Information Systems Frontiers (2023) 25:473–492
https://doi.org/10.1007/s10796-022-10327-9

A Theory-based Deep-Learning Approach to Detecting Disinformation in Financial Social Media

Wingyan Chung[1] · **Yinqiang Zhang**[2] · **Jia Pan**[2]

Accepted: 12 August 2022 / Published online: 12 September 2022

Abstract
The spreading of disinformation in social media threatens cybersecurity and undermines market efficiency. Detecting disinformation is challenging due to large volumes of social media content and a rapidly changing environment. This research developed and validated a theory-based, novel deep-learning approach (called TRNN) to disinformation detection. Grounded in social and psychological theories, TRNN uses deep-learning and data-centric augmentation to enhance disinformation detection in financial social media. Temporal and contextual information is encoded as specific knowledge about human-validated disinformation, which was identified from our unique collection of 745,139 financial social media messages about four U.S. high-tech company stocks and their fine-grained trading data. TRNN uses multiple series of long short-term memory (LSTM) recurrent neurons to learn dynamic and hidden patterns to support disinformation detection. Our experimental findings show that TRNN significantly outperformed widely-used machine learning techniques in terms of precision, recall, F-score and accuracy, achieving consistently better classification performance in disinformation detection. A case study of Apple Inc.'s stock price movement demonstrates the potential usability of TRNN for secure knowledge management. The research contributes to developing novel approach and model, producing new information systems artifacts and dataset, and providing empirical findings of detecting online disinformation.

Keywords Design science · Machine learning · Deep learning · Temporal recurrent neural network · Social media · Sequence prediction · Disinformation detection · Cybersecurity · Secure knowledge management · Financial market

1 Introduction

Concern about disinformation in social media is rising. Results of 150 interviews of industry practitioners, subject-matter experts, and government officials across nine countries show that disinformation campaigns on social media will likely increase in the future (Cohen et al., 2021). The spread of disinformation about COVID-19 vaccination (Bond, 2021), financial market (Commission, U.S.E, 2015), and political election (Abrams, 2019) (among others) threatens cybersecurity and seriously undermines social confidence (Chung, 2016). However, detecting disinformation in social media can be challenging. The large volumes of social media content and rapidly changing market indicators (e.g., stock prices, sales) make it difficult to accurately identify disinformation, defined as "false information that is purposedly spread to deceive people" (Lazer et al., 2018). In recent years, artificial intelligence (AI) technologies (machine learning and deep learning in particular) have been used to detect fake news and misinformation (e.g., Reis et al., 2019; Ducci et al., 2020). Despite their capability to automatically learn from data, a lack of data sense-making and data-centric augmentation is prevalent among various high-stake AI applications (Sambasivan et al., 2021). Social and psychological theories can be used to explain human behavior when faced with disinformation. But their use to enhance the application of AI to detect disinformation is not widely available.

This research seeks to answer several questions: (1) How can machine learning techniques and social and

✉ Wingyan Chung
wchung@uttyler.edu

Yinqiang Zhang
zyq507@connect.hku.hk

Jia Pan
jpan@cs.hku.hk

[1] The University of Texas at Tyler, Texas, USA

[2] The University of Hong Kong, Pokfulam, Hong Kong

psychological theories be used to support detection of online disinformation? (2) How can a theory-based, data-driven AI approach be developed to detect disinformation in social media? (3) Compared with widely-used machine-learning techniques, how does the approach perform in detecting disinformation in financial social media?

To answer these questions, we developed and validated a novel, theory-based deep learning approach, called temporal recurrent neural network (TRNN), to address the challenges in disinformation detection from social media. TRNN advances the classification capability of traditional AI techniques and was developed based on social and psychological theories and data-centric augmentation. To evaluate the usability and effectiveness of TRNN, we built a unique collection of 745,139 social media messages and fine-grained stock price data (of 277 contiguous trading days) of four U.S. high-tech companies. Our experiments empirically compared TRNN with widely-used machine learning techniques. We provide a case study to demonstrate the potential application to secure knowledge management. The results indicate a high generalizability of TRNN in other data types and domains than finance, and have strong implication for design science research and information systems practice.

2 Literature Review

Disinformation is used increasingly to manipulate human perception (Cybenko & Cybenko, 2018) and has raised concern in the academia, governments, and industries (Cohen et al., 2021; Chung, 2016; Del Vicario et al., 2016; Vosoughi et al., 2018). Social and psychological theories can be used to explain the proliferation of disinformation. Various techniques and methods have been developed to address the issues of disinformation. This review tries to identify strengths, weaknesses, and research gaps from the literature. Tables 1 and 2 summarize respectively methods of disinformation detection and components of systems used in disinformation detection research.

2.1 Theoretical Background on Disinformation

Theories can provide clues to explain human behavior when faced with disinformation (Kandhway & Kuri, 2017; Stage, 2013). Social, economic, and psychological theories postulate that humans tend to rely on heuristics and incentives when making judgments. One common heuristic is to follow the crowd in the social environment. Social Contagion Theory postulates that humans behave based on the information available to them (e.g., rational thought, experience) (Le Bon, 1895; Wheeler, 1966). In an online environment, information is often contradictory due to a lack of agreement, forcing the individuals to look for additional cue. Emergent Norm Theory further posits that new norms happen when group leaders and members agree on a new normative status or purpose for the group (Turner & Killian, 1957). These norms and cue become heuristics for judging the reliability of online information.

A second heuristic is to rely on neighbors, experts, or famous social actors. Theories based on social positions of humans describe persons as nodes in a social network and their relationship as links. Each node (or each link) can be characterized by common attributes within all the nodes (or all links). Homophily Theory states that network nodes may behave similarly if they share similar nodal attributes (McPherson et al., 2001). Social Impact Theory postulates that the influence of a node in a social network is a multiplicative function of the strength, immediacy, and count of all nodes in the network (Latané, 1981; Sedikides & Jackson, 1990). Social Interaction Theory states that people make decisions based on their social neighbors' decisions (Becker, 1974).

A third heuristic is to decide based on the expected rewards. Social Exchange Theory states that people engage in social interaction with an expectation that it will bring them some rewards, such as respect, approval, or recognition (Emerson, 1976). These heuristics are often exploited by malicious actors who may manipulate the online environment and messages to fabricate some consensus, to distort opinions or messages of famous persons, or to present lucrative payback (e.g., from stock investments). In addition, malicious actors manipulate incentives in a financial market to earn illegal profits. A cornerstone of the Capital Asset Pricing Model (CAPM) (Sharpe, 1964) relates market systematic risk and investors' expected return of an asset (e.g., stock portfolio), and can be used to explain investors' expectation of seeking abnormal returns that compensate for risk and the time value of money.

Several differences exist between online communication and offline (in-person) communication that may facilitate the creation and spread of disinformation: the quality of human interaction, and the speed and geographic spread of messages (Li et al., 2017; Quan-Haase, 2016). As a result, online communication changes the human perception of time fundamentally. Elements of time that are relevant to the spread of information throughout mass groups include recency and primacy effects (Hovland, 1957; Miller & Campbell, 1959). It has been found that people are more likely to share information when they are exposed to that information recently (recency) and when it is important to them (primacy) (Gino et al., 2009; Ngai et al., 2015). Therefore, malicious actors may use temporal information strategically in online messages to spread disinformation. Techniques that can use temporal information strategically from online content and context can possibly help to detect disinformation. The following review AI techniques for detecting malicious content.

Table 1 A summary of methods for disinformation detection

Category	Sub-Category	Description	Strength	Weakness	Work
Manual Pattern Matching	Rule based methods	Rule-based methods utilize user white lists or keyword blacklists and manually crafted rules to detect disinformation	Rule based methods are able to easily incorporate expert's domain knowledge into the blacklists to guide the disinformation detection	Trivial blacklist keywords matching tend to be error-prone when the context contains negation, sarcasm, etc.	(Lee et al., 2018; Ribeiro et al., 2018)
Network Based Methods	Semantic network methods	Semantic based methods capture the structure of the knowledge network and use it to infer the truthfulness of given information	Network based methods are promising in accuracy of statements of the form "A is B" and it also reveals the topology dynamics of the social connections	If the entity to be checked is not in the existing database, the disinformation detection can not be done	(Ciampaglia et al., 2015; Del Vicario et al., 2016; Ruchansky et al., 2017; Shi & Weninger, 2016)
	Diffusion based methods	Diffusion based methods look for critical network links or nodes to control the spread of disinformation	Once suspicious accounts and initial spread of disinformation are identified, the epidemic of devastating consequences can be avoided	Extremely computationally heavy in the context of billions of nodes and links	(Kuhlman et al., 2013; Nguyen et al., 2019; Pham et al., 2019; Vosoughi et al., 2018)
Machine Learning (ML) Methods	Traditional machine learning methods	Traditional ML methods use a collection of labeled instances to train a classifier such as support vector machine (SVM), Decision Tree and logistic regression to learn the disinformation patterns	With different variations and kernel tricks, traditional machine learning methods are flexible in different situations to handle the disinformation features	Traditional ML methods might not model the complex social dynamics exhibited in social media.	(Delort et al., 2011; Feng et al., 2012; Shu et al., 2017; Reis et al., 2019; Giasemidis et al., 2018; Langley et al., 2021)
	Deep Learning (DL) Methods	DL methods use multi-layered massive computational units to learn disinformation features with back-propagation algorithm. Representative techniques include RNN, LSTM, and CNN.	Deep learning methods are powerful in modeling complex non-linear social dynamics	Difficulty in model interpretation and explanation. Need large amount of labeled training data	(Zhang et al., 2019; Zhang et al., 2015; Volkova et al., 2017; Kumar et al., 2021)

2.2 Artificial Intelligence Techniques

Prior research uses rule-based methods that rely on user white lists, keyword blacklists, and hand-crafted rules to detect disinformation (Lee et al., 2018; Owda et al., 2017). Word usage, part-of-speech tags, syntax, and bag-of-word approaches are used to learn the patterns of disinformation messages (Feng et al., 2012; Markowitz & Hancock, 2016). Since rule-based methods rely primarily on n-grams or syntactical analysis, contextual meaning in the word sequence may not be captured (Conroy et al., 2015). Creating and maintaining hand-crafted rules is time-consuming and lacks generalizability.

2.2.1 Network-based Methods

Aside from content aspect of the online messages, network-based methods capture behaviors and structure of the online communities to help detect disinformation. Semantic network represents semantic relations between entities in a network (Sowa, 1987). By aggregating the existing information in network (e.g., profiles, labeled users, and confirmed statements), upcoming new messages can be fact-checked quickly with high accuracy (Dave, 2013). By analyzing the network topology, scores can be assigned to entities based on their relevance and distance to classify disinformation (e.g., Ciampaglia et al., 2015; Ruchansky et al., 2017).

Table 2 Components of systems used in disinformation detection research

Category	Article	System Input	Feature	Technique	Results
Textual feature	(Giasemidis et al., 2018)	Twitter messages	n-grams, part-of-speech	Semi-supervised learning algorithm	Improved speed with less labeled data for stance classification
	(Wang et al., 2018)	Weibo messages	Event features	Adversarial neural network	Improved accuracy on fake news detection
	(Vosoughi et al., 2017)	Twitter messages	Liguistic and network features	Hidden Markov model	Improved accuracy on unverified rumours
	(Liu et al., 2018)	Twitter and Weibo messages	Linguistic features and temporal feature	Neural network method	Improved performance in detecting disinformation
Network feature	(Nguyen et al., 2019)	Twitter, Pokec, DBLP nodes and edges	Network node activation feature	Linear threshold model	Reduced complexity in stopping cyber-epidemics
	(Tong et al., 2017)	Wiki, YouTube, Epinions nodes and edges	Neighbour influence	Randomized Algorithm	Reduced complexity in rumor blocking
	(Yan et al., 2019)	Wikipedia, Slashdot, Google+ nodes and edges	Node disseminating influence	Link deletion algorithm	Improved approximation of minimizing rumour spread
	(Zhang et al., 2016)	Twitter, Epinion, Slashdot nodes and edges	Network propagation feature	Network monitor placement algorithm	Reduced # of monitors to place in social network in detecting online misinformation

Diffusion-based methods study the propagation pattern of disinformation in social networks. A random walk algorithm is used to remove most effective links in online social network to prevent the spread of disinformation (Nguyen et al., 2019). A study using differential diffusion found that false news diffused significantly faster, deeper and more broadly than legitimate news (Vosoughi et al., 2018). Edge removal was used in Kuhlman et al. (2013) as a heuristic to limit the disinformation diffusion.

2.2.2 Machine Learning Methods

Machine learning (ML) is the set of theoretical and practical approaches for designing machines that learn autonomously from data without explicitly being programmed (Mitchell, 1997). A subcategory of ML, supervised ML techniques use labeled data instances, each consisting of a feature vector X and an output label y, to infer a mathematical function that maps from X (e.g., sales transactions) to y (e.g., fraud or non-fraud). These techniques have been used to detect fake news (Cybenko & Cybenko, 2018).

Traditional ML techniques include such diverse methods as k-nearest neighbor classifier, support vector machine (SVM), random forests (RN), and XGBoost (Reis et al., 2019). For instance, a Naive Bayes classifier was used to detect information that violates community guidelines (Delort et al., 2011). SVM was used with syntactic features to distinguish deception from benign information (Feng et al., 2012) and was shown to outperform logistic regression, decision tree, and neural networks in classifying textual news headlines into true or fake news (Langley et al., 2021). "Event adversarial neural networks," a supervised ML technique that uses massive interconnected computational units, was used to perform multi-modal fake news detection (Wang et al., 2018). Convolutional neural network (CNN) achieves good performance in general sentence classification tasks such as sentiment analysis (e.g., online reviews) (Zhang et al., 2015, 2019). Sequence models, such as Markov models and Kalman filters, deal with sequential data but are ill-equipped to learn long-range dependencies (Alzaidy et al., 2019).

2.2.3 Deep Learning

Deep learning – the use of multi-layered, interconnected computational units to infer non-linear functions (as found in CNN, RNN, and LSTM) – has dramatically advanced different application domains, most notably computer vision and speech recognition (LeCun et al., 2015). Deep learning (DL) models have been developed and applied to detecting different forms of false information, such as rumors, fake news, and misinformation. For example, a tree-structured classifier, known as "cascade-LSTM," was developed to learn from retweet behavior and to predict the veracity of 2,156 Twitter cascades that contain misinformation, giving a 2.8% improvement over the best baseline classifier (Ducci

et al., 2020). To detect algorithmically modified images and videos (or "deepfakes"), a decentralized blockchain framework uses multiple LSTM networks to support tracing and tracking of a digital content's historical provenance (Chan et al., 2020). In a study of classifying hate speeches on Twitter, a CNN-based model and a pre-trained VGG-16 network were used to process text (encoded with Glove embedding vectors) and image data respectively (Langley et al., 2021). Another application called "F-NAD" uses an ensemble technique of recurrent neural networks (LSTM and GRU) to classify the origins of news articles into either fake or real sources (Barua et al., 2019). Multiple neural network approaches (CNN, LSTM, bidirectional LSTM) were compared in detecting fake news (collected from Twitter and PolitiFact), finding that CNN plus bidirectional LSTM ensembled network with attention mechanism achieved the highest accuracy of 88.78% (Kumar et al., 2020). Another study comparing CNN, RNN, and LSTM and a tree-structured RNN produced similar findings, showing superior performance of bi-directional LSTM model over other methods (Bahad et al., 2019). The aforementioned studies indicate superior performance of bi-directional LSTM among other DL techniques due to its ability to address vanishing gradient and long-term memory problems. Other than using variations of RNN, the attention mechanism has gained much traction due to its high performance in language translation (Vaswani et al., 2017). However, due to a model-centric design, attention-based DL algorithms are limited by the quality and quantity of available data, and their adoption is limited by the level of trust afforded by human users (Gennatas et al., 2020). Disinformation detection presents additional challenges due to the intentional deception found in communities seeking profits (e.g., financial investment).

2.2.4 Feature Representation Learning

Disinformation detection requires understanding both the content and context of the information being used to deceive recipients. Prior research considered temporal characteristics such as content freshness and the period of time to classify rumors into different categories (Knapp, 1944). Temporal features play a role during breaking news events. In early stages of news release, people tend to support unverified rumor but as time goes on, a shift occurs to debunk false rumors (Zubiaga et al., 2016). Burstiness and linguistic, temporal, and structural features of rumor propagation were studied, finding that the popularity of rumors fluctuates over time in different platforms of social media (Kwon & Cha, 2014; Kwon et al., 2013). In a feature stability analysis, structural and temporal features were found to distinguish rumors from non-rumors over a long-term window, whereas user and linguistic features performed well in the early stages of rumor propagation (Kwon et al., 2017).

Prior work has also used textual features in disinformation detection. By using typical text phrases to express skepticism about factual claims, a study found that rumor clusters can be detected at about a third of the top 50 clusters in Twitter (Zhao et al., 2015). Bigram-based textual features were used to identify rumors in microblogs (Qazvinian et al., 2011). LSTM (Hochreiter & Schmidhuber, 1997; Mikolov et al., 2010; Yu et al., 2019) has been used to represent features using a word encoder, a sentence encoder and a headline-body encoder in detecting fake news in 2016 US election (Singhania et al., 2017). CNN and RNN were applied to detecting fake news in the event of Sydney siege, Ottawa shooting, Germanwings crash, etc. (Ajao et al., 2018). While prior studies examined different aspects of rumors and fake news, disinformation features that are highly interrelated (e.g., market prices, media content, and temporal features) such as those in financial social media are not studied widely.

2.3 Summary of Research Gaps

The literature review has identified a diverse set of theories, methods, techniques, and features used in detecting fake news, misinformation, and deceptive information. Table 1 provides a summary of different categories of methods. Table 2 summarizes various system input, features, and techniques. Manual pattern matching can yield accurate and intuitive results, but does not scale up to rapid growth of online data. Network-based methods can reveal online community structure and user behavior, but do not help to identify whether the behavior constitutes disinformation.

Deep learning techniques, and RNN and LSTM in particular, have shown promise in detecting rumors and fake news (Chan et al., 2020; Ducci et al., 2020). However, current studies on recurrent neural networks are mainly empirical explorations and lack explicit knowledge (p. 1261, Yu et al., 2019), thus requiring a richer and clearer representation of the complex features that may appear in disinformation. Although various features were studied in previous works (Singhania et al., 2017; Wang et al., 2018), prior research does not consider temporal dependencies of textual features in social media and does not incorporate features whose values interrelate highly with the economy (e.g., financial market). These requirements call for new DL approaches and representation that simultaneously address the temporal and contextual needs of disinformation detection.

Another research gap is inadequate application-domain expertise among artificial intelligence (AI) practitioners to support data sense-making. Results of interviews with 53 AI practitioners in high-stake domains show that a lack of domain expertise is experienced by 43.5% of the practitioners and can cause negative downstream data issues

in deployment (Sambasivan et al., 2021). A lack of well-defined ground truth and high-quality data can hamper the training and application of DL models to detect disinformation. Unfortunately, existing research focuses primarily on model development and does not use data-centric augmentation (Gennatas et al., 2020) to enhance accurate detection of disinformation.

3 Temporal Recurrent Neural Network

This section describes a novel theory-based, data-driven temporal recurrent neural network (TRNN) approach that is developed based on a design science paradigm (Ericsson & Simon, 1993) to address the aforementioned gaps. Grounded in social and psychological theories, TRNN dramatically expands the power of traditional recurrent neural networks (RNNs) by incorporating contextual and temporal information from social media, financial stock prices, general market trends, and the complex interactions among these factors. A unique representation of temporal and contextual information in disinformation allows TRNN to encode specific knowledge for the detection. The design artifacts include the TRNN approach and model, an instantiation of the model with application to disinformation detection in financial social media, and the related disinformation dataset.

3.1 Design Rationale

The design rationale of TRNN is three-fold: (1) to advance the architecture of traditional approaches by using theories and specific temporal and contextual information, (2) to enrich and clarify the knowledge representation of complex features found in disinformation, and (3) to enhance detection performance by using data-centric augmentation. Addressing the needs for modeling complex disinformation in social media, TRNN captures temporal and contextual information by using the posting times of and neighboring words appearing in messages, and considers various behavioral attributes as postulated by social, economic, and psychological theories (Becker, 1974; Hovland, 1957; Li et al., 2017; Miller & Campbell, 1959; Quan-Haase, 2016; Turner & Killian, 1957) reviewed in Section 2.1.

Different from prior work that uses time windows of stock trading information (e.g., Islam et al., 2018), TRNN learns from social media messages organized into contiguous, time-based scenarios, each spanning five minutes and containing all messages posted during that time span. Only scenarios in which an abnormal return of the interested financial stocks is observed are considered in disinformation detection, because malicious hackers often launch cyber attacks amidst market turbulence to gain illegal profits (e.g., Commission, U.S.E, 2015). Based on social exchange theory (Emerson, 1976), TRNN considers abnormal returns in identifying disinformation because people are lured by these returns to engage in a social media environment. To support the learning of dynamic information from scenarios that span a long time, TRNN concatenates multiple RNNs to enable deep learning of long-range textual and temporal dependencies for disinformation detection. Different from prior research that relies primarily on generic word embeddings and trading data (e.g., Seth & Chaudhary, 2020), TRNN considers enriched data from social media, financial market, temporal information, and multiple independent human annotations to increase prediction accuracy. Contextual information is obtained from time-based scenarios and can uniquely model changes in social media discussion and their impact on financial stock prices. The use of positive pointwise mutual information (PPMI, to be explained below) helps to extract semantic content from noisy messages (Jurafsky & Martin, 2020). We implemented the TRNN model in a proof-of-concept system, whose three modules and architecture are shown in Fig. 1 as explained below.

3.2 Module 1 - Feature Representation with PPMI

Module 1 converts raw textual data into numerical features that encode content and contextual information. To capture contextual information from voluminous text, TRNN uses a context of 11 words (five words before and five words after a target word, plus the target word) to identify neighboring words and to compute the positive pointwise mutual information (PPMI) of a target word. The size of the context was determined based on empirical testing that balances between the extent of context being considered in a word of a social media post and specific information presented in the word. Based on social contagion theory, TRNN uses the context of social media text to identify potential disinformation because humans behave based on the information they receive (Wheeler, 1966; Le Bon, 1895). TRNN also uses emergent norm theory (Turner & Killian, 1957) to characterize humans' behavior of following opinion leaders in a social media environment.

PPMI is a measure of the contextual information of a word with reference to the collection of words used in the corpus. Each word is represented as a vector of numeric values that reflect the word usage in relation to other words. Shown in Eq. 3, PPMI_{ij} is a measure of the likelihood of co-occurrence of words i (target word) and j (contextual word), compared with what would be expected if they were independent (Jurafsky & Martin, 2016). The likelihood that word i occurs in the context of word j is computed as P_i, whereas the likelihood that word j occurs in the context of word i is computed as P_j, as shown in Eq. 2, in which f_{ij} is the frequency of co-occurrence of words i and j in the same context. PPMI, an improved version of Pointwise Mutual

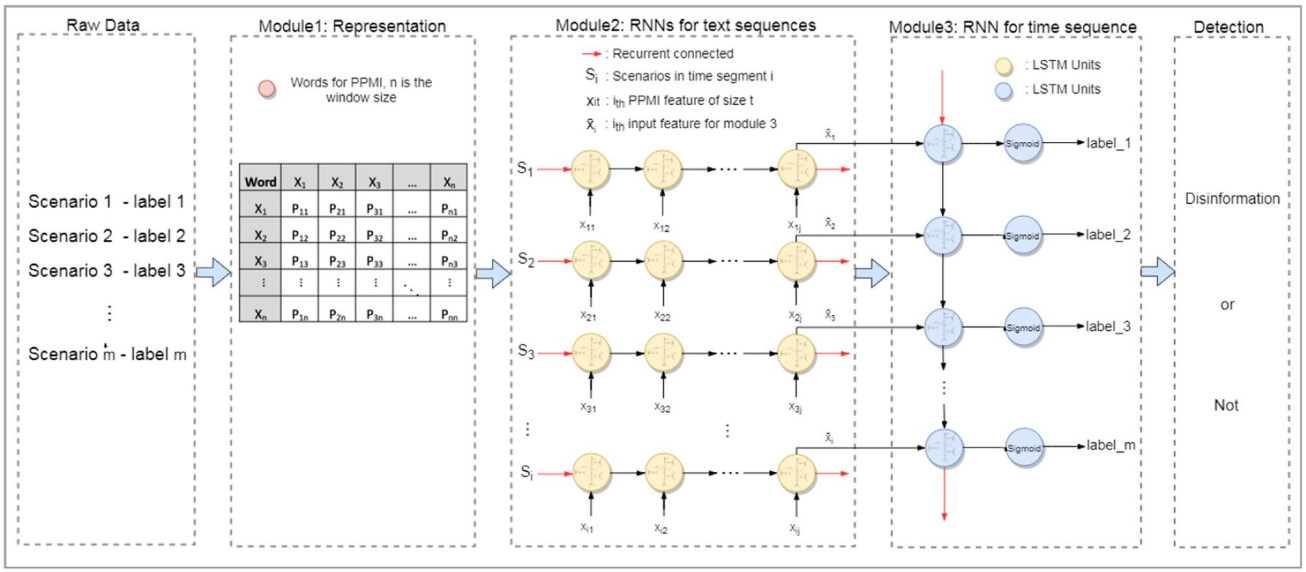

Fig. 1 Architecture of the temporal recurrent neural network approach

Information (PMI), ranges from zero to infinity and replaces negative PMI values (which carry no semantic meaning) by zeros. PPMI overcomes some limitations arising from the use of word frequency alone that may ignore word context (e.g., "Apple" occuring near "computer" carries a different meaning than "Apple" occuring near "juice."). Our human-annotated data collection (as described below) allows PPMI to encode more specific knowledge about disinformation than general word embeddings (e.g., GloVe) or traditional information retrieval methods (e.g., tf-idf) (Kowsari et al., 2019; Salton & McGill, 1983).

$$P_{ij} = \frac{f_{ij}}{\sum_i \sum_j f_{ij}} \tag{1}$$

$$P_i = \frac{\sum_j f_{ij}}{\sum_i \sum_j f_{ij}}; P_j = \frac{\sum_i f_{ij}}{\sum_i \sum_j f_{ij}} \tag{2}$$

$$\text{PPMI}_{ij} = \max\left(\log_2 \frac{P_{ij}}{P_i \times P_j}, 0\right) \tag{3}$$

3.3 Module 2 - RNN Using PPMI Textual Features

Module 2 transforms the output of Module 1 into sequences of activation values computed by recurrent neural network (RNN) cells. Figure 1 shows that the PPMI values of textual features are fed into multiple RNNs, each representing a time segment containing all the tweets in their time period. Two same-length segments of trading times are considered: (1) from 9:30am to 12:45pm (morning) and (2) from 12:45pm to 4:00pm (afternoon). Each RNN node represents one time segment. A single long short-term memory (LSTM) node is used in the RNN cells. The formulas that are used in Module 2 to compute the output values are given in Eqs. 4, 5, 6, 7, 8 and 9, in which x_t is the input feature vector of PPMI values; f_t is the forget gate that controls the extent to which textual/temporal information is not stored in the RNN cells; o_t is the output gate that controls the extent to which the output values in Module 2 are used to feed in Module 3 and activation function for final classification respectively; x_t is the PPMI input feature vector; h_t is the hidden state vector that holds previous textual/temporal information the neural network has been presented during model training; and c_t is the cell state vector that transfers relative information as a highway in textual/temporal sequences. W, U and b are the weights matrices and bias vectors for disinformation detection that are learned in the training process. σ_g and σ_c are the sigmoid and hyperbolic tangent activation functions (Han & Moraga, 1995) to map the output values to probabilities.

Compared with the standard feedforward neural network, LSTM can be used to memorize and learn the feedback information in both text and time sequences. LSTM overcomes the "vanishing gradient problem" in traditional RNN by applying multiplicative gates that enable information to pass through the internal states of the memory cells. To train TRNN (which uses multiple LSTM units as shown in Fig. 2), the weights are updated corresponding to the gradient of an error function (the cross entropy is used because it can achieve a maximum likelihood in the disinformation/benign binary classification) in every training iteration. The "vanishing gradients problem" happens when the gradient's value is extremely small such that the weights either change too slowly or do not change at all

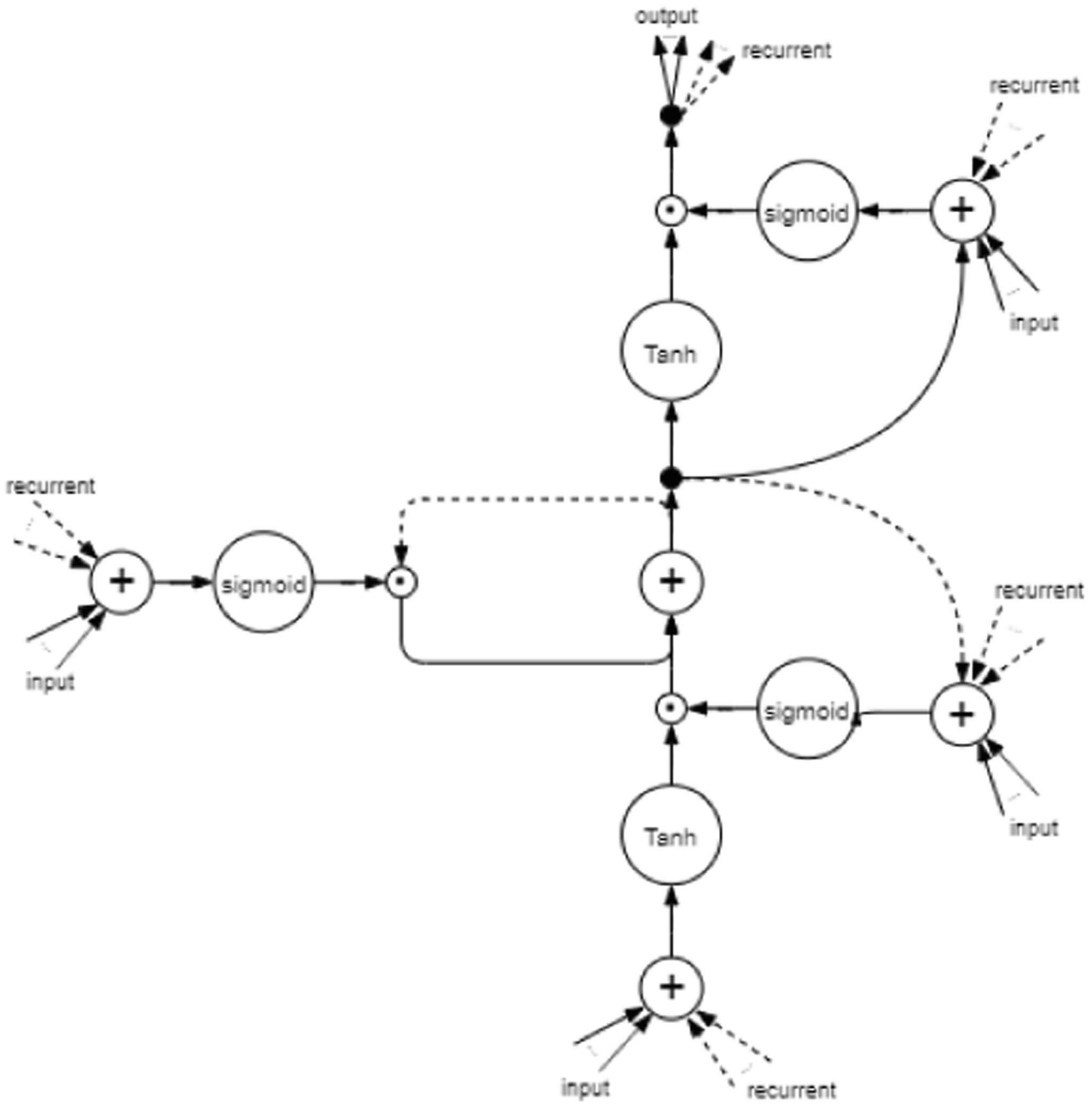

Fig. 2 A long short-term memory unit

in subsequent iterations of training, preventing the neural network from learning correctly from long-range textual features. The input gates, output gates, and forget gates are designed to keep selected values in states unmodified to achieve memorization and correct prediction in long sequences (Hochreiter & Schmidhuber, 1997).

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{4}$$

$$i_t = \sigma_g(W_i x_t + U_f h_{t-1} + b_i) \tag{5}$$

$$o_t = \sigma_g(W_o x_t + U_f h_{t-1} + b_o) \tag{6}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{7}$$

$$h_t = o_t \cdot \sigma_c(c_t) \tag{8}$$

$$\bar{x}_t = h_t \tag{9}$$

3.4 Module 3 - Recurrent Neural Network Using Temporal Features

Module 3 takes the output from Module 2 to produce a single time-series RNN that incorporates the temporal information of the scenarios. The formulas given in Eqs. 4, 5, 6, 7, 8, 9 are also used in the RNN of Module 3, with different input sizes than those of Module 2 (the i^{th} RNN takes as input textual feature vectors $\bar{x}_i$). Psychological theories have shown that people are more likely to share information when they are exposed to that information recently (recency) or when it is important to them (primacy) (Gino et al., 2009; Ngai et al., 2015). Therefore, TRNN models each input feature vector by including its time segment (morning or afternoon) to incorporate temporal information. A sigmoid function is used in the final layer to produce a probability to indicate the likelihood that a given scenario contains disinformation.

3.5 Novelty of TRNN

The novelty of the TRNN approach includes its theoretical foundation, comprehensiveness in modeling disinformation, and innovative data-centric augmentation of DL techniques. *First*, social and psychological theories were used to explain human behavior in spreading and detecting disinformation. Disinformation in financial social media reflects malicious behavior of illegal profit-seeking by exploiting people's motivation to gain from abnormal market movements and to follow the "crowd" in an uncertain social environment. As explained in Sections 3 and 4.2, TRNN uses Capital Asset Pricing Model (Sharpe, 1964) (as a data augmentation method) to characterize abnormal price movements of stock portfolios, Social Contagion Theory to model human behavior based on the information they receive (Wheeler, 1966; Le Bon, 1895), Emergent Norm Theory (Turner & Killian, 1957) to represent investors' behavior to follow the "crowd," and Social Exchange Theory (Emerson, 1976) to represent malicious actors' behavior of seeking illegal profit by posting disinformation social media messages. These theories are unique in TRNN and have not been used in prior DL techniques and in disinformation detection. *Second*, the TRNN approach is designed specifically to capture temporal and contextual information from messages that may contain disinformation, thus enriching knowledge representation of the complex features. To our knowledge, TRNN is the first approach that models human temporal perception of information by considering the importance of recent occurrences and past memory according to psychological theories (Gino et al., 2009; Ngai et al., 2015). The approach also advances IS research by designing and validating new information technology artifacts (Hevner et al., 2004) for detecting disinformation in financial social media. *Third*, the TRNN approach advances traditional DL methods (such as RNN and LSTM) by integrating multiple layers of RNN and LSTM cells and combining financial information (modeled by CAPM (Sharpe, 1964) and abnormal price movements), textual information (using PPMI), and interactions of market signals and timed trading patterns in the prediction of disinformation. The integration enables dynamic modeling of disinformation in financial social media that is beyond the predictive capabilities of traditional RNN techniques.

4 Experimental Design

To understand the usability and performance of the TRNN approach to disinformation detection, we conducted a series of experiments to compare the TRNN approach with different machine learning techniques. The U.S. financial market is chosen as the domain of the experiments because malicious hackers often spread disinformation to create abnormal price movements and to gain illegal profits. The study used a unique dataset that was built in this research to capture disinformation in social media. The following sections describe the data collection, augmentation, and experimental hypotheses.

4.1 Data Collection

The research developed an automated system to collect social media messages and stock prices of four Dow Jones Industrial Average (DJIA) technology companies: Apple, Microsoft, Intel, and Cisco (Chung & Sura, 2019). These companies were selected based on their important roles of providing technology products (e.g., smart phones), services (e.g., office productivity software), infrastructure (e.g., integrated circuits), and network hardware (e.g., routers) to the global economy. The system consists of multiple components to transform raw data into feature values (see Fig. 3): The *crawler* is a general-purpose collection agent that crawls publicly accessible web resources. The *scraper* extracts relevant data from HTML and JSON files and can be configured for different data content. The *featurizer* transforms raw text and financial data into input values for TRNN, using formulas as explained in Sections 3.2 and 4.2.1. The *scheduler* is the starting point of the system and regulates various day-to-day operations as mentioned above. The *learner* supports experimentation with different algorithms and approaches for machine learning. Using the collected data, a research test bed was built for use in the experimental evaluation.

The stock prices and social media messages were recorded once every five minutes on each U.S. trading day

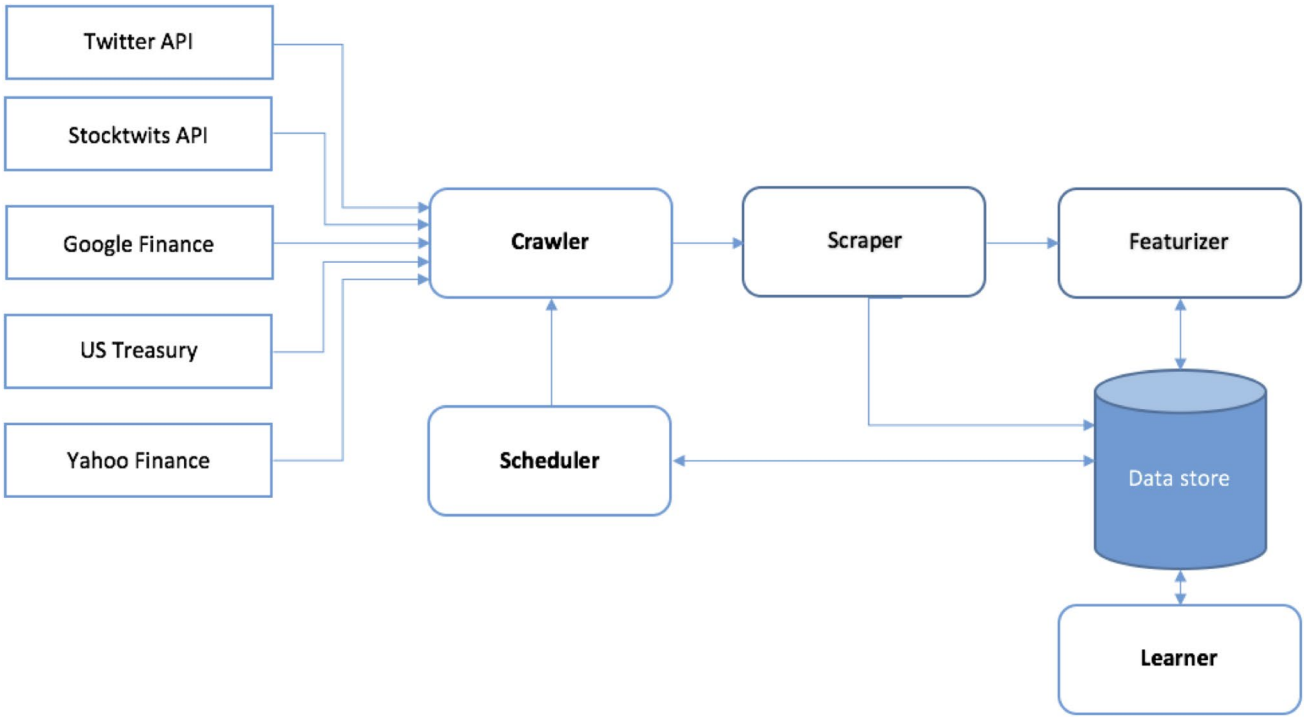

Fig. 3 An automated system to collect and transform data to support the TRNN approach

during 11-July-2017 – 15-August-2018 (277 trading days). The social media messages were collected from the sites Twitter and StockTwits (a social media platform for sharing ideas between investors, traders, and entrepreneurs) by using their public APIs. The total number of messages is 745,139, in which 560,062 messages are from Twitter and 185,077 messages are from StockTwits. The raw data includes (for each message) full text, timestamp, weekday, daytime (morning/afternoon), publication source, number of "likes," target respondent(s), sentiment (positive/negative), and author's total number of "likes." Selected sample messages collected from StockTwit are shown in Fig. 4.

We collected stock market data from three sources to enable computation of abnormal returns of stocks (to be explained in Section 4.2.1): (1) Google Finance provides real-time stock prices and S &P 500 index values that were used to indicate market performance. (2) The U.S. Treasury provides the latest risk-free rates of return for different maturities. (3) Yahoo Finance provides historical monthly closing stock prices. Due to its popularity among investors who also use financial social media, Google Finance and Yahoo Finance were chosen instead of other professional services (e.g., CRSP database). On each trading day, our system automatically collected data continuously (over 5-minute intervals) from the start (9:30 am) to close (4:00 pm) of the stock market (Jeong, 1999).

4.2 Data Augmentation

We performed data augmentation on our collected data to (1) transform the social media messages into contiguous message sets (called scenarios) and to derive labels from abnormal rates of return of the stock prices and to (2) produce human-validated labels of disinformation for each scenario.

4.2.1 Data Transformation

The raw data were transformed to scenarios and labels that indicate abnormal stock price movements. Each scenario is a concatenation of all social media messages posted during a 5-minute time frame. The total number of scenarios extracted from the raw data is 10,455 (37.74 scenarios per day). A label is assigned automatically to each scenario to indicate abnormal return (up, down, or none) of the market-weighed price of the selected companies' stocks.

The abnormal rate of return was calculated using Eq. 10 (according to Capital Asset Pricing Model (Sharpe, 1964), where R_i is Stock i's actual rate of return; R_f is the market return based on S &P 500 index normalized to the time span of the scenario; β is Stock i's price volatility relative to the overall market and is computed as the ratio of the covariance between the rate of return of Stock i and market rate of return (R_f) divided by the variance of market rate of

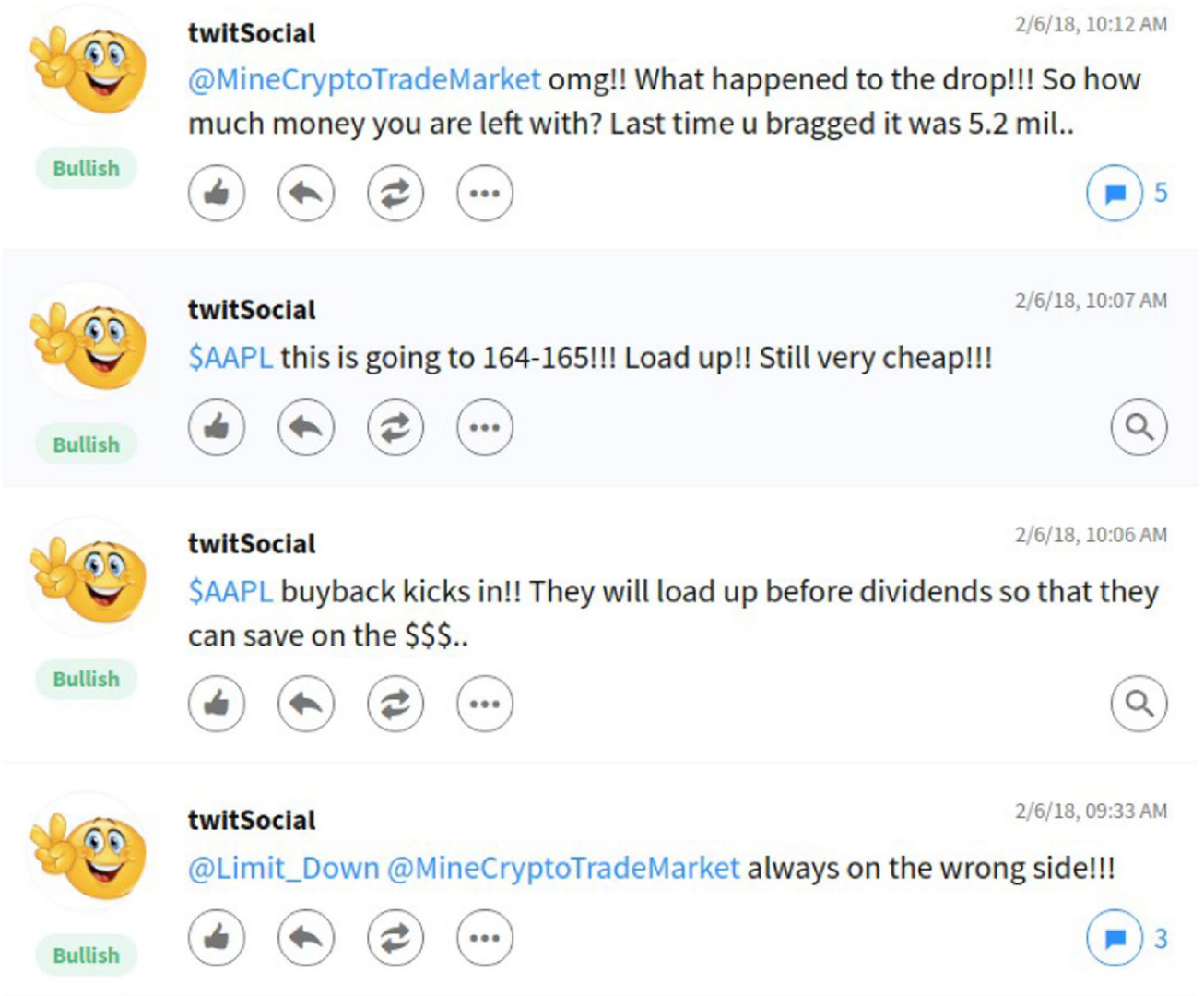

Fig. 4 A sample of social media discussions about stock price movements

return$\left(\beta = \frac{\text{cov}(R_i, R_m)}{\text{var}(R_m)}\right)$. In Eq. 10, the stock portfolio consists of the four selected stocks. The portfolio's rate of return (R_i) is computed as the weighted sum of the component stocks' rates of return (in which the "weight" of a stock is the ratio of the stock's market capitalization to the total market capitalization of all the portfolio's stocks). Consequently, a scenario is labeled as one of the following: "normal," "abnormal up," or "abnormal down."

$$\text{Abnormal Return of Stock Portfolio } i = R_i - [R_f + \beta * (R_m - R_f)] \quad (10)$$

Based on the calculation, 10,170 scenarios (97.27%) have normal price movement whereas 137 scenarios (1.31%) have abnormal upward movements (abnormal rate of return = 0.5% or above), and 148 scenarios (1.42%) have abnormal downward movements (abnormal rate of return = −0.5% or below). Only scenarios that are labeled "abnormal up" or "abnormal down" were used to study whether disinformation exists. On average, each abnormal up scenario has 62.58 messages and each abnormal down scenario has 70.42 messages (overall average = 66.25 messages per scenario).

While disinformation may possibly appear in "normal" scenarios and in "abnormal" scenarios, this research focuses only on disinformation found in scenarios with abnormal price movements because of two reasons. *First*, malicious hackers often leverage abnormal stock price movements to gain illicit profits. Therefore, focusing only on abnormal scenarios would help to create a useful filter of the input data (e.g., normal scenarios) that may be less likely to contain disinformation. *Second*, abnormal scenarios caused by

disinformation are often investigated by financial security regulators and by intelligence specialists to devise strategies to combat cyber attacks. The practical value of detecting disinformation in abnormal scenarios is thus far higher than in normal scenarios. *Third*, disinformation detection from abnormal scenarios (that are a minority among all scenarios) is not found in the literature. Related studies (such as this one) should provide new findings to support future research developments.

To develop the research test bed for use in disinformation detection, we manually built two message sets labeled as "benign" and "disinformation" respectively by drawing messages randomly from the "abnormal" scenarios as explained above. This manual process consists of two steps. *First*, we randomly drew messages from the "abnormal up" and "abnormal down" scenarios and extracted from each message these contextual and temporal feature values: timestamp, weekday, source, message's count of "likes," count of messages in the scenario, author's count of "likes," and message sentiment score (calculated by using the tool described in Hutto & Gilbert (2014). *Second*, we used the aforementioned feature values and the message content to assign an initial label to indicate whether the message is potentially disinformation or not (i.e., benign) (the initial label was later validated by independent human annotators as explained in Section 4.2.2). The two-step process resulted in a balanced dataset consisting of 2,000 messages categorized into four groups (each having 500 messages): (1) abnormal upward / benign, (2) abnormal upward / disinformation, (3) abnormal downward / benign, (4) abnormal downward / disinformation (see Table 3). An even distribution among the four groups ensures the same probability of selecting among the four types of messages in data validation.

4.2.2 Data Validation

The research test bed was validated by five human annotators who independently evaluated the labeling of disinformation in the messages. The validation required each annotator to indicate, for each message of the 50 randomly-sampled from the 2,000 messages (see Table 3), whether they agree on the initial labeling (i.e., benign or disinformation). Each sampled message was displayed to the annotator together with the contextual and temporal feature values as explained above and its initial label (half of the 50 messages were labeled initially as "disinformation" and the other half as "benign" using the two-step process explained above).

All annotators possess academic degrees from U.S. universities – four annotators have master's degrees and one has a bachelor's degree. Each annotator was given a survey with background introduction and a tutorial along with training examples to guide them to perform the task. The use of 5 annotators is aligned with research guidelines stating that generally 3-5 annotators are sufficient to validate the dataset labels to produce reliable results (Burmania et al., 2015). The research was certified by the Institutional Review Board of the investigators' university to comply with all regulations for protecting human subjects and data privacy.

Each annotator classified the same set of 50 messages that contain 25 messages with an initial label of "disinformation" and 25 with an initial label of "benign message." Out of the 50 messages, 48 messages (24 disinformation and 24 benign messages) were classified (by 3 or more annotators) as having the same label as their initial labels. Based on the results of the annotation, we also calculated three reliability measures of internal consistency among the five annotators: Cronbach's alpha (Cronbach, 1951), Cronbach's Alpha Based on Standardized Items, and Guttman's Lambda 6 (Guttman, 1945). The formulas are given in Eqs. 11, 12, 13. The notation and meaning are shown in Table 4. The results of these three measures are given in Table 5, showing that over 90% of the responses have at least four (out of 5) agreements among the independent human annotators.

Table 4 Notation and its Meaning in the Eqs. 11 – 13

Notation	Meaning
α	Cronbach's alpha $\in [0,1]$ that indicates inter-rater reliability
N	Total number of responses from the annotators
$\hat{c}$	Average of all covariances between pairs of responses
$\hat{v}$	Average variance of each response
e_j^2	Variance of the errors of estimate response j on the rest of the responses
s^2	Variance in each response that can be accounted for the linear regression of all of the other responses
$\hat{r}$	Mean of correlation coefficients
λ_6	Guttman's Lambda-6 estimate of reliability

Table 3 Categorization (and count) of Messages in Abnormal Stock Price Movements

Categorization	Abnormal upward movement	Abnormal downward movement
Benign Messages	Benign messages in abnormal upward scenarios (500 messages)	Benign messages in abnormal downward scenarios (500 messages)
Disinformation	Disinformation in abnormal upward scenarios (500 messages)	Disinformation in abnormal downward scenarios (500 messages)

Table 5 Reliability test results

Measure	Value
Cronbach's alpha, α	0.9094
Cronbach's standardized alpha, $\alpha_{standardized}$	0.9074
Guttman's Lambda-6, λ_6	0.9059

$$\alpha = \frac{N \times \hat{c}}{\hat{v} + (N-1)\hat{c}} \tag{11}$$

$$\alpha_{standardized} = \frac{N \times \hat{r}}{1 + (N-1)\hat{r}} \tag{12}$$

$$\lambda_6 = 1 - \frac{\sum_{j=1}^{N} e_j^2}{s^2} \tag{13}$$

In the sample messages shown in Fig. 4, the highlighted message (listed on the second line) was identified as disinformation, according to verification by independent human annotators described above. In that message, the author *twitSocial* speculated that the stock price of *Apple Inc.* would go up dramatically and suggested a timely purchase of the stock. The annotators based their independent decisions on the definition of disinformation (Lazer et al., 2018), which was provided to them in the beginning of the annotation process. Our validation shows that all three measures of reliability of the annotation exceed 90% (see Table 5).

4.3 Experimental Benchmarks and Hypothesis Testing

We selected four benchmark techniques to compare against the TRNN approach: artificial neural network (ANN) (Rumelhart et al. 1994), recurrent neural network (RNN), long short-term memory RNN (LSTM) (Yu et al., 2019), and convolutional RNN (CRNN) (Wang et al., 2019). ANN and RNN were chosen because they are among the most popular machine learning techniques for text classification. LSTM is among major deep-learning techniques that overcome the problems of large input gaps and long-term dependencies (Yu et al., 2019). CRNN uses convolutional and pooling layers to extract multiple sets of features that are then used as input to the LSTM neural network. CRNN has been shown to outperform other state-of-the-art text classification methods (fastText developed by Meta Platforms Inc. (formerly known as Facebook) (Joulin et al., 2016) and HAN (Yang et al., 2016) across different datasets (Wang et al., 2019).

The benchmark techniques use as their input the textual features extracted from our experimental dataset, whereas TRNN uses as input both textual and temporal features from the same dataset. ANN and RNN use two hidden layers, each having 128 nodes chosen empirically based on the size of the input and output vectors. Both LSTM and CRNN use a batch size of 64 for training and testing, and their architecture and hyperparameters are listed in Table 6.

We developed three hypotheses to evaluate the performance of TRNN against the four benchmark techniques. *First*, TRNN is hypothesized to outperform the benchmarks in detecting disinformation in accuracy due to its novel theory-based development that should enable a deeper understanding of human social and psychological features. *Second*, TRNN is hypothesized to outperform benchmark techniques in classifying disinformation messages in upward and downward scenarios due to its theory-based prediction of human behavior when faced with either type of scenarios. *Third*, the relatively better performance of TRNN is hypothesized to achieve a statistically-significant validity due to its theory-based representation and data-driven detection of disinformation. We used pairwise two-sample t-tests to test the hypotheses. We used these performance measures to evaluate the hypotheses: precision, recall, F-score and accuracy, whose formulas are given in Eqs. 14–17.

In our experiments, we randomly sampled from our research test bed 1,600 messages for training and 400 messages

Table 6 Hyperparameters Used in LSTM and CRNN

Technique	Hyperparameter	Description	Value
LSTM (Yu et al., 2019)	Dense Units	Fully-connected sequential layers, each having a specified number of computational nodes	[32, 32, 1]
	Activation function	Functions to transform input values to output values	['relu', 'relu', 'sigmoid']
CRNN (Wang et al., 2019)	CNN Channels	Number of channels used in CNN layers	[8, 16, 32]
	CNN Kernel	Same kernel size used in all CNN layers	[3, 3, 3]
	Pooling Layer	Input window size for max-pooling in CNN layers	[[2, 2], [2, 2], [1, 1]]
	LSTM Units	Number of units in bidirectional LSTM layers	[64, 64]
	LSTM Dropout rate	Fraction of the units to drop for linear transformation	0.5

for testing. Furthermore, we randomly sampled 320 messages from the training set for validation. The random sampling was conducted without replacement.

$$\text{Accuracy} = \frac{|\text{Messages Correctly Classified as Disinformation or Benign by an Algorithm}|}{|\text{All Benign and Disinformation Messages}|} \tag{14}$$

$$\text{Precision} = \frac{|\text{Messages Correctly Classified as Disinformation by an Algorithm}|}{|\text{All Messages Classified as Disinformation by the Algorithm}|} \tag{15}$$

$$\text{Recall} = \frac{|\text{Messages Correctly Classified as Disinformation by an Algorithm}|}{|\text{All Messages Classified as Disinformation by the Human Annotation}|} \tag{16}$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

5 Experimental Findings and Case Study

This section describes the experimental findings and provides a case to illustrate a potential real-world application of the approach to disinformation detection.

5.1 Accuracy of Disinformation Detection

As shown in Fig. 5, TRNN achieved the highest overall accuracy in disinformation detection. LSTM achieved the second-highest accuracy, followed by RNN, ANN, and CRNN in third, fourth, and fifth places respectively. Because TRNN specifically learns from temporal information of disinformation messages in addition to textual features, the TRNN model was able to accurately capture the dynamic signals unique to disinformation. The models of LSTM and RNN were also able to capture long-term dependencies in textual features that relate to temporal information, but their capabilities are lower than that of TRNN because the dependencies are not the most relevant temporal information for the detection. Similarly, ANN models the information by encoding the relationship in network weights, but do not possess as high a capability as TRNN does due to its lack of theoretical representation and of temporal modeling. Surprisingly, CRNN achieved the lowest accuracy – we believe its use of various convolutional operations may produce too much noise from the data, and its lack of temporal information also caused the lower accuracy.

5.2 Detecting Disinformation in Upward and Downward Scenarios

According to the literature, little is known about the dynamics of spreading disinformation in social media (Zubiaga et al., 2016; Kwon & Cha, 2014). This lack of understanding presents significant risk to the economy due to vulnerabilities of malicious use of financial social media. Therefore, we were interested in finding whether the use of TRNN enables better detection of disinformation in abnormal upward and downward price movement scenarios than benchmark techniques.

Table 7 shows the results comparing TRNN to the four benchmarks in upward and downward scenarios. The results show that TRNN consistently outperformed the benchmarks in terms of precision and F-score in both upward and downward scenarios; TRNN also outperformed all benchmarks in terms of recall in downward scenarios, and obtained the second-highest recall among all models in upward scenarios (in which CRNN obtained the highest recall). Several observations can be found from the results. *First*, the differences between the performance of TRNN and the performance of benchmarks are significantly larger than those between RNN and ANN. It is because the use of deep learning and temporal features (explicitly modeled in TRNN) supports a more timely and precise identification of disinformation, which is often spread by malicious hackers to create a "cognitive hack" that may work only for a short time frame (e.g., approximately 20 minutes in the hack reported in Lauricella et al. (2013). *Second*, the gaps between precision and recall in upward scenarios are generally larger than those in downward scenarios across all deep-learning models (TRNN, CRNN, LSTM). In addition, the scores of recall in upward scenarios are higher than those in downward scenarios across all models. This indicates that the effect of disinformation messages in upward scenarios is more readily identified by the models than in downward scenarios, due to the fact that malicious hackers tend to profit from abnormal gains. *Third*, in upward scenarios, TRNN has a relatively narrower gap between recall and precision than RNN, ANN, and CRNN have. This can be attributed to the combined effect of the two aforementioned reasons, i.e., temporal effect modeled by TRNN and TRNN's relatively stronger ability to identify "cognitive hack" in upward scenarios, in which hackers may use pump-and-dump schemes (Xu & Livshits, 2018) that have significant upward price movements signals (e.g., terms such as "BUY"). *Fourth*, in downward scenarios, TRNN is the only model having all the three measures above 80%, whereas both RNN and ANN have all these measures below 80%, and both LSTM and CRNN have precision and F-score below 80%. This is attributed to the effectiveness of using temporal features in TRNN to identify disinformation when prices go down dramatically (e.g., speculative short-selling).

5.3 Statistical Validity

This section reports results of statistical tests that compared between TRNN and each of the four benchmark techniques. We created 30 random samples with different proportions of training (80%) and validation (20%) in the dataset to evaluate the models' performance.

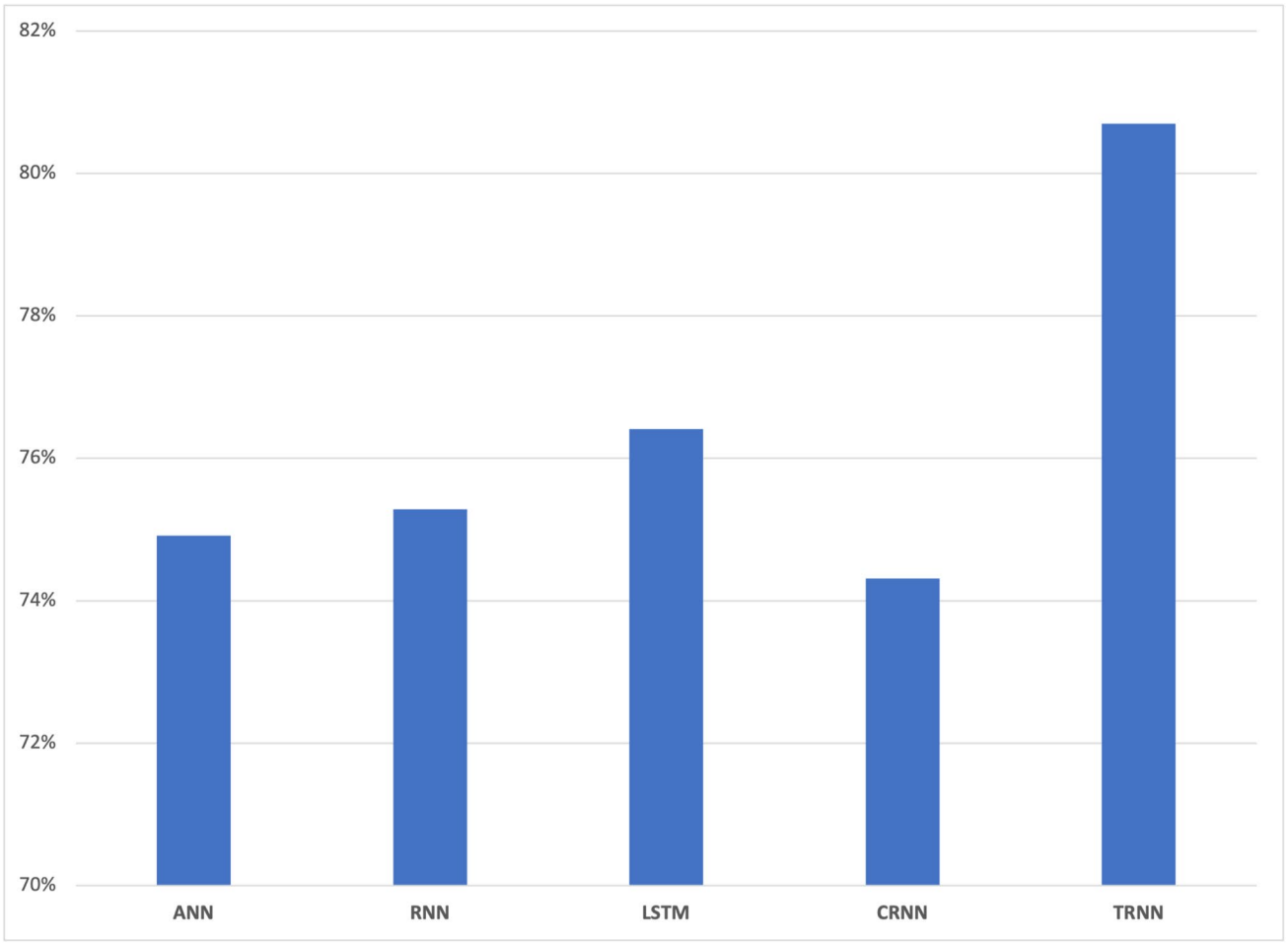

Fig. 5 Accuracies achieved by different techniques on disinformation detection

Table 7 Performance of disinformation detection in downward and upward scenarios

Scenario	Model	Precision	Recall	F-score
Downward	TRNN	**0.8036**	**0.8223**	**0.8113**
	CRNN	0.7290	0.8211	0.7717
	LSTM	0.7527	0.8013	0.7752
	RNN	0.7549	0.7426	0.7465
	ANN	0.7428	0.7578	0.7485
Upward	TRNN	**0.7789**	0.8481	**0.8108**
	CRNN	0.7178	**0.8573**	0.7808
	LSTM	0.7697	0.8287	0.7975
	RNN	0.7270	0.8024	0.7620
	ANN	0.7202	0.8116	0.7623

The bold numbers indicate the best performance achieved by a model among all models' performances in downward or upward scenarios

Note: the values are averaged from 30 random samples

Table 8 shows the hypotheses and their testing results. All p-values are significantly below 0.05, indicating that the F-scores of TRNN are significantly higher than those of the benchmark techniques. In addition, the Shapiro-Wilks tests for normality show that the F-scores achieved by all models do not differ significantly from a normal distribution, thus confirming the validity of the two-sample t-tests. Therefore, *we conclude that TRNN achieved a significantly higher performance in detecting disinformation from financial social media than all benchmark techniques did.* Among the results, the p-value of the test comparing F-scores of TRNN and of LSTM is the highest (still below 0.05 significantly), indicating that LSTM is a close contender to TRNN among all the benchmarks. The results show that the design of TRNN of being able to simultaneously process temporal and textual features can significantly increase the performance in disinformation detection than the benchmarks that use textual features only. Since temporal and textual features are widely available in a variety

Table 8 Pairwise Two-Sample t-Test of Models Using F-score

The bold numbers indicate the best performance achieved by a model among all models' performances in downward or upward scenarios

Hypothesis	textitp-value	Significant?	Conclusion
F_score(TRNN) > F_score(RNN)	6.0974e-08	Yes	Hypothesis confirmed
F_score(TRNN) > F_score(ANN)	2.2614e-08	Yes	Hypothesis confirmed
F_score(TRNN) > F_score(LSTM)	0.0001319	Yes	Hypothesis confirmed
F_score(TRNN) > F_score(CRNN)	8.5647e-07	Yes	Hypothesis confirmed

of social media, we believe TRNN will not only work well in the financial domain, but also in other areas such as political campaign or sport events.

5.4 Implication and Explanation

The results provide several implications for detecting disinformation with application to financial social media. *First*, the superior performance of TRNN across different measures indicates the importance of representing the social, economic, and psychological factors at play in the composition and spread of disinformation. The level of complexity in detecting disinformation can be adequately captured by TRNN, while other machine learning / DL models (RNN, ANN, LSTM and CRNN) may either overfit the data or contain too much bias. Therefore, TRNN is shown to be suitable for detecting disinformation in financial social media, thus demonstrating the promise of AI-driven secure knowledge management. The results provide new insights to examine social, economic, and psychological theories to detect malicious online behavior.

Second, TRNN supports rational explanation of disinformation detection by using financial data, market movements, and textual information, thus overcoming the interpretability problems due to a "black-box" nature commonly found in other DL techniques (Savage, 2022). A case of disinformation detection related to the stock of Apple Inc. is provided in Section 5.5. TRNN's use of social contagion theory, emergent norm theory, social exchange theory, and psychological theories on timeliness of human decisions provide theoretical guidance of its detection of disinformation from social media.

Third, compared with other approaches using ML/DL methods to detect disinformation (e.g., CRNN, ANN (Wang et al., 2019; Bahad et al., 2019), TRNN provides several advantages: (1) TRNN represents dynamic market signals by considering the temporal and contextual information in each financial social media scenario. (2) TRNN is developed based on social, economic, and psychological theories and can be used to explain its predictions from a theoretical perspective. By contrast, prior work in detecting disinformation does not examine these theories; other ML applications are also not grounded in these theories. (3) TRNN uses data-centric augmentation in its representation of complex features found in disinformation, thus advancing traditional DL techniques that focus primarily on model building and architectural sophistication. These advantages explain the generalizability of TRNN in domains other than finance. As the use of social media is prevalent across different domains, the highly encouraging results obtained from our experiments demonstrate a strong potential of TRNN to contribute to any domains involving textual social media, human decision making, and valuable assets.

5.5 A Case on Apple Inc.'s Stock Price Movement

To understand the potential application of TRNN to detecting disinformation in financial social media, we conducted an analysis of the tweets associated with abnormal price movements. We presented an empirical observation of how disinformation correlates with the stock price movement. Figure 6 shows a sample tweet identified from the dataset.

Similar to the tweet posted by the Scottish trader (Patrick, 2015), this aggressive tweet was posted at 2018-02-02 09:24:39 by the well known financial agency Phil's Stock World. This tweet claimed "a stock failure day" by stating that U.S. Federal Government and President Donald Trump failed to boost the markets. Before this tweet, Apple's stock price was steady at around $166 since 2018-01-29 for 4 consecutive days. Without any influence of the company's

Fig. 6 A tweet about abnormal stock price movement

earning release or major announcement, Apple's stock price still went down from $166.34 to $164.9 immediately (in less than 9 minutes after the tweet), and then to a closing price of $160.5 that day, resulting in a 3.64% drop. While other factors might have contributed to the change, the dramatic price drop following the tweet signaled the powerful and potentially malicious impact brought by the tweet.

6 Conclusion

Disinformation in social media poses significant threats to cybersecurity and efficient market operations (Cohen et al., 2021). However, the large volume of social media and rapidly changing environment make it challenging for AI techniques to accurately detect disinformation in social media. This research developed and validated a temporal recurrent neural network (TRNN) approach to addressing the needs. Grounded in social and psychological theories, TRNN incorporates contextual and temporal information from human-annotated social media data and from fine-grained financial market data that are synchronized with the social media data. Findings from our experiments on detecting disinformation in financial social media about four U.S. high-tech companies show that TRNN significantly outperformed the benchmark techniques in both accuracy and classification performance. A case study of social media messages and Apple Inc.'s stock price movement demonstrates a strong potential to apply TRNN to disinformation detection.

6.1 Research Contributions

This research makes several contributions. *First*, this research is the first attempt to develop a theory-based, deep learning (DL) approach that combines contextual, textual, financial, and temporal information in disinformation detection. Grounded in social and psychological theories, the TRNN approach and model have advanced the understanding of disinformation and of ways to detect disinformation from social media. As managers and decision makers face rapidly-growing challenges from online disinformation, the approach and model provide useful tools and techniques for secure knowledge management. *Second*, this research provides the first data-centric augmentation to existing DL methods for disinformation detection in financial social media. While existing DL methods focus primarily on depth and sophistication of neural network architecture, our findings enrich the understanding of human behavioral data used in training and applying DL methods. *Third*, the research contributes to new information systems (IS) artifacts and reusable datasets for disinformation detection research in financial social media. While prior research uses standardized datasets for testing DL models, our research breaks new ground by producing a unique disinformation dataset annotated by multiple independent human raters (with empirically-confirmed reliability) and novel IS artifacts in the forms of DL method and its instantiation for detecting disinformation in financial social media. These artifacts and dataset can benefit researchers, practitioners, and people interested in design science research and in related fields (Hevner et al., 2004; Peffers et al., 2007). *Fourth*, this research contributes to building a generalizable tool for classifying complex instances that involve textual content, temporal information, user activities, and financial assets. With suitable domain adaptation, the tool can be generalized to other applications, such as market prediction based on cryptocurrency movement, rumor detection in online forums, strategic planning in marketing campaigns, and intelligence filtering in adversarial settings (e.g., launching of competing products), among others.

6.2 Limitations and Future Directions

There are several limitations in this research. *First*, the use of only Twitter, StockTwits, and U.S. financial market data in our datasets limits the sources of social media and financial data used in the experiments. Using additional platforms and data sources (e.g., datasets other than DJIA component stocks) will provide new opportunities for deeper understanding of disinformation detection across different markets and online platforms. *Second*, the design of TRNN assumes daily collections and five-minute scenarios of social media messages to represent contextual information. This design may limit the study of non-linear message propagation in social media networks. These lengths can be dynamically adjusted to extend TRNN to incorporate more contextual information, such as message propagation patterns and network topology. *Third*, the parameters and structure of TRNN are selected based on our empirical testing, which was limited by our available resources. To address this limitation, extensive optimization and tesing can be used to improve model performance and computational efficiency.

Declarations

Conflicts of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Abrams, A. (2019). Here's what we know so far about russia's 2016 meddling. Time, https://time.com/5565991/russia-influence-2016-election/.

Ajao, O., Bhowmik, D., & Zargari, S. (2018). Fake news identification on twitter with hybrid CNN and RNN models. In *Proceedings*

of the 9th international conference on social media and society (pp.226–230).

Alzaidy, R., Caragea, C., & Giles, C. L. (2019). Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents. In *The World Wide Web conference*(pp.2551–2557). ACM

Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional lstm-recurrent neural network. In *2nd International conference on recent trends in advanced computing, ICRTAC 2019, November 11, 2019 - November 12, 2019* , vol.165 of *Procedia Computer Science* (pp.74–82). Elsevier B.V.

Barua, R., Maity, R., Minj, D., Barua, T., & Layek, A. K. (2019). F-nad: An application for fake news article detection using machine learning techniques. In *2019 IEEE Bombay section signature conference (IBSSC), 26-28 July 2019* (p. 6). IEEE

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy, 82*(6), 1063–1093.

Bond, S. (2021). Just 12 people are behind most vaccine hoaxes on social media, research shows. NPR News.

Burmania, A., Parthasarathy, S., & Busso, C. (2015). Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing, 7*(4), 374–388.

Chan, C. C. K., Kumar, V., Delaney, S., & Gochoo, M. (2020). Combating deepfakes: Multi-lstm and blockchain as proof of authenticity for digital media. In *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G), 21-25 Sept. 2020* (pp. 55–62). IEEE

Chung, W., & Sura, A. R. (2019). Asmods: Intelligent detection of abnormal stock price movements in response to social media postings. In *Recent developments in intelligent computing, communication and devices* (pp. 1169–1175). Springer

Chung, W. (2016). Social media analytics: Security and privacy issues. *Journal of Informaiton Privacy and Security, 12*(3), 105–106.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS ONE, 10*(6), e0128193.

Cohen, R. S., Beauchamp-Mustafaga, N., Cheravitch, J., Demus, A., Harold, S. W., Hornung, J. W., Jun, J., Schwille, M., Treyger, E., & Vest, N. (2021). Combating Foreign Disinformation on Social Media. RAND Corporation.

Commission, U. S. E. (2015). SEC charges: False tweets sent two stocks reeling in market manipulation. U.S. Security Exchange Commission:

Conroy, N. J., 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.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cybenko, A. K., & Cybenko, G. (2018). Ai and fake news. *IEEE Intelligent Systems, 33*(5), 3–7.

Dave, O. (2013). 50,000 lessons on how to read: a relation extraction corpus. Google AI Blog, https://ai.googleblog.com/2013/04/50000-lessons-on-how-to-read-relation.html.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., et al. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences, 113*(3), 554–559.

Delort, J.-Y., Arunasalam, B., & Paris, C. (2011). Automatic moderation of online discussion sites. *International Journal of Electronic Commerce, 15*(3), 9–30.

Ducci, F., Kraus, M., & Feuerriegel, S. (2020). Cascade-lstm: A tree-structured neural classifier for detecting misinformation cascades. In *Proceedings of the 26th ACM SIGKDD conference on knowledge discovery and data mining, 6-10 July 2020* (pp. 2666–76). ACM

Emerson, R. M. (1976). Social exchange theory. *Annual Review of Sociology, 2*(1), 335–362.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. Cambridge, MA: MIT Press.

Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 171–175). Association for Computational Linguistics

Gennatas, E. D., Friedman, J. H., Ungar, L. H., Pirracchio, R., Eaton, E., Reichmann, L. G., et al. (2020). Expert-augmented machine learning. *Proceedings of the National Academy of Sciences, 117*(9), 4571.

Giasemidis, G., Kaplis, N., Agrafiotis, I., & Nurse, J. (2018). A semi-supervised approach to message stance classification. *IEEE Transactions on Knowledge and Data Engineering*.

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and differentiation in unethical behavior: the effect of one bad apple on the barrel. *Psychol Sci, 20*(3), 393–8.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255–282.

Han, J., & Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks* (pp. 195–201). Springer

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Management Information Systems Quarterly, 28*(1), 75–105.

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

Hovland, C.I. (1957). *The order of presentation in persuasion*. Yale University Press, Inc.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Islam, S. R., Ghafoor, S. K., & Eberle, W. (2018). Mining illegal insider trading of stocks: A proactive approach. In *2018 IEEE international conference on big data (Big Data)* (pp. 1397–1406).

Jeong, J.-G. (1999). Cross-border transmission of stock price volatility: evidence from the overlapping trading hours. *Global Finance Journal, 10*(1), 53–70.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016.) Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 427–431). Association for Computational Linguistics

Jurafsky, D., & Martin, J.H. (2016). Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (draft 3rd edition)*.

Kandhway, K., & Kuri, J. (2017). Using node centrality and optimal control to maximize information diffusion in social networks. *IEEE Transactions on Systems Man Cybernetics-Systems, 47*(7), 1099–1110.

Knapp, R. H. (1944). A psychology of rumor. *Public Opinion Quarterly, 8*(1), 22–37.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text classification algorithms: A survey. *Information, 10,* 150.

Kuhlman, C. J., Tuli, G., Swarup, S., Marathe, M. V., & Ravi, S. (2013). Blocking simple and complex contagion by edge removal. In *2013 IEEE 13th international conference on data mining* (pp. 399–408). IEEE

Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., & Akbar, M. (2020). Fake news detection using deep learning models: a novel

approach. *Transactions on Emerging Telecommunications Technologies, 31*(2), e3767 (23 pp.).
Kumar, G., Singh, J. P., & Kumar, A. (2021). A deep multi-modal neural network for the identification of hate speech from social media. In *Responsible AI and analytics for an ethical and inclusive digitized society* (pp. 670–680). Springer International Publishing
Kwon, S., & Cha, M. (2014). Modeling bursty temporal pattern of rumors. In *Eighth international AAAI conference on weblogs and social media*.
Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining* (pp. 1103–1108). IEEE
Kwon, S., Cha, M., & Jung, K. (2017). Rumor detection over varying time windows. *PloS ONE, 12*(1), e0168344.
Langley, D., Reidy, C., Towey, M., Manisha, & Dennehy, D. (2021). Developing machine learning model for predicting social media induced fake news. In *Responsible AI and analytics for an ethical and inclusive digitized society* (pp. 656–669). Springer International Publishing
Latané, B. (1981). The psychology of social impact. *American Psychologist, 36*(4), 343–356.
Lauricella, T., Stewart, C. S., & Ovide, S. (2013). Twitter hoax sparks swift stock swoon. *The Wall Street Journal, 23*.
Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.
Le Bon, G. (1895). *The crowd: A study of the popular mind*. New York, NY: The MacMillan Co.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
Lee, P.S., Owda, M., & Crockett, K. (2018). The detection of fraud activities on the stock market through forward analysis methodology of financial discussion boards. In *Future of information and communication conference* (pp. 212–220). Springer
Li, Q. Z., Nourbakhsh, A., Shah, S., & Liu, X. M. (2017). Real-time novel event detection from social media. *IEEE 3rd International Conference on Data Engineering*, 1129–1139.
Liu, Q., Yu, F., Wu, S., & Wang, L. (2018). Mining significant microblogs for misinformation identification: an attention-based approach. *ACM Transactions on Intelligent Systems and Technology (TIST), 9*(5), 50.
Markowitz, D. M., & Hancock, J. T. (2016). Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology, 35*(4), 435–445.
McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27,* 415–444.
Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association* .
Miller, N., & Campbell, D. T. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurements. *The Journal of Abnormal and Social Psychology, 59*(1), 1–9.
Mitchell, T. (1997). *Machine Learning*. New York: McGraw-Hill.
Ngai, E. W. T., Tao, S. S. C., & Moon, K. K. L. (2015). Social media research: theories, constructs, and conceptual frameworks. *International Journal of Information Management, 35*(1), 33–44.
Nguyen, H. T., Cano, A., Tam, V., & Dinh, T. N. (2019). Blocking self-avoiding walks stops cyber-epidemics: A scalable gpu-based approach. *IEEE Transactions on Knowledge and Data Engineering*.
Owda, M., Lee, P. S., & Crockett, K. (2017). Financial discussion boards irregularities detection system (fdbs-ids) using information extraction. In *2017 Intelligent Systems Conference (IntelliSys)* (pp. 1078–1082). IEEE
Patrick, M. (2015). Sec charges scottish trader for twitter fraud. Financial Times, https://www.ft.com/content/03f828a0-8420-11e5-8e80-1574112844fd.
Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77.
Pham, C. V., Phu, Q. V., Hoang, H. X., Pei, J., & Thai, M. T. (2019). Minimum budget for misinformation blocking in online social networks. *Journal of Combinatorial Optimization, 38*(4), 1101–1127.
Qazvinian, V., Rosengren, E., Radev, D. R., & 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). Association for Computational Linguistics
Quan-Haase, A. (2016). *Technology and Society* (2nd ed.). Oxford, UK: Oxford University Press.
Reis, J. C., Correia, A., Murai, F., Veloso, A., Benevenuto, F., & Cambria, E. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems, 34*(2), 76–81.
Reis, J. C. S., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems, 34*(2), 76–81.
Ribeiro, F. N., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., & Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth international AAAI conference on web and social media*.
Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 797–806). ACM
Rumelhart, D. E., Widrow, B., & Lehr, M. A. (1994). The basic ideas in neural networks. *Communication of the ACM, 37*(3), 87–92.
Salton, G., & McGill, M. (1983). *An introduction to modern information retrieval*. NY: McGraw-Hill.
Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI. In *Proceedings of the ACM conference on human factors in computing systems*. ACM Press, p. Article 39.
Savage, N. (2022). Breaking into the black box of artificial intelligence. Nature. Savage, Neil eng News England 2022/03/31 Nature. 2022 Mar 29. pii: https://doi.org/10.1038/d41586-022-00858-1.
Sedikides, C., & Jackson, J. M. (1990). Social impact theory: A field test of source strength, source immediacy and number of targets. *Basic and Applied Social Psychology, 11*(3), 273–281.
Seth, T., & Chaudhary, V. (2020). A predictive analytics framework for insider trading events. In *2020 IEEE international conference on big data (Big Data)* (pp. 218–225).
Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance, 19*(3), 425–442.
Shi, B., & Weninger, T. (2016). Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems, 104,* 123–133.
Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22–36.
Singhania, S., Fernandez, N., & Rao, S. (2017). 3han: A deep neural network for fake news detection. In *International conference on neural information processing* (pp. 572–581). Springer
Sowa, J. F. (1987). *Semantic networks*. Citeseer.

Stage, C. (2013). The online crowd: A contradiction in terms? On the potentials of Gustave Le Bon's crowd psychology in an analysis of affective blogging. *Distinktion: Journal of Social Theory, 14*(2), 211–226.

Tong, G., Wu, W., Guo, L., Li, D., Liu, C., Liu, B., & Du, D.-Z. (2017). An efficient randomized algorithm for rumor blocking in online social networks. *IEEE Transactions on Network Science and Engineering*.

Turner, R. H., & Killian, L. M. (1957). *Collective behavior*. Prentice-Hall sociology series. Prentice-Hall Englewood Cliffs, N.J.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30) Curran Associates, Inc.

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: Short Papers, pp. 647–653).

Vosoughi, S., Mohsenvand, M., & Roy, D. (2017). Rumor gauge: predicting the veracity of rumors on twitter. *ACM Transactions on Knowledge Discovery from Data (TKDD), 11*(4), 50.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(638), 1146–1151.

Wang, R., Li, Z., Cao, J., Chen, T., & Wang, L. (2019). Convolutional recurrent neural networks for text classification. In *Proceedings of the international joint conference on neural networks*. IEEE Press.

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). ACM

Wheeler, L. (1966). Toward a theory of behavioral contagion. *Psychological Review, 73*(2), 179–192.

Xu, J., & Livshits, B. (2018). The anatomy of a cryptocurrency pump-and-dump scheme. arXiv:1811.10109 [q-fin.TR].

Yan, R., Li, Y., Wu, W., Li, D., & Wang, Y. (2019). Rumor blocking through online link deletion on social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD), 13*(2), 16.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the NA Chapter of the ACL: Human Language Tech.* (pp. 1480–1489). ACL

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation, 31*(7), 1235–1270.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649–657).

Zhang, H., Alim, M. A., Li, X., Thai, M. T., & Nguyen, H. T. (2016). Misinformation in online social networks: Detect them all with a limited budget. *ACM Transactions on Information Systems (TOIS), 34*(3), 18.

Zhang, Y., Zhang, Z., Miao, D., & Wang, J. (2019). Three-way enhanced convolutional neural networks for sentence-level sentiment classification. *Information Sciences, 477,* 55–64.

Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on World Wide Web* (pp. 1395–1405). International World Wide Web Conferences Steering Committee,

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS ONE, 11*(3), e0150989.

Wingyan Chung Dr. Wingyan Chung is an information systems scholar and educator with expertise in business analytics, social media analytics, machine learning, cybersecurity, data and text mining, knowledge management, and human-computer interaction. He is a professor at the Department of Computer Science in Soules College of Business at The University of Texas at Tyler.

Yinqiang Zhang is a Ph.D. student in the Department of Computer Science at The University of Hong Kong. His research interests include machine learning, visual tracking, medical image processing, LiDAR perception, reinforcement learning, and motion planning.

Jia Pan is an associate professor at the Department of Computer Science at The University of Hong Kong. His research interests cover robotics and artificial intelligence, with a focus on developing intelligent algorithms, sensors, and machines to accomplish fully autonomous robots.