additionally, the window size of 10 and a vector size of 100 were used as parameters for the doc2vec document embedding. The dimension of the weights in the Context and Content Analysis modules were configured to 100.

3.3 | The design of DismisInfoVis

It remains challenging for researchers and others to develop deeper insights into plots involving how these tweets spread through social media. After detecting a false news event in social media, a detailed investigation is necessary to understand the pattern, characteristics, or significant individuals, who might be responsible for diffusing dis/misinformation. In this visualisation, Twitter users are categorized into one of the following two groups:

- Trusters: users who trust the rumor as the truth/facts
- Non-Trusters: users who realize that the rumor is actually false

To explore the challenges and key tasks associated with analysing the dis/misinformation tweets, two researchers were asked to identify several characteristics of dis/misinformation data according to multiple attributes during the initial analysis session. However, in the absence of visualisation tools, it was difficult for the researchers to explore and analyse attributes of the labelled tweets while considering the overall propagation of tweets through social networks.

Thus, the main motivation for developing a visualisation tool came from the significant challenges of analysing social media data from diverse angles—the difficulty of which can be attributed to their multivariate and multidimensional characteristics. An analysis task of this nature typically requires an individual to examine an overview of tweets, while also conducting detailed analyses of the content of tweets and their attributes to understand the propagation of dis/misinformation in social networks over time. It is important for users to be able to view the social network structure, as well as the temporal changes of tweets associated with dis/misinformation.

Figure 4 presents the overall user interface of *DismisInfoVis*. Throughout the views in Figure 4, the same colour coding was used for data items: 'orange' corresponds to the trusters (of fake news) and 'blue' to the non-trusters. *DismisInfoVis* consists of seven interconnected views:

(a) The Social Network view: It allows the user to investigate the network structure of dis/misinformation events spreading on social media (Figure 4a). It displays nodes for both trusters and non-trusters, which are labelled before the analysis, as well as such nodes' retweet relationships in a node-link graph representation. A truster/non-truster is represented as a circle-shaped node and relationships are shown as lines connecting nodes.
(b) The Map view: It visualizes the distribution and frequency (using the size of the circle) of tweets with respect to the geographic positions of their posts (Figure 4b). In the Map view, an analyst can identify and track where the two types of dis/misinformation tweets were generated geographically on a map with coloured dots. The size of each dot at specific geospatial positions on the Map represents the relative number of tweets generated at the respective position.
(c) The Timeline view: It assists the user in visualising the temporal aspects of dis/misinformation-related tweets via a stacked bar chart enabling the analyst to observe quantitative changes associated with tweets over time (e.g., changes in the tweet volume distribution over 30-minute terms, hours, days, or weeks) (Figure 4c). Each stacked bar along the axis is divided into two sub-bars: one corresponding to the number of trusters and the other to the number of non-trusters. Each stacked bar also depicts how many tweets were created within a particular timeframe across the two categorical variables.
(d) The Sentiment view: It shows the sentiments computed from tweets throughout the timeline (Figure 4d). The line graph is

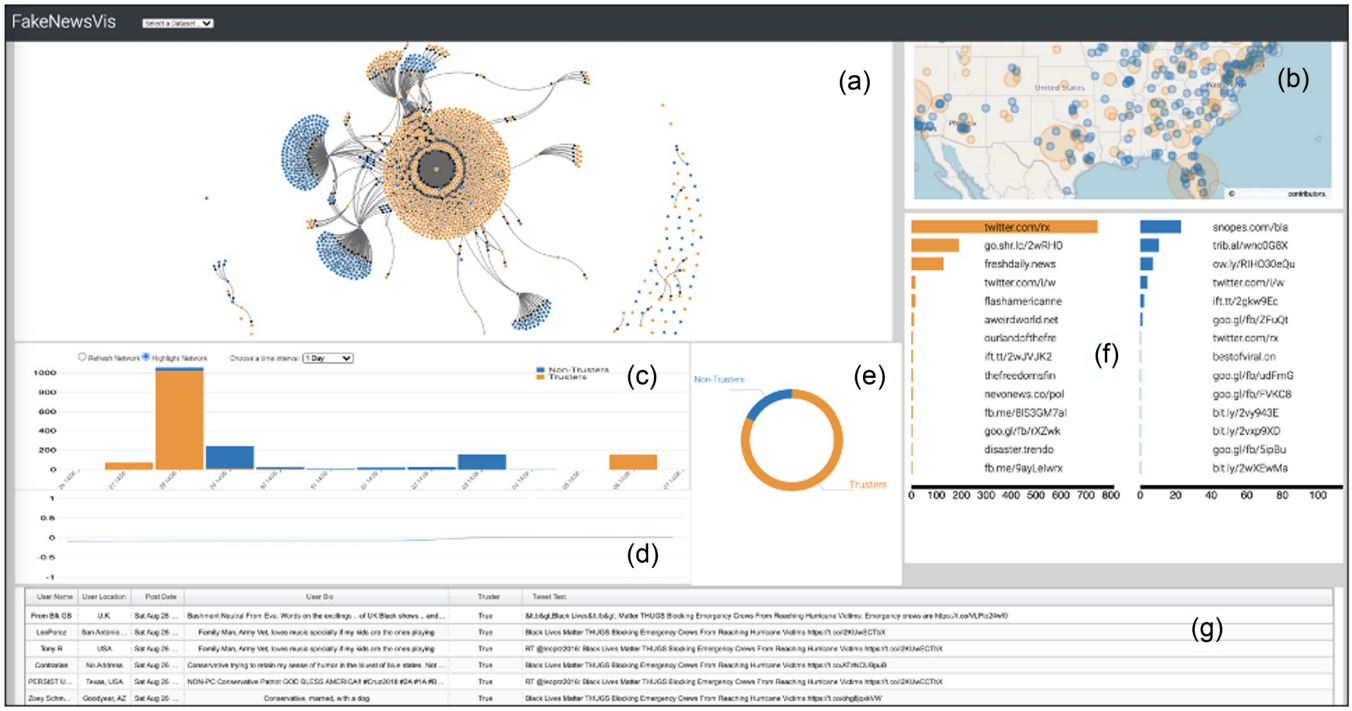

FIGURE 4 Overview of *DismisInfoVis* user interface. Two different colours indicate trusters (orange) and non-trusters (blue). (a) Social Network view, (b) Map view, (c) Timeline view, (d) Sentiment view, (e) Pie Chart view, (f) URL view, and (g) Detailed view.

represented on a scale of -1 to 1, based on the 'compound score' based on the popular VADER (Hutto & Gilbert, 2014) model for analysis of polarity (positive/negative/neutral) and intensity of emotion found in social media texts.

(e) The Pie Chart view: It enables users to have a quick insight into the overall ratio between the trusters and non-trusters (Figure 4e).

(f) The URL view: It shows the two groups of bar graphs for listing frequently shared URLs extracted from the two groups of tweets (trusters and non-trusters). It should be noted that a URL of a debunking article from snopes.com is often located on top of the URL list for non-trusters (Figure 4f).

(g) The Detailed view: It enables users to view detailed information about the labelled tweets with respect to the content of the text and multiple other attributes, in the form of a spreadsheet table (Figure 4g).

All of the views are connected via brushing and linking. Selecting data items in one view will highlight the corresponding data representations in the other views.

4 | RESULTS AND ANALYSIS

We conducted a use case study for the analysis and detection of dis/misinformation using a real-world social media data set, based on our presented two-step process of detecting and understanding dis/misinformation events occurred on social media.

4.1 | Performance of the detection algorithm

In Figure 5, the accuracies and F-1 scores are presented. The two graphs were obtained with our model that was trained using various cumulative event periods. The different cumulative time lengths of event periods, which were marked on the x-axis from 6 to 48 hours, are used to identify the best size of the training data set and to observe the characteristics of our model training.

The best accuracy of 91.47% and the best F-1 score of 90.89% were achieved when we trained the model only with the first 28 hours of the training data set (See the red vertical line in Figure 5).

4.1.1 | TL

TL was applied only to the Content Analysis module portion of the model. To understand the effect of different model configurations and TL in the model performance, we computed accuracies with different combinations as shown in Table 2.

When the Context Analysis module was integrated with the Content Analysis module, we could achieve an accuracy increase of 7.24% on average when this setting was compared against the Content Analysis-only setup, regardless of the TL application. This may show us that our Context Analysis module could positively contribute to the overall model performance. The TL application to the Content Analysis module also showed increased accuracy of 5.3% on average whether we add the Context Aware module or not. From this, we could say that the problem of an insufficient amount of

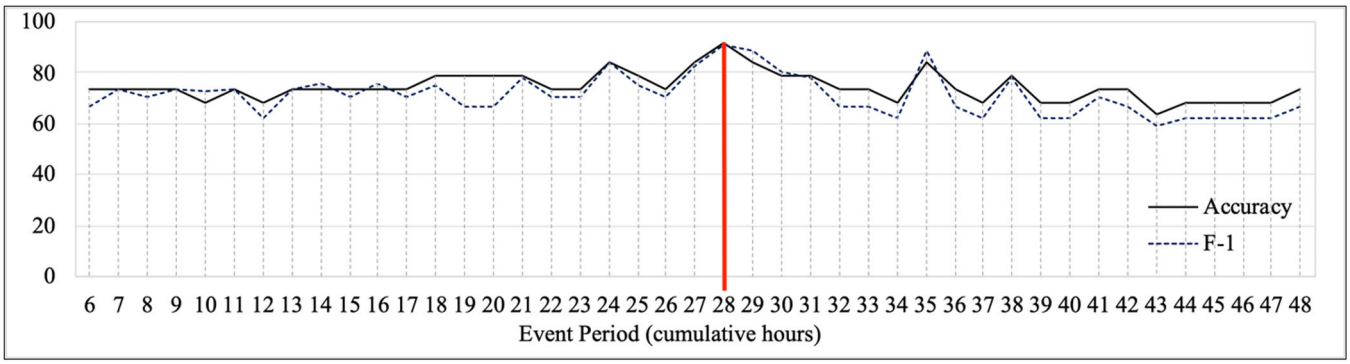

FIGURE 5 Accuracies and F-1 scores of our detection algorithm. Best performance achieved at the event period of the first 28 hours of data, marked with a red vertical line.

TABLE 2 Accuracies for different model setup and with or without the application of TL.

Model setup	Without TL	With TL
Content Analysis only	78.94	84.21
Content + Context Analysis	86.15	91.47

training data in this study could be addressed successfully due to the application of TL. The accuracy improvement is much more significant (12.53%) when the Content Analysis-only model devoid of the application of TL (accuracy 78.94%) is compared against the Content + Context Analysis model with the application of TL (accuracy 91.47%). This fact may support that our approach of using the Twitter profile data and the application of TL in this study was successful.

4.2 | Analysis of a dis/misinformation event using DismisInfoVis

We present how *DismisInfoVis* could be applied to analyse a real-world Twitter data set to demonstrate its utility. The real-world data set that we used was a Black Lives Matter (BLM)-related rumor, and it had a total of 1780 tweets. These tweets contained the text, 'BLM protesters are blocking emergency crews from reaching hurricane victims'. Among several dis/misinformation events that occurred in Twitter during Hurricane Harvey, we selected this BLM rumor event considering that it was one of the most viral events, and thus allowed us to analyse and present different aspects of the events with our *DismisInfoVis* tool.

Understanding dis/misinformation events occurring on Twitter during a large-scale disaster is not always easy due to the large volume of data sets (Sadri et al., 2018), as well as the multi-dimensional aspects of social media data, which often include textual messages with hashtags and embedded external links, information about the temporal progression, user-related information, network information such as retweets or mentions, or geolocation features (Maddock et al., 2015). By examining patterns that have been uncovered from these multi-dimensional 'signatures', we may have enhanced understanding regarding the propagation of dis/mis-information events on Twitter, especially under the context of disasters where complex communication could occur among multiple parties about various topics (Maddock et al., 2015; Pourebrahim et al., 2019). For this reason, our overarching question that we would like to address using *DismisInfoVis* is:

4.2.1 | Our overarching question

W*hat are the distinct patterns identified from the dis/misinformation event which occurred on Twitter during Hurricane Harvey in 2017?*

Based on this question, we also formulated three specific questions that we would like to investigate using *DismisInfoVis*.

In Question #1, we seek to understand the overall volume of tweets and sentiment changes over time by both trusters and non-trusters of the dis/misinformation, and the pattern of overall social networks. Understanding the tweet posting behavior (e.g., volume) could be important since it helps identify key periods of the disaster event and track the spread of information (Vieweg et al., 2010). Also, monitoring sentiment throughout the periods could help emergency responders understand the psychological impact on the people and affected communities and plan their responses accordingly (Beigi et al., 2016; Buscaldi & Hernandez-Farias, 2015).

Question #1: *What is the pattern of posting tweets over time, the communication pattern, and the ratio of people who trusted (or did not trust) the false news?*

Specifically, we examined the visualisations presented on three different views.

4.2.2 | Tweet volume over time

In Figure 6, the height of the bars represents the frequency of posted tweets for each time unit (e.g., *day* in this graph) on the x-axis that

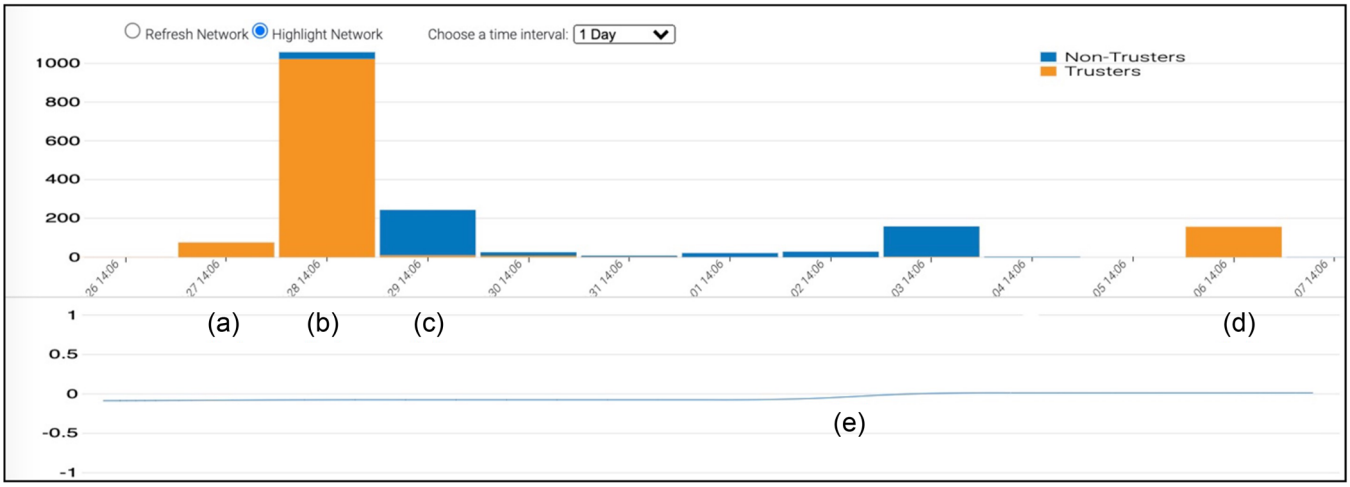

FIGURE 6 The volume of tweets per day as bar graphs, along with the sentiment scores as a line graph, in the Timeline view. (a) tweet volume of Day 1, (b) tweet volume of Day 2, (c) tweet volume of Day 3, (d) tweet volume of Day 11, and (e) sentiment change.

represents a total of 12 days. This false news, 'Black Lives Matter protesters are blocking emergency crews from reaching hurricane victims…', was probably designed to denigrate a specific political movement. The news gained enormous popularity within the first 2 days (Figure 6a,b) of its posting, as demonstrated by the highest orange bar labelled in Figure 6b. During those 2 days, the trusters, whose tweet frequencies are marked with orange, believed that the news was true and shared it on their Twitter network. From the third day at Figure 6c, a large group of non-trusters started realising that this news was actually false. Following Figure 6c, these non-trusters continued posting their tweets, which mentioned that this popular news was false. Often, a URL pointing to a debunking article from snopes.com or some other fact-checking website was embedded as part of their tweet content (Figure 7). This act by the non-trusters was an effort to fight against the widespread false belief about the rumor. The re-emergence of the orange bar on the 11th day at Figure 6d illustrates that this false news remained persistent and was still being retweeted among the trusters even after a while. Overall, it was observed that the amount of Twitter postings made by the trusters dominated the social media in volume compared to those made by non-trusters, at least in the beginning phase of the news spread.

We could also observe the sentiment changes, which is visualized beneath the Timeline view as a blue line graph (Figure 6e). The sentiments appear on a slightly negative side (approximately -0.2) for the first 6 or 7 days. However, the sentiment score changed closer to a neural score of zero (approximately −0.1) around the 7th day. This slight change could be due to the continued efforts of the nontrusters (starting at Figure 6c) who posted and retweeted the fake news debunking articles published by snopes.com or FactCheck.org. An example of such a tweet, identified with the Detailed view in Figure 4g, was 'RT @factcheckdotorg: FAKE NEWS ALERT: BLM protesters did not block emergency crews from helping Hurricane Harvey victims' and the verb 'did not block' should have been interpreted as a non-negative sentiment score.

4.2.3 | Social network structures

As shown in Figure 8, one of the distinct characteristics of the social networks was the presence of a large orange cluster of trusters (1), which also had a connection to a smaller orange cluster (2) that had a half-circle shape. Four blue clusters of nontrusters (3)–(6) were shown as well. The size of these blue clusters was relatively smaller compared to the large orange cluster. There were also orange and blue nodes scattered around individually or having a small number of connections on the outskirts of the social network graph.

4.2.4 | The overall ratio of the two types of tweets

Figure 9 presents a quick insight into the overall ratio of the trusters and non-trusters. The total number of trusters was approximately six times more than that of nontrusters.

In Question #2, our focus is on the influential Twitter users who are responsible for spreading the dis/misinformation and also sharing correct/debunked information. Such users play an important role in disseminating both dis/misinformation and accurate information (Yang et al., 2019).

> Question #2: *Who are the significant Twitter users mainly responsible for spreading rumors the most? Who are spreading the debunking news articles in the social network to counter the spread of dis/misinformation?*

This question is centred around the people—represented as nodes in social networks—who consume, spread, and/or attempt to counter rumors that are being circulated in social media. We examined the centroid nodes in each cluster shown in the social network graph by using the mouse-over feature of *DismisInfoVis*. As shown in Figure 10, when a cursor is placed over a specific node in the social network graph, the

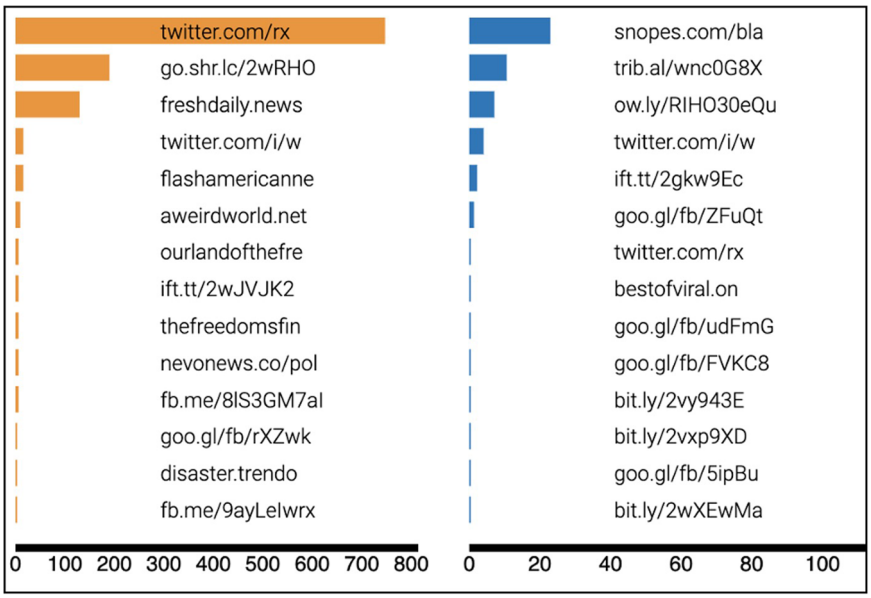

FIGURE 7 The URL view showing the frequently shared URLs among the trusters and the non-trusters. Note that snopes.com is on top of the URLs shared by non-trusters.

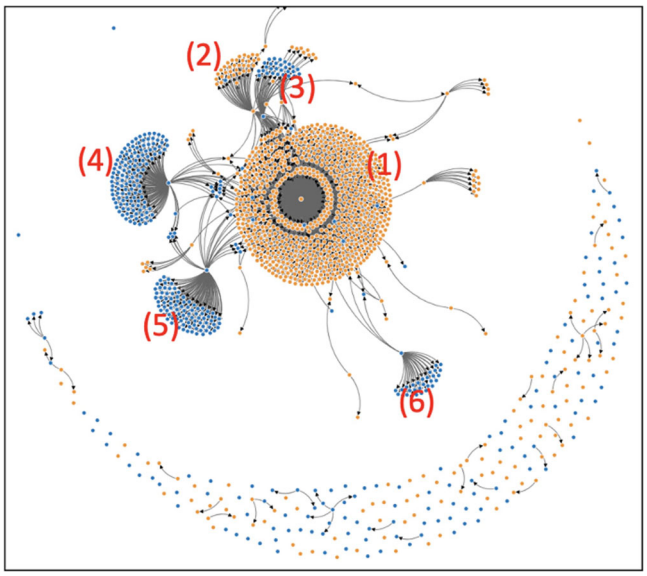

FIGURE 8 Social networks showing the clusters of trusters and non-trusters. Cluster numbers (1)–(6) are assigned.

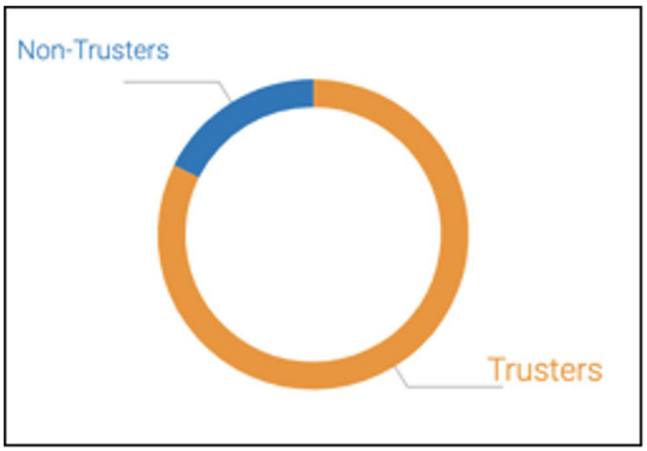

FIGURE 9 The Pie Chart views showing the overall ratio of trusters and nontrusters.

username and bio extracted from the person's Twitter user profile, and the information whether the person is a truster (e.g., Believes True > Believes Not True in Figure 10) or not, are displayed in a pop-up window. This feature enabled a quick look-up of node details, and it helped uncover the bases of the user behaviors for posting false news or debunking articles to counter the effect of false news.

4.2.5 | Twitter users responsible for spreading rumors

We examined the two significant centroid nodes from the clusters of trusters, which consisted of mostly orange nodes. Centroid nodes are those figures who potentially have a high level of influence on others in the cluster. Additionally, we examined four centroid nodes as well from the clusters of nontrusters, and these clusters consisted of mostly blue nodes (Table 3).

One of the common characteristics of the two Twitter users #1 and #2 in Table 3 was that they turned out to be Trump supporters when we examined their Twitter user bios. The centroid node #1 in Table 3, which was in the centre of the largest orange cluster, was a Canadian who supported Brexit and Trump. He included a link to a news article that originally had content about the false news (https://flashamericannews.com/news/black-lives-matter-thugs-blocking-emergency-crews-reaching-hurricane-victims/). However, this article was not accessible on March 8, 2021. The user bio of the other centroid node #2, which was in the center of the smaller cluster, described that he was a Christian and a Trump supporter by using multiple hashtags.

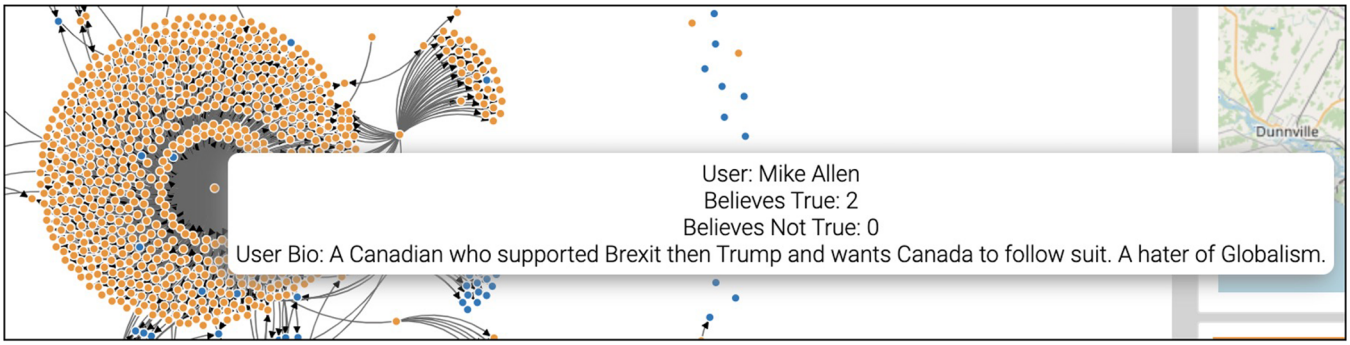

FIGURE 10 Username and bio are displayed on the pop-up window by hovering the mouse pointer over a specific node in the Social Network view.

TABLE 3 Twitter usernames from the centroid nodes of clusters.

No.	Cluster type	Cluster no.	Centroid user
1	Trusters	Figure 8 (1)	User: Mike Allen
2	Trusters	Figure 8 (2)	User: Deplorable Jim K
3	Nontrusters	Figure 8 (3)	User: snopes.com
4	Nontrusters	Figure 8 (4)	User: FactCheck.org
5	Non-trusters	Figure 8 (5)	User: Lead Stories
6	Non-trusters	Figure 8 (6)	User: Brasilmagic

4.2.6 | Twitter users spreading the *debunking* news articles

In contrast to the centroid nodes #1 and #2 from the orange clusters of trusters, who were individual people, the three centroid nodes #3-#5 from the blue clusters were the false news debunking websites such as snopes.com, factcheck.org, and Lead Stories. Only one centroid node #6 seemed to be an individual person. From this, we might postulate that investigating the veracity of news being circulated in social networks could be a very difficult and demanding task for individuals to undertake. Thus, such debunking of false news seemed to be conducted by organisations or at least a group of investigators. The first three—snopes.com, FactCheck.org, and Lead Stories—are rather non-partisan. However, node #6 was on the liberal democrat side based on the person's Twitter user bio. Considering that the trusters continuously posted and shared false news on Twitter until these fact-checking websites debunked the rumor and shared such debunking articles on Twitter, the role of these debunking sites in the current social media environment could be essential in fighting against dis/misinformation.

In Question #3, we examine the geographical distribution of trusters and non-trusters in the United States. Disaster management requires geospatial data throughout the four phases of the disaster lifecycle in that people who share their geospatial data with emergency management agencies may receive timely support, including accurate information about the disaster (Haworth & Bruce, 2015).

Question #3: *(1) Which geographic regions mainly have the trusters of dis/misinformation news? (2) What may have caused such geographic distribution of users* (metropolitan, urban, rural, or overseas areas)?

The false news in the BLM Rumor data set inaccurately blamed the BLM movement, and this type of news would provoke anger towards BLM among conservatives and possibly make liberals search for the truth of the news. Figure 11 shows the locations of the Twitter users—both trusters and non-trusters—in the United States. Fairly large orange circles, denoting that many people posted tweets around that area, appeared in Florida, Texas, Georgia, California, Michigan, and New Jersey areas. The circles in Texas and Florida were exceptionally large, which may mean the larger population trusted the false news as a true story. Many nontrusters were also shown on the Map view; however, they were rather dispersed all over the United States without forming clusters. Another characteristic of Figure 11 was that the nontrusters (blue) were appearing on both the east and west coast areas, which corresponded with the consensus that the East and West Coast people are more liberal and supportive of the Democratic Party.

5 | DISCUSSION AND LIMITATIONS

In our two-step methodology of detecting and understanding dis/misinformation events in Twitter, we had challenges in each step. In the detection step, we had the insufficient amount of training data set, which was specifically built to the rumor cases during a natural disaster. If we used only this training data set, we may have an overfitting model. To overcome this challenge, we incorporated TL for the Content Analysis module and improved the model performance (Table 2). Considering that a larger training data set could be built by collecting dis/misinformation events during natural disasters in the future, we expect that our model could be further fine-tuned to achieve better performances as we build a larger training set.

Another challenge during the detection model development was to understand what might be the optimal length of the event

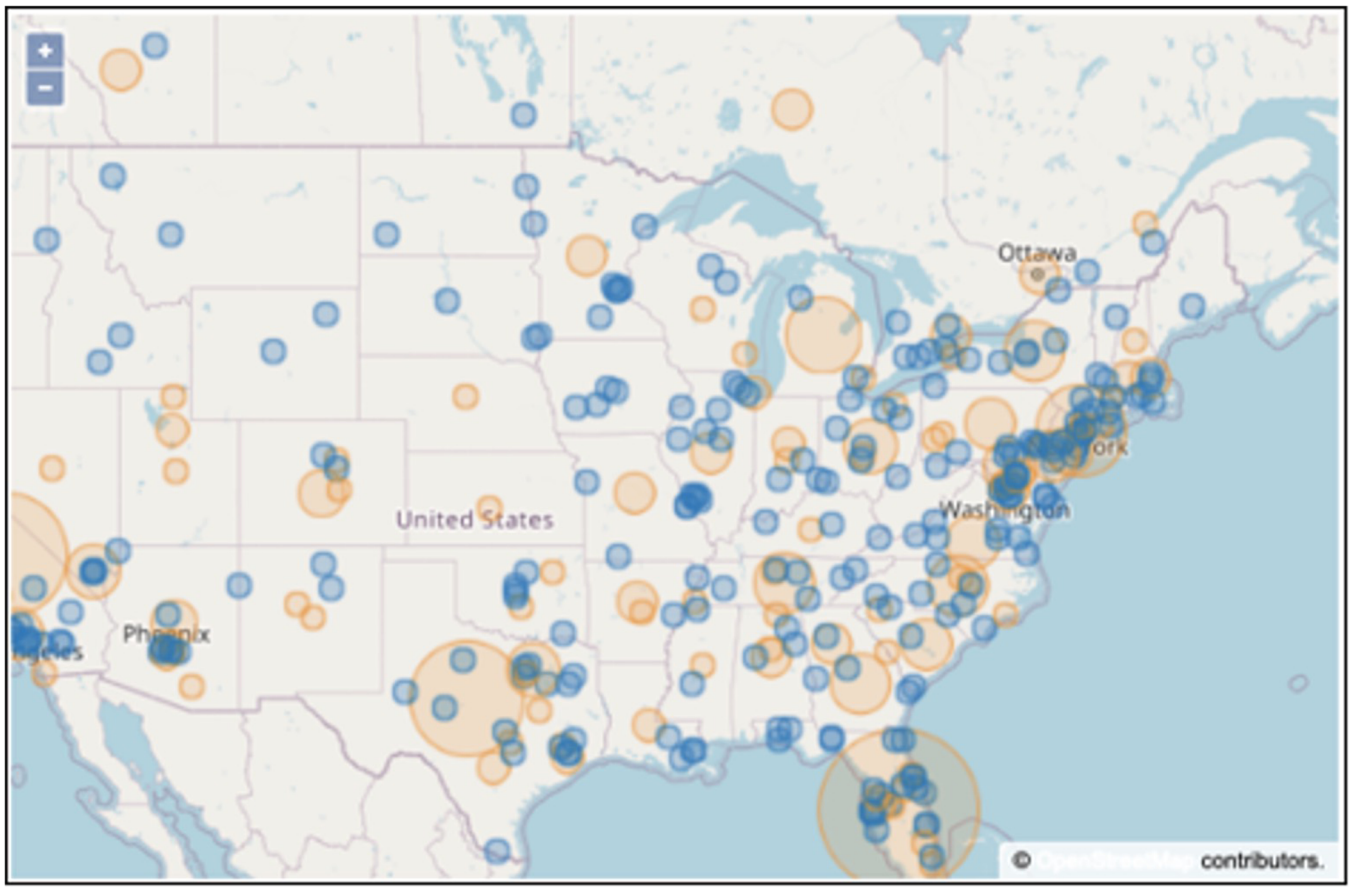

FIGURE 11 Twitter user locations visualized by the Map view.

period in the training data set for fine-tuning our model to achieve the best performance. When our detection model was trained with the fine-tuning data set, which had the event period length of 28-hour partitions, the model showed the best performance (accuracy of 91.47% and F-1 score of 90.89%) (Figure 5). When we trained the model with an event period longer than 28-hour partitions, we observed an adverse effect on the model performance without improving the accuracy. One possibility for this performance degradation is the influx of noisy data following the 28-hour point considering that marketers in social media platforms often abuse hashtags in dis/misinformation tweets for their marketing goals. Unraveling such time-dependent behavior might be an interesting future project for dis/misinformation detection studies.

In the understanding step, one of the challenges was to effectively present social media data related to false news events, which often have multiple dimensions. Our approach was to design a multi-view visualisation tool, *DismisInfoVis*. Various features in our tool provide effective navigation and interaction with the data, allowing users to uncover details of multi-faceted events. Distinguishing social media users into two types (trusters and non-trusters) and assigning colours to each type in orange and blue, respectively, was also an effective approach in understanding the overall distribution of users and their dynamics.

During this study, several limitations have been identified. First, our detection model has been trained with TL, which had a fine-tuning process with a relatively small data set. For this reason, our detection model might have been biased toward the false news events that occurred during the specific natural disaster event that was selected for this study. With a different set of false news events from different natural disasters, our detection model may perform differently. Additionally, in our visual analysis, we used a real-life dis/misinformation event with a political aspect to understand the network patterns, centroid users, and their locations, and so on. If a dis/misinformation event was selected from other domains—such as celebrities or the stock market—our visualisations might have presented vastly different results. Also, we did not capture a long-term pattern of the user interactions with dis/misinformation in this study. Considering that certain false news and images have been recurring in disasters, understanding the long-term impact of such false news and user interactions remain as our future study. Lastly, since information obtained from visualisations may significantly affect an analyst's sensemaking and decision-making processes with respect to dis/misinformation, it is critical to ensure the reliability and validity of dis/misinformation discovered with our data visualisations. However, ensuring such 'trust' in visualisations is still quite complex and remains underexplored (Mayr et al., 2019). Thomas and Kielman (2009) sought to shed light on this issue, concluding that measuring

and verifying the reliability of data visualisations remain major challenges for the visual analytics and data visualisation community. Thus, our future research plans can entail investigating the development of new metrics to measure trust for dis/misinformation visualisation; concurrently, we will seek to improve a systematic understanding of trust in dis/misinformation and fake news visualisations to ensure the reliability of presented information.

As for the practical and theoretical implications of this study, our approach could potentially improve crisis communication and management during disasters by providing awareness and multi-dimensional information. For example, early detection, identification of responsible parties, and spreading patterns) regarding various dis/misinformation events in social media. Starbird et al. (2018) suggest that the dissemination of dis/misinformation often occurs rapidly in social media compared to efforts to correct it, and crisis management teams should monitor social media streams and identify rumors early to proactively correct dis/misinformation and reduce its impact. Karami et al. (2020) also emphasize the importance of situational awareness in disaster preparedness, response, and recovery actions and present their analytical system, Twitter Situational Awareness.

Additionally, media literacy educators and media professionals have great interests in educating people to fight against the epidemic of dis/misinformation by promoting critical thinking abilities in literacy education. It is known that visualisation or visual analytics software tools can improve the critical thinking abilities of students. For example, Bodén and Stenliden (2019) found that the visual literacy aspects of their classes generated more discussions among students and those discussions could reflect students' logical thinking. Additionally, Shatri and Buza (2017) examined the effect of using visualisation tools in teaching and learning to develop critical thinking abilities. They measured students' understanding of complex computer science concepts that required critical thinking from lecture test scores. The test results showed an improvement in the performance of students who had taken lectures that had been prepared with various visualisations.

For these reasons, we expect that students who were educated with visualisation tools (e.g., *DismisInfoVis*) should be able to improve their critical thinking skills and potentially become better at discerning dis/misinformation that they may encounter in an online environment.

6 | CONCLUSION

In this study, we presented our two-step approach of detecting and understanding dis/misinformation events occurring in social media, especially during critical times such as disasters and crisis events. For the detection, our approach was to develop an RNN-based algorithm trained with TL so that it could make decisions whether a given cascade of tweets about a specific news item might be a dis/misinformation event or not. To better understand the multi-faceted nature of the dis/misinformation event, we presented *DismisInfoVis*, which integrates multiple views consisting of existing visualisations and charts. By combining these visualisation techniques, we could enable deeper insight into dis/misinformation tweets focusing on unforeseen connections of keypersons and causal relationships of different aspects of the social media data. With such deeper insight gained from the multiple views of our *DismisInfoVis* tool, users should be able to 'describe' an identified dis/misinformation event from multiple angles in more detail.

As for future work on our detection algorithm, we plan to create training data sets based on dis/misinformation events occurring in other types of disasters (man-made disasters or pandemics). This larger data set would allow us to further fine-tune our detection model, making it perform better for identifying the dis/misinformation events occurring in various types of natural and/or man-made disasters. Regarding our visualisations, we will recast these visualisations as not just analytic tools for experts and researchers, but social spaces for the general public. For this goal, we will contribute to the design and implementation of infographics and user-friendly visual representations for the public, which will enable nonexpert users to understand such dis/misinformation readily.

Our hope is that this study will contribute to improving the quality of information that is generated and shared on social media amid critical times, eventually helping both the affected and the general public recover from the impacts of a disaster.

ACKNOWLEDGEMENT

This work has been supported by the Russell B. Long Professorship in the College of Human Sciences and Education at Louisiana State University. It was also supported partially by the National Science Foundation Award HCC-2146523 and National Institute of Justice Award 2019-75-CX-K00.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon request.

There are two data sets supporting the findings of this study:

1. Rumdect data set (publicly available)
2. Disaster data set (developed for this study)

For #1, only the Twitter data portion, excluding the Weibo data, was used in this study. Considering the terms of Twitter data use, only the tweet IDs were shared and they need to be hydrated for accessing the entire data. For #2, both input data and label data are provided as two NumPy binary files.

The data sets can be accessed at:

https://drive.google.com/file/d/1iDcXn5mnNTWf9OO58NHKhzg0SkqVceUQ/view?usp=sharing

ORCID

Seungwon Yang http://orcid.org/0000-0003-4175-2316

REFERENCES

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. In W. Pedrycz, & S. M. Chen (Eds.), *Sentiment analysis and ontology engineering. Studies in computational intelligence*, (Vol 639). Springer, Cham. https://doi.org/10.1007/978-3-319-30319-2_13

Bodén, U., & Stenliden, L. (2019). Emerging visual literacy through enactments by visual analytics and students. *Designs for Learning*, *11*(1), 40–51.

Brandes, U. & Nick, B. (2011). Asymmetric relations in longitudinal social networks. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2283–2290.

Buscaldi, D. & Hernandez-Farias, I. (2015). Sentiment analysis on microblogs for natural disasters management: A study on the 2014 genoa floodings, *Proceedings of the 24th International Conference on World Wide Web* (pp. 1185–1188).

Cao, C., Yu-Ru lin, L., Xiaohua Sun, S., Lazer, D., Shixia Liu, L., & Huamin Qu, Q. u (2012). Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2649–2658.

Cao, N., Shi, C., Lin, S., Lu, J., Lin, Y. R., & Lin, C. Y. (2016). Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE transactions on Visualization and Computer Graphics*, *22*(1), 280–289.

Cao, R., Lee, R. K. W., & Hoang, T. A. (2020). DeepHate: Hate speech detection via multi-faceted text representations, *12th ACM conference on web science* (pp. 11–20).

Chen, S., Chen, S., Wang, Z., Liang, J., Yuan, X., Cao, N., & Wu, Y. (2016). D-map: Visual analysis of ego-centric information diffusion patterns in social media, *In IEEE conference on visual analytics science and technology (VAST)* (pp. 41–50). IEEE.

Chen, S., Lin, L., & Yuan, X. (2017). Social media visual analytics, *In Computer Graphics Forum* (Vol. 36, No. 3, pp. 563–587).

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259*.

Chua, A., Marcheggiani, E., Servillo, L., & Moere, A. V. (2014). FlowSampler: Visual analysis of urban flows in geolocated social media data, *International conference on social informatics* (pp. 5–17). Springer, Cham.

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4.

Cook, J., Ecker, U., & Lewandowsky, S. (2015). Misinformation and how to correct it, *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* (pp. 1–17).

Egelhofer, J. L., & Lecheler, S. (2019 Apr 3). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, *43*(2), 97–116.

Fallis, D. (2014). The varieties of disinformation. In L. Floridi, & P. Illari (Eds.), *The philosophy of information quality. Synthese library* (Vol 358). Springer, Cham. https://doi.org/10.1007/978-3-319-07121-3_8

Figueira, Á. & Oliveira, L. (2017). The current state of fake news: challenges and opportunities. *Procedia Computer Science*, *121*, 817–825.

Graves, A., Jaitly, N., & Mohamed, A. R. (2013). Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE workshop on automatic speech recognition and understanding* (pp. 273–278).

Haworth, B. & Bruce, E. (2015). A review of volunteered geographic information for disaster management. *Geography Compass*, *9*(5), 237–250.

Heer, J. & Boyd, D. (2005). Vizster: Visualizing online social networks, *IEEE Symposium on Information Visualization* (pp. 32–39). IEEE.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Horne, B. & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, *In Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1).

Hosseinimotlagh, S. & Papalexakis, E. E. (2018). Unsupervised content-based identification of fake news articles with tensor decomposition ensembles, *In Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*.

Hunt, K., Agarwal, P., Al Aziz, R., & Zhuang, J. (2020). Fighting fake news during disasters. *OR/MS Today*, *47*(1), 34–39.

Hutto, C. & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, p. 1).

Jain, G., Sharma, M., & Agarwal, B. (2018). Spam detection on social media using semantic convolutional neural network. *International Journal of Knowledge Discovery in Bioinformatics*, *8*(1), 12–26.

Karami, A., Shah, V., Vaezi, R., & Bansal, A. (2020). Twitter speaks: A case of national disaster situational awareness. *Journal of Information Science*, *46*(3), 313–324.

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents, *In International conference on machine learning* (pp. 1188–1196). PMLR.

Liu, Y. & Wu, Y. F. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks, *In the Thirty-second AAAI conference on artificial intelligence*.

Loyola-González, O., López-Cuevas, A., Medina-Pérez, M. A., Camiña, B., Ramírez-Márquez, J. E., & Monroy, R. (2019 Mar 1). Fusing pattern discovery and visual analytics approaches in tweet propagation. *Information Fusion*, *46*, 91–101.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K. F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th international joint conference on artificial intelligence (IJCAI 2016)* (pp. 3818–3824). Research Collection School Of Computing and Information Systems. https://ink.library.smu.edu.sg/sis_research/4630

Maddock, J., Starbird, K., Al-Hassani, H. J., Sandoval, D. E., Orand, M., & Mason, R. M. (2015). Characterizing online rumoring behavior using multi-dimensional signatures, *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 228–241).

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C (2011). Twitinfo: Aggregating and visualizing microblogs for event exploration, *In Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 227–236).

Mayr, E., Hynek, N., Salisu, S., & Windhager, F. (2019). Trust in Information Visualization, *TrustVis@ EuroVis* (pp. 25–29).

Naeem, S. B., Bhatti, R., & Khan, A. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information and Libraries Journal*, *38*(2), 143–149.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th international conference on computational linguistics* (pp. 3391–3401). Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pierri, F. & Ceri, S. (2019). False news on social media: A data-driven survey, *ACM SIGMOD Record 48*(2), 18–27.

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). *Declare: Debunking fake news and false claims using evidence-aware deep learning. arXiv preprint arXiv:1809.06416.*

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of 56th annual meeting of the association for computational linguistics (ACL 18)*.

Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., & Mohanty, S. (2019). Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. *International Journal of Disaster Risk Reduction*, *37*, 101176.

Qazvinian, V., Rosengren, E., Radev, D., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs, *In Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1589–1599).

Rubin, V. L., Chen, Y., & Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4.

Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection, *In Proceedings of the ACM on Conference on Information and Knowledge Management* (pp. 797–806).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Sadri, A. M., Hasan, S., Ukkusuri, S. V., & Cebrian, M. (2018). Crisis communication patterns in social media during Hurricane Sandy. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(1), 125–137.

Sapir, D. G., & Lechat, M. F. (1986). Reducing the impact of natural disasters: Why aren't we better prepared? *Health Policy and Planning*, *1*(2), 118–126.

Seref, O., Seref, M. M., & Hong, S. (2020). Context Map Analysis of Fake News in Social Media: A Contextualized Visualization Approach, *HICSS* (pp. 1–9).

Shao, C., Hui, P. M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLoS One*, *13*(4), e0196087.

Shatri, K., & Buza, K. (2017). The use of visualization in teaching and learning process for developing critical thinking of students. *European Journal of Social Sciences Education and Research*, *4*(1), 71–74.

Shu, K., Mahudeswaran, D., & Liu, H. (2019). FakeNewsTracker: A tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, *25*(1), 60–71.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017 Sep 1). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Starbird, K., Dailey, D., Mohamed, O., Lee, G. & Spiro, E. S. (2018). Engage early, correct more: How journalists participate in false rumors online during crisis events. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–12).

Sun, D., Wang, M., & Li, A. (2019). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(3), 841–850.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506.

Tandoc, Jr., E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news" A typology of scholarly definitions. *Digital journalism*, *6*(2), 137–153.

Thomas, J. & Kielman, J. (2009). Challenges for visual analytics. *Information Visualization*, *8*(4), 309–314.

Viégas, F., Wattenberg, M., Hebert, J., Borggaard, G., Cichowlas, A., Feinberg, J., Orwant, J., & Wren, C. (2013). Google+ ripples: A native visualization of information flow, *In Proceedings of the 22nd international conference on World Wide Web* (pp. 1389–1398).

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1079–1088).

Volkova, S., & Jang, J. Y. (2018). Misleading or falsification: Inferring deceptive strategies and types in online news and social media, *In Companion Proceedings of the The Web Conference* (pp. 575–583).

Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter, *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 647–653).

Vosoughi, S., Roy, D., & Aral, S. (2018 Mar 9). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wang, W. Y. (2017). *"liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648*.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection, *In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 849–857).

Wardle, C. & Derakhshan, H. (2018). Thinking about 'information disorder': Formats of misinformation, disinformation, and mal-information. In C. Ireton, & J. Posetti (Eds.), *Journalism,'fake news' & disinformation* (pp. 43–54). UNESCO.

Wu, L. & Liu, H. (2018). Tracing fake-news footprints: Characterizing social media messages by how they propagate, *In Proceedings of the eleventh ACM international conference on Web Search and Data Mining* (pp. 637–645).

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019 Nov 26). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, *21*(2), 80–90.

Wu, Y., Cao, N., Gotz, D., Tan, Y. P., & Keim, D. A. (2016 Sep 27). A survey on visual analytics of social media data. *IEEE Transactions on Multimedia*, *18*(11), 2135–2148.

Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., & Hu, X. (2019). Xfake: Explainable fake news detector with visualizations, *In The World Wid e Web Conference* (pp. 3600–3604).

Yang, J., & Luttrell, R. (2022). Digital Misinformation & Disinformation: The Global War of Words, In *The Emerald Handbook of Computer-Mediated Communication and Social Media* (pp. 511–529). Emerald Publishing Limited.

Yang, Y., Zhang, C., Fan, C., Yao, W., Huang, R., & Mostafavi, A. (2019). Exploring the emergence of influential users on social media during natural disasters. *International Journal of Disaster Risk Reduction*, *38*, 101204.

Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., Li, D., Wu, J., & Havlin, S. (2020). Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, *9*(1), 7.

How to cite this article: Yang, S., Chung, H., Singh, D., & Shams, S. (2023). A two-step approach to detect and understand dismisinformation events occurring in social media: A case study with critical times. *Journal of Contingencies and Crisis Management*, 1–17. https://doi.org/10.1111/1468-5973.12483