This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Social Bots and the Spread of Disinformation in Social Media: The Challenges of Artificial Intelligence

Abstract

Artificial intelligence (AI) is creating a revolution in business and society at large as well as challenges for organisations. AI-powered social bots can sense, think, and act on social media platforms in ways similar to humans. The challenge is that social bots can perform many harmful actions, such as providing wrong information to people, escalating arguments, perpetrating scams, and exploiting the stock market. As such, an understanding of different kinds of social bots and their authors' intentions is vital from the management perspectives. Drawing from the actor-network theory (ANT), this study investigates human and non-human actors' role in social media, particularly Twitter. We use text mining and machine learning techniques, and after applying different pre-processing techniques, we applied the bag of words model to a dataset of 30,000 English-language tweets. The present research is among the few studies to use a theory-based focus to look, through experimental research, at the role of social bots and the spread of disinformation in social media. Firms can use our tool for the early detection of harmful social bots before they can spread misinformation on social media about their organisations.

Keywords: Artificial Intelligence; Social bots; Actor-network theory; Machine Learning; Deep Learning, Infodemic; Disinformation; Fake news

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

1. Introduction

Existing studies emphasise the value of social networks for firms and people (Leonardi, 2017; Candi et al., 2018; Mangold and Faulds, 2009; Sigfusson and Chetty, 2013; Williams et al., 2020). Product reviews through social media, for example, are becoming a rich source of consumer information (Moon and Kamakura, 2017). However, other research highlights social media's potentially harmful consequences for public discourse (Miranda et al., 2016). Information produced by individuals on social networking sites (SNSs), for example, has ethical consequences for business, such as spreading misinformation and compromising privacy, security, and trust (Nadeem et al., 2019; Wang et al., 2019). With new AI advancements, particularly the emergence of deep machine learning, we face new challenges; despite the opportunities AI presents for firms (Ross et al., 2019). Social bots within SNSs are autonomous actors driven by algorithms and software that post content on these platforms. Malicious bots are developed with the intention to harm. Research show these types of bots deceive and manipulate the stock market and also influence social media dialogues with fake news and misinformation (Kudugunta and Ferrara, 2018; Shi, Zhang, and Choo, 2019).

Twitter has about 23 million social bots, accounting for 8.5% of total users (Lima Salge and Berente, 2017). In addition, over two-thirds of tweets come from social bots. A 2018 Pew research study examined 1.2 million English-language tweets over a period of 47 days. The result showed that 66% of the tweets are through suspected bots (Pew Research Center, 2018). These days, Twitter has become a vector for spreading misinformation (Oh et al., 2013).

The above indicates that a variety of AI agents are classified by unique functions (Russell and Norvig, 2016). The social actor, who enters the problem situation in interaction with others (networked actors) and with events themselves, works to set goals and find procedures that can

achieve the desired endpoint (Cantor and Kihlstrom, 2000). AI appears in different applications: information recovery, text mining, expert systems, machine learning, computational intelligence, computer vision, optimisation, decision support, and automation actions via robots or intelligent agents (Russell and Norvig, 2016). The latter is the focus of this study. These are the positive side of AI robots. However, we need to look at the other side as well. Social bots play vital roles in society; they can also be altered to perform a wide range of malicious activities targeting business enterprises, government agencies, NGOs, political parties and SNSs. Information and data manipulation is an example of the dark side of social robots. For example, as Kudugunta and Ferrara (2018) mentioned during the 2010 U.S. midterm elections, malicious social bots employed to support some candidates with fake news. Also, the debate on information manipulation has intensified today with bots in social media and increased cyber-attacks associated with the 2016 US presidential election. The allegations that Russia' meddled in all big social media to harm the US election and increase political or social discord in the US highlight that networked actors (humans, robots or machines) play important roles in framing situations for political gain.

Another example is Russian-linked groups, reported by the European Commission that tried to undermine the credibility of the 2019 EU elections by disseminating false information through SNSs to influence votes (Satariano, 2019). Another study argues some SNSs provide incentives for content contribution (Tang et al., 2012), persuading people to create content for peers. Social bots are also available to purchase.

There are firms active on the Net that sell fake followers to generate false popularity of a tweet account. As noted by Yang et al. (2019), people can buy fake followers at a very low price. The majority of celebrities are among those who purchased fake followers on Twitter (Yang et al., 2019:49). Figure 1 below shows a snapshot of a tweet in XML format. Our dataset's detected bot

has tweeted 100 times, encouraging people to buy followers by offering them different sites to visit.

```
<document>
  - <![CDATA[
        If you trying to get more followers go to http://ohurl.com/clR .You will get 206 followers fast!
    ]]>
</document>
```

Figure 1: A hyperlink to buy followers

The examples mentioned above reveal the challenging side of social bots through the spread of misinformation via social media, highlighting the dark side of recent technological advancements. It is important to note that humans are also behind many cases of spreading infodemic or fake news.

To look at the challenging side of social bots, we use actor-network theory to support our argument describing how malicious social bots manipulate social media (Shao et al., 2017) and influence their audience. As such, it is essential to investigate how social bots impact people by triggering actors' actions in this context. We use ANT as a "toolbox to study meaning production, going from abstract structure - actants - to concrete ones – actors (Latour, 1996:373)." In this context, ANT allows us to understand better the relationship between actors, meaning production, or discourse, or text (Latour, 1996) generated by the socially intelligent actors through the machine and deep learning approaches. In particular, we are interested in investigating the role of agents (humans and bots) in spreading false information about various socio-political and economic events to influence public opinion on the Twitter platform.

Czarniawska (2006) noted that the notion of social had been extended from 'humans only' to 'all actants that can be associated' (p.65). The latter is the core concept of ANT, as articulated by Latour (2005). In this context, SNSs as a platform for social networking and knowledge sharing fits well within the framework of ANT. Couldry (2012) mentions, the new media whose history

has been so important to modernity's shared world is a platform for transforming what was formerly called 'mass' media into the interface between person-to-person, person-to-object, and object-to-object (e.g., bots retweet other bots). Moreover, this interaction within the context of an interweaved network of objects and humans makes ANT a useful tool to study social bots as entities aimed to generate content and "interact with humans to emulate and possibly alter their behaviour" (Ferrara et al., 2016:96). Social bots are able to explore information to fill their profiles (Ferrara et al., 2016:99). In this context, using the ANT theory, we examine the role of social bots through the following research questions:

RQ1: What is the role of social bots in shaping public opinion on social networking sites?

RQ2: What mechanism can be used to accurately determine if the author of a given Tweet is a bot or a human?

We developed the paper by reviewing the related studies. In section three, we discuss the research method, followed by section four, to analyse the data analysis using the machine learning approach to detect malicious social bots. Finally, section five concludes our study with a discussion as well as limitations and recommendations for future work.

2. Review of Related works

Bots or software robots appear in different realities and arrangements, including social media bots, chatbots, and conversational AI. Bots play important roles in human life today. Different AI applications have been developed for industrial manufacturing, healthcare, transportation, aviation, financial institutes, and governmental and public initiatives such as smart cities. In fact, the emergence of new technologies as enablers and multipliers is the backbone of today's AI applications in that the Internet has made it possible to connect massively parallel AI systems to

support these businesses. As mentioned earlier, malicious bots are explicitly designed with the purpose to harm.

2.1 Actor-Network Theory (ANT)

In recent years ANT has been applied in a wide range of scholarly research from Information Systems (Walsham, 1997; Mwenya and Brown, 2019) to Healthcare (Iyamu and Mgudlwa, 2018; Lutz and Tamò, 2016) business (Sarker et al., 2006; Effah, 2012; David and Halbert, 2014; Murdock and Varnes, 2018) and education (Fenwick and Edwards 2010) among others. Mwenya and Brown (2019) argue that a "Key tenet of ANT is generalised symmetry, which advocates that human and material actors be viewed on the same analytic plane" (p. 1). Their empirical study covering 36 recent scholarly publications in the IS domain shows that a wide range of IS studies has adopted ANT in several different ways. These studies highlight human and non-human actors' vital role in information management research. Iyamu and Mgudlwa (2018) reviewed big healthcare data from the lens of ANT. Using ANT, Lutz et al. (2016) investigate social and healthcare robots' role as a threat to patients' privacy within the healthcare systems. They emphasised that ANT is a descriptive, constructivist approach that considers the agency of objects, concepts, ideas, and the rationality of technology and society (p.2).

Latour (1996) mentioned that earlier studies of social networks, no matter how interesting, concern themselves with the social relations of individual human actors - their frequency, distribution, homogeneity, and proximity. But to do so, it does not limit itself to the individual human actor; it extends the word actor - or actant - to non-human, non-individual entities. It does not wish to add social networks to social theory but to rebuild social theory out of networks (p. 369).

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

As effective multipurpose communication platforms, SNSs such as Twitter, Facebook, Instagram, LinkedIn, Myspace, and others play important roles in our daily lives. Christakis and Fowler (2009) view SNS as "a kind of human superorganism. They grow and evolve. All sorts of things flow and move within them. This superorganism has its structure and function, and we became obsessed with understanding both. Seeing ourselves as part of a superorganism allows us to understand our actions, choices, and experiences in a new light" (p. xii). In this study, we view SNSs or social media in general as a collection of human, non-human objects as detailed by Actor-Network-Theory. This view supports us in understanding the role of social bots in spreading misinformation, rumours, and fake news within these sites and employing mechanisms to detect them promptly. This study debates that AI are integral parts of ANT. The integration will help us understand better human and non-human behaviour when they use technology-mediated social settings (Kling et al., 2003).

ANT also extends semiotics, which investigates how meaning is created and how meaning is communicated. Actor-network theory is a disparate family of material-semiotic tools, sensibilities, and methods of analysis that treat everything in the social and natural worlds as a continuously generated effect of the webs of relations within which they are located. Nevertheless, actor-networks do connect, and by connecting with one another, it provides an explanation of themselves, the only one there is for ANT (Latour, 1996). In this context, ANT extends the study of how signs and symbols (visual and linguistic) create meaning but linking semiotics to things instead of limiting them to meaning (Latour, 1996). Law (2008) argues that the ingredients of ANT include "semiotic relationality (it's a network whose elements define and shape one another), heterogeneity (there are different kinds of actors, human and otherwise), and materiality not just "the social" (p. 146). In this context, an actor is always a network of elements that the social and

technical elements are embedded in (Law, 2008). This means that it simply isn't possible to explore the social without at the same time studying the “hows” of relationships (p. 147); if all the world is relational, then so too are texts (Law, 2008). The Actor network's material semiotics explore the "hows" to articulate new intellectual tools, sensibilities, and questions (p.148).

Semiotics is an integral part of the current study. As noted by Mattozzi, 2019), an icon or likeness (i.e., a resemblance, as with a figurative image); an index (i.e., physical or causal relations); a symbol like a hashtag, an object or referent; a representation (i.e., the actual sign which represents the object), provides a common ground for the exchange between ANT and semiotics (Law 2009). In this context, semiotics as a "method" allows describing the "interdefinition of actors and the chains of translations" (Latour1988: 11). Or as precisely mentioned by Mol (2010:257) "In semiotics, words do not point directly to a referent, but form part of a network of words. They acquire their meaning relationally, through their similarities with and differences from other words. In ANT this semiotic understanding of relatedness has been shifted on from language to the rest of reality. Thus it is not simply the term, but the very phenomenon of its relations" (cf. Mattozzi, 2019:90). Finally, as mentioned by Mattozzi, the main task of the empirical level of ANT is the material study of objects at the language level (Mattozzi, 2019).

2.2 AI-powered Social Bots

Digital technologies such as AI are progressively becoming key to reaching a competitive advantage in business. However, simultaneously, firms are threatened with a range of challenges of these technologies in the market. One of these challenges is malicious social bots. These bots are a new form of bots that use social media to create content (Boshmaf et al., 2011; Lee, Eoff & Caverlee, 2011; Russell & Norvig, 2016; Woolley, 2016; Varol et al., 2017). However, Twitter is often used for research because it is open and easy to use (Ross et al., 2019). User-generated

content on social media influences organisations (Sheng et al., 2019), making social media an essential part of strategic management. As part of the realm of social intelligence (Russell & Norvig, 2016), social bots are important in this context. As Chu et al. (2012) and Ferrara et al. (2016) have mentioned, their proliferation has had both bad and good outcomes. In the case of the COVID-19 crisis, they can provide information useful for protecting society. Automated bots can also be useful for combining data from multiple sources for further analysis. However, malicious bots are sources of disinformation[1]. For example, spammers and rogue agents (e.g., hackers) can manipulate bots to appeal to current profiles (Ferrara et al., 2016; Ghosh et al., 2012; Hu et al., 2013).

Furthermore, hackers can use malicious bots to produce more severe effects, such as generating anxiety and panic in emergencies like COVID-19 (Chakraborty et al., 2020; Shi et al., 2020), harming a company's status, swaying political opinions (Chu et al., 2012; Wang, 2010), and/or spreading rumours and fake news. According to MIT researchers, robots have successfully faster the spread of both news true and false in social media: in fact, the latter spread more easily because robots are more likely to be behind them (MIT Media Lab, 2018). Research shows social bots play the spread of fake news online (Shao et al., 2017). Moreover, the fact that "relatively few accounts are responsible for a large share of the traffic that carries misinformation" (p.11). Therefore, this article uncovers how offensive malicious social bots pose a threat by evaluating them using detection techniques and suggesting possible future study paths.

2.3 Our ANT Approach

[1] According to the Merriam-Webster dictionary, disinformation is defined as "false information deliberately and often covertly spread (as by the planting of rumors) in order to influence public opinion or obscure the truth."

The four main components of ANT relevant to this study are actors, network interactions (e.g., social media networks), meanings, and action. The strength of ANT lies in its emphasis on humans and objects as participants in generating meanings (deliberation or manipulation) in the course of actions (Latour, 2005). ANT does not alter or transform meanings; it aims to reveal the meanings without any alteration. As Latour (2005) emphasised that the "black box" notion of ANT is acting as an intermediary role in which it aims to transport meaning without transformation (Latour, 2005). The volume of information exchange that various agents generate is an interesting phenomenon of consideration, as they may directly or indirectly be associated with the characteristics of agents (e.g., agents' risk aversion or risk-seeking behaviour, the agent's responsive behaviour, self-interest, and so on).

We aim to analyse tweeter posts as they are without altering, transforming, distorting, or modifying the meaning. In addition, there are two broad categories of meanings (content) disseminated in social media by actors, deliberation (information) and manipulation (disinformation), which in turn trigger actions. The latter is caused by bad actors (humans or malicious social bots) to distort communication discourse.

In the context of social media, the actors are integrated within the network platform. As such, ANT helps us separate AI-powered bots from the pool of existing actors active in social media such as Twitter. This allows us to investigate further the malicious activities of social bots in spreading misinformation.

Broadly speaking, fake news has been studied based on multiple theoretical perspectives, including the style of the content (Zuckerman, DePaulo and Rosenthal, 1981), distribution patterns, and attributes of the user in creating fake news (Zhou and Zafarani, 2018). Some of the approaches to tackle these issues have been fact-checking with experts, a machine learning

algorithm, information comparisons, etc. Unique emotional patterns and signals between fake news and real news also study as manipulation requires emotional language cues (Ghanem, Rosso and Rangel, 2020).

3. Research Method

Figure 2 outlines distinct phases of our approach to investigate our research questions.

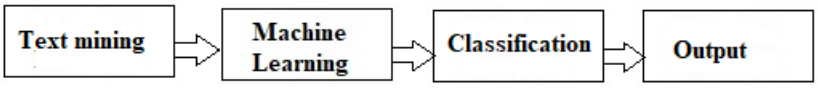


Figure 2. Social bot analytics processes

As mentioned by Adams (2017), the rapidly advancing machine learning areas and artificial intelligence contributed to powerful AI-enabled multi-modal social bots. The proposal developed below builds upon the multi-perspective framework for analysing agents' behaviour in a networked environment and particularly Twitter, where we aim to generate a new understanding of social bots' accelerated activities in spreading both real and false news through social media. The latter is of distinct interest to this study. The problem at hand strongly generalises this setting. As Figure 2 shows, our goal is to formulate a methodological and technically supported model that includes text mining and machine-learning-based classification of agents in Twitter and social bots' role in this context.

As Zeng et al. (2010) have mentioned, social media analytics develops tools and techniques for collecting, analysing, and visualising social media data, motivated by the target application's specific requirements. We extend this definition by emphasising the importance of sharing the results with other stakeholders and the community at large.

3.1 Data

The PAN-20 evaluation lab provided data for this empirical study to detect fake news (PAN, 2020). Each set of data was shipped with the related labelled meta-file as depicted in figures 3 and 4. Our

dataset was mixed with fake news generated by malicious social bots, humans, and true stories. In addition, we had two sets of data in English and Spanish languages; we selected the English-based tweets.

4. Data Analysis: Detecting AI-enabled Social bot on Twitter

For detecting social bots and the spread of fake news, we received a dataset of 30,000 tweets in the English language from the PAN-20 project (Dataset, 2020). These tweets were provided in the form of XML files. Each file represented one user and contained 100 tweets for that user. Separate files were provided that contained meta-data for the XML files.

The first challenge we faced was related to the XML format shipped files (see Figure 3) not suitable for natural language processing (NLP). As such, we processed data formatting to generate data frames[2] suitable for our analysis. We split data randomly into two subsets comprising 80% of data or 24,000 tweets for training. For testing, we use the remaining 20% (6,000 tweets) to validating the data set.

We performed three sets of algorithms in order to mine the tweet text and do the classification of content (RQ2) and finally, analyse the classified data. For this section, we apply different machine learning algorithms to determine the best accuracy rates in response to our RQ2, as described below.

Text Mining

We use text mining in this research. This methodology discovers patterns and relationships in the text (Sheng et al., 2019; Batistič & van der Laken, 2019; Thorpe et al., 2018). Mostafa (2013) argues that natural language processing (NLP) applications made text mining possible. NLP is performed on text collections composed of Tweets, known as a corpus. The corpus inherits many

[2] The pandas library was used for generating DataFrame from XML files.

hidden patterns in which further analysis is needed to confirm their relevance to decision-making factors. Text mining uses machine-learning techniques to gain knowledge from a large dataset that can then be used in decision making. Using SNSs data, text mining processes offer scholars understanding about the opinions in the text collection (Sharda et al., 2013)

Machine Learning

Machine learning is one popular methodology for studying user-generated content (Larsen et al., 2019; Lukyanenko et al., 2017; Ptaszynski et al., 2019). Research shows that this method can support companies to recognise applicable content generated by social media (Vermeer et al., 2019). In machine learning, information processors sort, assemble, simulate and classify information.

Classification

Classification is a process in which objects are divided into conceptually meaningful groups. The decision tree is among common classification techniques. Decision tree algorithms come from the family of supervised machine learning algorithms, where data is continuously split based on a parameter. There are generally two types of decision trees: classification trees and regression trees. For our problem, we used classification trees.

To solve our research problems outlined earlier in RQ2, we used various text pre-processing techniques and applied multiple machine learning as well as deep learning algorithms. We fed twitter data as inputs into our machine learning systems for the file reading and processing to achieve accuracy, as described below.

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

```
<author lang="en">
    <documents>
        <document>Tweet 1 textual contents</document>
        <document>Tweet 2 textual contents</document>
        ...
    </documents>
</author>
```

Figure 3: XML format.
Source: PAN, 2020

Figure 4 shows the meta-data file in the form of a text file

```
b2d5748083d6fdffec6c2d68d4d4442d:::0
2bed15d46872169dc7deaf8d2b43a56:::0
8234ac5cca1aed3f9029277b2cb851b:::1
5ccd228e21485568016b4ee82deb0d28:::0
60d068f9cafb656431e62a6542de2dc0:::1
...
```

Figure 4: Metafile in text format.
Source: PAN, 2020

As mentioned in the literature, fake news needs to preserve stylistic features in the tweet to record information that can add value in evaluation. Hence, we used the pre-processing of content (Baziotis et al., 2017) in order to tokenise the content into unique tags before passing it into our machine learning models. In this phase, we go through these steps:

- Removing stop words using natural language toolkit,
- Removing repeated words,
- Covering emojis to text (e.g., a Happy face emoji was converted to <HAPPY>)
- Removing punctuating marks,
- Removing HTML tags,
- Removing the numbers, kept only alphabets,
- Replacing user mentions @username with <USERMENTION>,
- Performing stemming by using snowball stemmer,
- Converting the content into a long string,
- Lowercasing the text.

This is an important feature because humans tend to have more diverse writing styles than bots, and bots tend to retweet more than humans. Despite the fact that both human and malicious social bats are behind the spread of fake news, however, there are differences between the two when it comes to the use of language and symbols. As mentioned by Varol et al. (2017) mentioned bots tend to retweet each other, meaning that simple bots frequently mention sophisticated bots and sophisticated bots retweet but do not mention humans due to their inabilities to involve in expressive discourse with individuals. Also, individuals may retweet bots, but those humans do it by posting interesting content. As observed by Gilani et al. (2017), malicious social bots tend to share URLs and upload media content more frequently than humans, and humans use more hashtags than bots. As such, indicators such as the average number of sentences, the average number of words, the average number of # symbols, @ symbol, frequency of unique words and are important indicators for distinguishing humans from malicious social bots (Gilani et al., 2017).

4.1 Results

After cleaning data for detecting fake news spread via malicious social bots, we deployed a text analytic technique called n-gram composed of an n-character share of a longer text (Cavnar and Trenkle, 1994). To obtain the highest accuracy rates (RQ2), we conducted six different machine learning and classification approaches, along with two deep learning methods according to the literature. The machine learning and classification approaches include Multinomial naïve Bayes (Su, Shirab, & Matwin, 2011), Logistic Regression (Dreiseitl & Ohno-Machado, 2002), Support Vector Machine (Tong & Koller 2001), Decision Tree (Kohavi & Quinlan, 2002), Multilayer Perceptron Neural Network (Pham et al., 2019), and K-Nearest Neighbor (Maillo et al., 2017). The deep learning approach includes Long Short-Term Memory (LSTM) (Liu, Mi, & Li, 2018) and

Bi-directional Long Short-Term Memory (Bi- LSTM) with the self-attention option (Zhou et al., 2016).

In standard machine learning, multinomial naïve Bayes returned the best result in its category, while Bi-LSTM (with self-attention) returned the best deep learning results. In the following sections, we explain the deep learning approaches.

Deep Learning Approaches

It is possible to generate output that mimics human behaviour or the so-called AI-powered (enabled) social bots with companion generative networks (Foysal et al., 2019). LSTM is an artificial Recurrent Neural Networks (RNNs) architecture classified as a deep feedforward neural network for classification of sound and signal categories (Le et al., 2019; Sak, Senior & Beaufays 2014), handwriting recognition (Graves et al., 2009), speech tagging (Xu & Yu, 2015) and key phrase extraction as part of natural language processing (Alzaidy, Caragea, & Giles, 2019). LSTM has a memory structure with self-connection links that store the network's temporal state at each time step. The bi-directional LSTM (Bi-LSTM) networks operate on the input sequence in forwarding and backward directions to decide the local context. The attention-based Bi-LSTM is designed to collect the most important semantic information in a sentence (Zhou et al., 2016).

Measuring Accuracy Rates

Accuracy is a measure of performance and is calculated as the percentage of correct predictions as the ratio of the correct predictions by the number of total predictions. As depicted in Figure 5, a single case problem has two classifications: positive and negative, in which the true positive (TP) and true negative (TN) are desired.

	Positive	Negative
True	True Positive	True Negative
False	False Positive	False Negative

Figure 5: A single classification model

However, a false positive (FP) happens when the outcome is incorrectly predicted as positive but it is negative. And a false negative (FN) occurs when the outcome is incorrectly predicted as negative, but it is positive (Schwenke & Schering, 2007; Lin & Chen, 2009; Yin et al., 2017). As a measure of performance, formula (1) calculates the accuracy rate as the proportion of correctly predicted (TP and TN) on the total values of all predictions.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (1)$$

4.2 Findings and Discussion

The above-mentioned developed algorithms were run on the training and test sets and achieved accuracies for the tasks as reported in Table1. As indicated, the machine learning methods using popular classifiers, including Logistic Regression and Multinomial naïve Bayes still perform well for the text mining studies described above. However, using a proper method such as Bi-LSTM with the deep machine learning with self-attention option offers the best possible results within a reasonable time frame. As mentioned by Xu and Yu (2015), within the context of bidirectional LSTM architecture, data in this system is using back-propagation through time (BPTT) which leads to the improved organisation, search, retrieval, and recommendation of various documents (Alzaidy, Caragea, & Giles, 2019).

Our attention-based Bi-LSTM model was trained with eight epochs; the hidden layers were set to size 50, dropout was set to 0.2, and "AdagradTrainer" was selected. The Python AdagradTrainer function is essentially an optimiser that performs stochastic gradient descent procedure for deep neural networks.

Table 1: Accuracy rates for different classification methods

Model	**Accuracy (Validation set)**
Multilayer perceptron NN	68
SVM	72.2
LSTM	74.2
KNN	74.7
Decision Tree	76.7
Logistic Regression	78.2
Multinominal NB	79.1
Bi-LSTM with Self-Attention	79.7

The empirical studies show that the deep learning Bi-LSTM model outperforms different algorithms designed for text classification and analysis, in particular when it comes to larger datasets (Jang et al., 2016; Minaee, Azimi, & Abdolrashidi, 2019; AlKhwiter & Al-Twairesh, 2020; Shaid, Zammer, & Muneeb, 2020).

As indicated by ANT, social bots and humans are integrated with a web of relationships. Each actor plays an important role in farming situations to impact public opinion and policy-making processes regarding social, economic or political matters. However, as mentioned earlier, many bots are created solely to spread misinformation, rumours, and spams. One of the key aspects used as an indicator is social context (Zhou & Zafarani, 2018) and the distribution pattern of real and fake news. We are using a text segmentation algorithm such as n-gram to effectively classify the text writing styles for the classification of social bots and humans (RQ2). Our findings answer our research question RQ2 in identifying fake news written by humans versus bots with an

accuracy rate of 79.7% due to the use of a more efficient predictive analysis algorithm. This forms one of our key findings that the algorithm could filter out almost 80% of malicious social-bot-generated tweets intended to harm the public trust and/or spread misinformation.

To highlight this, Figure 6 shows a snapshot of a malicious social bot that we detected in our dataset. It is important to note that the timing of our collected data set was linked to the 2020 US presidential election; as such, many of the tweet messages were related to this event. The following figure shows an example of these messages.

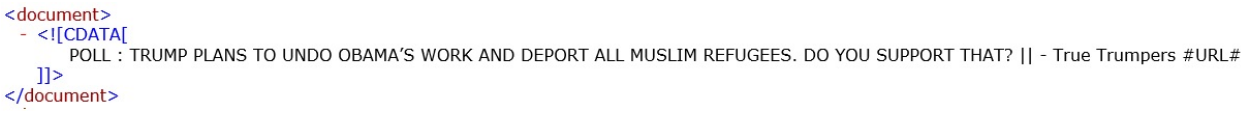
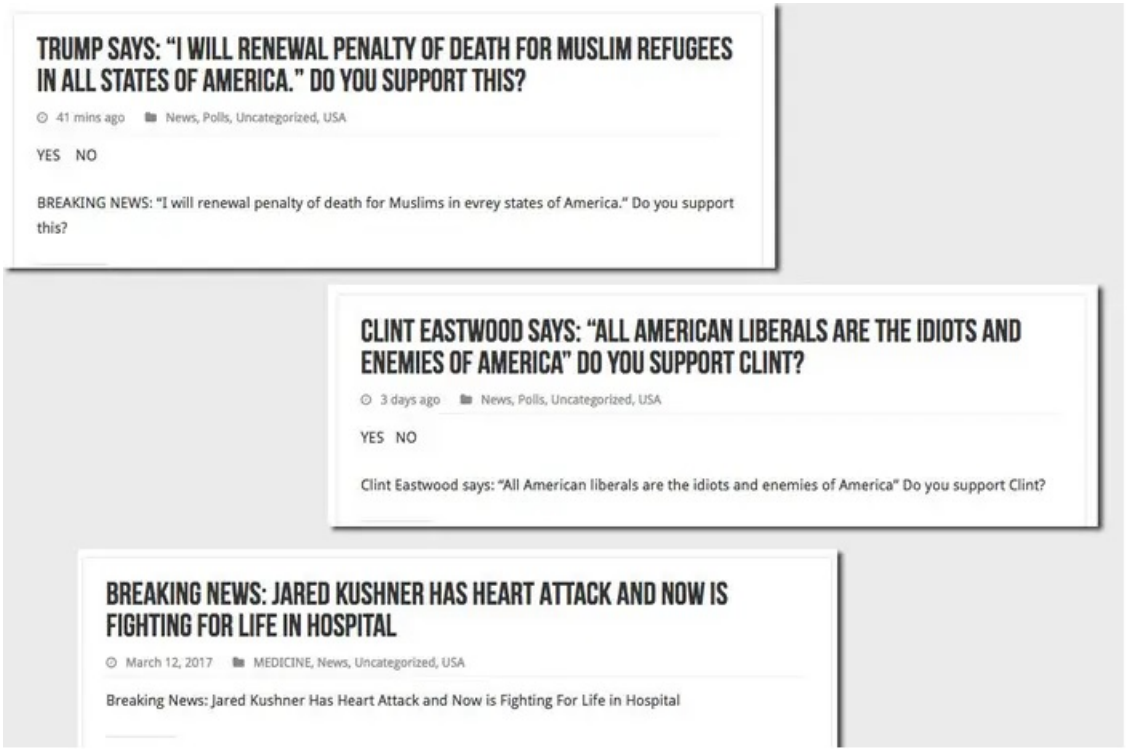


Figure 6: A Malicious social bot

We traced the *True Trumpers* in Tweeter. As shown in Figure 6, the account was created in August 2017, with no followers but 4,423 tweets. It does not have a human profile picture either.

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Figure 7: A typical social bot account



This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Figure 8: A sample of posts by True Trumpers
Source: Buzzfeed

Figure 7 shows completely false stories tweeted collections by *True Trumpers*. In fact, this malicious social bot operated from an Eastern European country, according to Buzzfeed (2017).

While much research emphasises the value of social media for people and organisations (Leonardi, 2017; Candi et al., 2018; Mangold and Faulds, 2009; Sigfusson and Chetty, 2013; Williams et al., 2020), other research highlights unavoidably malicious counterparts to the benefits of social media that bring about harmful consequences in community discourse (Miranda et al., 2016). Emerging technologies such as artificial intelligence empower bots on social media platforms, causing them to develop into social bots. The proliferation of bots has both bad and good outcomes. Our research provides an algorithm to differentiate tweets written by a human from those bots with an accuracy rate of 79.7% due to the use of a more efficient predictive analysis algorithm. Public and private organisations, as well as individuals, can apply this algorithm to filter out almost 80% of tweets generated by social bots aimed to harm public trust and/or spread disinformation otherwise harmful to the public opinion about a brand, political organisations, political actors, celebrities or commercial enterprises.

As mentioned by Latour (1996), we used ANT as a "toolbox" to investigate our research question RQ1 in order "to study meaning production… going from abstract structure - actants - to concrete ones – actors, in which an actant can be anything provided; it is granted to be the action source" (Latour 1996:373) (see Figure 1). We deployed ANT to critically investigate the fact that SNS and social media is not just "a kind of human superorganism" (Christakis &Fowler, 2009), and/or the networking aspects of SNS "emphasises relationship initiation, often between strangers" (Boyd & Ellision, 2008:211). In response to RQ2, we empirically investigated the role of social bots

in SNS. We showed that the bots exist, but they are also primary sources of disseminating fake news, rumours, and disinformation. We also empirically tested a range of popular algorithms using 30,000 tweets to determine the most effective algorithm that provides us with the mechanism to detect the social bots (RQ2) with a high accuracy rate.

Within the context of ANT, automated bots have more power in the network as they are active 24/7 a week for tweeting and retweeting subjects that they are programmed to do. They can be the source of disinformation. Research shows (Ross et al., 2019) that when social bots encourage a definite view on social media, they might create a situation that gives the false impression that the "bot opinion" is shared by more individuals than in reality. Accordingly, individuals who agree with that information get confidence to express the matter in interactions with other network peers, whereas individuals who disagree keep soundless of fear of being socially isolated (Ross et al., 2019). The bottom line is the fact that social bots are one of the emerging issues in management. They are widely used to blackmail a wide range of objects and actors.

4.3 Theoretical and Practical Implications

Theoretical Implications

Plenty of studies have been undertaken to investigate the use of social bots in different fields (e.g., Abokhodair et al., 2015; Forelle et al., 2015). This research examines to what extent social bots affect an individual's opinion formation. It also contributes to the growing research in artificial intelligence, such as social bots, particularly how social bots spread a given online opinion over social media platforms, leading to the misconception that other humans share the 'bot opinion.' It explains an admissible mechanism of opinion manipulation based on the theory of opinion formation. Numerous studies have investigated different aspects of social bots, described

different techniques to recognise social bots (e.g., Subrahmanian et al., 2016) or explained their behaviour (e.g., Hegelich and Janetzko, 2016).

This research integrates AI and Actor-Network Theory to understand better human and non-human behaviour in the use of social media as a platform for communication discourse. Recent studies related to social bots in social media were mainly focused on the technological and software engineering aspects of that (Wang, 2010; Jang et al., 2016; Zhou et al., 2016; Minaee, Azimi, & Abdolrashidi, 2019; Ross et al., 2019; Shaid, Zammer, & Muneeb, 2020). Our research is among the few to utilise a theory-based focus to look at the role of malicious social bots and the spread of disinformation in social media through experimental research. Therefore, this research, with an interdisciplinary approach, brows Actor-Network Theory from information systems to investigate the role of artificial intelligence in management.

As Walsham (1997) mentioned, ANT looks at the actors who are linked by associations of heterogeneous networks of aligned interests. A vital feature of ANT is that performers include humans and non-human actors such as technological objects (Walsham, 1997). They are part of hybrid networks such as SNSs. Besides, ANT is both a "theory and methodology combined" (Walsham, 1997: 469). ANT not only provides theoretical concepts to view the fundamentals of the world. It also suggests that it is precisely these fundamentals that need to be outlined in empirical work (Walsham, 1997), as this study had its main focus, and this is, "of course, no small task for a complex network" (Walsham, 1997:470) such as the social media networks.

Practical Implications

As mentioned above, malicious social bots have generated millions of social media pages containing false, unreliable and misleading information aimed to harm brands. Early detection of

these malicious bots, some of which can change their behaviour and characteristics, is an important task for trust-building, whether related to a business or particular brand and/or socio-political matters. From a managerial perspective, users, social actors and organisations can use our deep learning tool for early detection of harmful social bots before they can spread misinformation on social media. We obtained almost 80% accuracy in our deep learning classification approach (human or bot) using an attention-based Bi-LSTM. Therefore, this tool is a useful tool for managing their information and disinformation on social networking sites.

Besides, the results of our study highlight the fact that extra security measures such as a two-factor identification (e.g., a combination of password and anti-bot measures such as smart captcha or token) should be in place to limit or block unwanted bots from spreading disinformation in SNSs including Twitter.

5. Conclusion, Limitations and the Future Work

This study extends the ANT theory to the Deep Learning model of text mining from the perspectives of the semiotics paradigm. We have also contributed to ANT by demonstrating the interrelationship between actors (human, bots) and a web of words, symbols, signs, tokens, icons, etc., to actors on the ANT network. As noted by Mattozzi (2019), ANT offers a "methodological middle ground in between the theoretical-conceptual and the empirical ones" (p.97), as provided by this study. In addition, as Mattozzi highlighted, the main task of the empirical level of ANT is the material study of objects at the language level (Mattozzi, 2019). In this context, RQ1 and RQ2 integrate to understand better the content of tweets generated by actors and, in particular, the malicious bots in shaping public opinions on social networking sites.

Our experimental research is set up to examine the role of artificial intelligence and social bots in spreading disinformation. Social bots can harm a company's status by spreading rumours and

fake news about a particular brand. Social bots have successfully accelerated the spread of both true and false news in social media; in fact, the latter spread more because robots are more likely behind it. These days, thoughts expressed in social networking sites play a key role in influencing actors across all domains, including business. Therefore, this article reveals the probable threats of offensive social bots by evaluating them using detection techniques. Overall, the findings indicate the theoretical potential for using automated robot accounts to form an online opinion. The result shows that willingness to express an opinion and form an online opinion is affected by social bots on SNSs. It has been discussed that although bots are considered less credible than humans, they still have a significant impact on online public opinion. This research emphasises that the propagation of disinformation increases social bots' activity aimed at spreading unverified information. Drawing from Actor-Network Theory, this research examines human and non-human actors' role in SNSs, particularly Twitter, to understand better social bots' role in spreading spam and false information.

This research has limitations like other studies. The lack of meta-data is one of them. As depicted in Figure 4, the meta-data was masked out and replaced with a code to anonymise the tweet author's actual name. The meta-data contains valuable information about the author's information like the authors' name, profile picture, ID number, physical location, followers, and the time stamp of posts or retweets, among others. If the meta-data and the content feed into our deep learning machine, the accuracy will be increased drastically due to more information provided to the system. Particularly meta-data provides valuable information for the purpose of authors' classification. As mentioned by Stieglitz et al. (2017a,b) many bots lack basic account information like name or profile pictures. While regular users get access from front-end websites, bots obtain access through a site's application programming interface (API). As the API of Twitter

is especially accessible, many social bots focus on this platform (Stieglitz et al., 2017a) because it offers a wide range of resources and a faster communication platform (Crains and Shetty, 2020).

We used only English-language tweets, and this is another limitation of our research. As mentioned in section 3.1, we had two sets of Twitter data, in English and Spanish; we had the option to select one of the possible languages or both. We selected only the English language. Our focus on this study was text-based communication, while our model is also designed to perform well with non-text communication such as voice (speech), image, and video classification, as the deep learning Bi-LSTM method is chartered to do so. In the future, new research may consider embedding TF-IDF weighting (Rangel and Rosso, 2019) and studying document embedding (Daelemans et al., 2019). Using syntactic n-grams is another option for future research direction (Potthast et al., 2019).

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Reference

Abokhodair, N., Yoo, D., and D. W. Mcdonald, (2015). 'Dissecting a social botnet', In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing – CSCW '15*, Vancouver, BC, Canada (pp. 839–851). New York, NY: ACM.

Abou-Assaleh, T., Cercone, N., Keselj, V., and R. Sweidan (2004, September). 'N-gram-based detection of new malicious code', In *Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004. COMPSAC 2004*. (Vol. 2, pp. 41-42). IEEE.

Adams (2017). 'AI-powered social bots', arXiv preprint arXiv:1706.05143.

Akhtar, P., Khan, Z., Frynas, J. G., Tse, Y. K., and R. Rao-Nicholson (2018). 'Essential micro-foundations for contemporary business operations: top management tangible competencies, relationship-based business networks and environmental sustainability', *British Journal of Management*, **29**(1), pp. 43-62.

AlKhwiter, W., and N. Al-Twairesh (2020). 'Part-of-speech tagging for Arabic tweets using CRF and Bi-LSTM', *Computer Speech & Language*, **65**, 101138.

Alzaidy, R., Caragea, C., and C. L. Giles (2019, May). 'Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents In The world wide web conference (pp. 2551-2557).

Ashfaq, M., Yun, J., Yu, S., and S. M. C. Loureiro, (2020). 'I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents' *Telematics & Informatics*, **54**, 101473.

Bakir, V., and A. McStay (2018). 'Fake news and the economy of emotions: Problems, causes, solutions', *Digital Journalism*, **6**(2), pp. 154–175.

Batistič, S. and P. van der Laken .(2019). 'History, evolution and future of big data and analytics: a bibliometric analysis of its relationship to performance in organisations', *British Journal of Management*, **30,** pp**.** 229-251.

Baziotis, C., Pelekis, N., and C. Doulkeridis (2017, August). 'Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis', In *Proceedings of the 11th international workshop on semantic evaluation* (SemEval-2017) (pp. 747-754).

Bibault, J. E., Chaix, B., Nectoux, P., Pienkowski, A., Guillemasé, A., and B. Brouard (2019). 'Healthcare ex Machina: Are conversational agents ready for prime time in oncology?', *Clinical and translational radiation oncology*, **16**, pp. 55-59.

Boshmaf, Y., Muslukhov, I., Beznosov, K. and M. Ripeanu (2011). 'The socialbot network: when bots socialise for fame and money', In *Proceedings of the 27th annual computer security applications conference* (pp. 93-102).

Boyd, D. M., and N. B. Ellison (2007). 'Social network sites: Definition, history, and scholarship', *Journal of computer-mediated Communication*, **13**(1), pp. 210-230.

Buzzfeed (April 2017). This Pro-Trump Website Run From Eastern Europe May Be The Worst Thing On The Internet. https://www.buzzfeednews.com/article/craigsilverman/anti-muslim-traffic-arbitrage-is-a-thing

Candi, M., D. L. Roberts, T. Marion and G. Barczak (2018). 'Social strategy to gain knowledge for innovation', *British Journal of Management*, **29,** pp. 731-749.

Cantor, N., and J. F. Kihlstrom (2000). 'Social intelligence', *Handbook of intelligence*, **2**, 359-379.

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., and A. E. Hassanien (2020). 'Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media', *Applied Soft Computing*, **97**, 106754.

Christakis, N. A., and J. H. Fowler (2009). '*Connected: The surprising power of our social networks and how they shape our lives*', Little, Brown Spark.

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Chu, Z., Gianvecchio, S., Wang, H., and S. Jajodia (2012). 'Detecting automation of twitter accounts: Are you a human, bot, or cyborg?', *IEEE Tran Dependable and Secure Comput*, **9**(6), pp. 811-824.

Chung, T. S., R. T. Rust, and M. Wedel (2009). 'My mobile music: An adaptive personalisation system for digital audio players', *Marketing Science*, **28**, **pp.** 52-68.

Chung, T. S., M. Wedel and R. T. Rust (2016). 'Adaptive personalisation using social networks', *Journal of the Academy of Marketing Science,* **44**, **pp.** 66-87.

Crains, I, and P. Shetty (August, 2020). 'Introducing a new and improved Twitter API', https://blog.twitter.com/developer/en_us/topics/tools/2020/introducing_new_twitter_api.html

Cavnar, W. B., and J. M. Trenkle (1994, April). 'N-gram-based text categorisation', In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (Vol. 161175).

Choudhury, S. R. (May, 2016). 'SoftBank's Pepper Robot Gets a Job Waiting Tables at Pizza Hut', *CNBC.* https://www.cnbc.com/2016/05/24/mastercard-teamed-up-with-pizza-hut-restaurants-asia-to-bring-robots-into-the-pizza-industry.html

CNBC News (April, 2015). 'Robot with $100 bitcoin buys drugs, gets arrested', https://www.cnbc.com/2015/04/21/robot-with-100-bitcoin-buys-drugs-gets-arrested.htmland

Couldry, N. (2012). 'Media, Society, World: Social Theory and Digital Media Practice', *Polity Press*, Cambridge, UK.

Czarniawska, B. (2006). 'Bruno Latour: Reassembling the Social: An Introduction to Actor-Network Theory', *Organization Studies*, **27**(10), pp. 1553-1557.

Daelemans, W., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., ... and E. Zangerle (2019, September). 'Overview of PAN 2019: bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection', In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 402-416). Springer, Cham.

Davenport, T., and R. Kalakota (2019). 'The potential for artificial intelligence in healthcare', *Future healthcare journal*, **6**(2), 94.

Dataset (2019). 'Authorship Analysis', Available at: https://pan.webis.de/data.html#pan19-authorship-attribution.

Dautenhahn, K. (1995). 'Getting to know each other-Artificial social intelligence for autonomous robots', *Robotics and Autonomous Systems*, **16** (2–4), pp. 333-356.

De Lima Salge, C. A., and N. Berente (2017). 'Is that social bot behaving unethically?', *Communications of the ACM*, **60**(9), pp. 29-31.

Dreiseitl, S., and L. Ohno-Machado (2002).' Logistic regression and artificial neural network classification models: a methodology review', *Journal of biomedical informatics*, **35**(5-6), pp. 352-359.

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

Eisenhardt, K. (1989). 'Agency Theory: An Assessment and Review', *The Academy of Management Review*, **14**(1), pp. 57-74.

Ekström, M., Lewis, S. C., and O. Westlund (2020). 'Epistemologies of digital journalism and the study of misinformation', *New Media & Society***, 22**(2), pp. 205–212.

Fadhil, A. (2018). 'Beyond patient monitoring: Conversational agents role in telemedicine & healthcare support for home-living elderly individuals', arXiv preprint arXiv:1803.06000.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and A. Flammini (2016). 'The rise of social bots', *Communications of the ACM*, ***59***(7), pp. 96-104.

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Forelle, M. C., Howard, P. N., Monroy-Hernandez, A., and S. Savage (2015). 'Political bots and the manipulation of public opinion in venezuela', SSRN *Electronic Journal*. **14**, pp. 57-74.

Foysal, A., Islam, S., and T. Rahaman (2019). 'Classification of AI Powered Social Bots on Twitter by Sentiment Analysis and Data Mining through SVM', *International Journal of Computer Applications*, **117**(25), pp. 13-19.

Gans, G., Jarke, M., Kethers, S., and G. Lakemeyer (2001, May). 'Modeling the impact of trust and distrust in agent networks', In *Proc. of AOIS'01* (pp. 45-58).

Gentsch, P. (2019). 'Conversational AI: how (chat) bots will reshape the digital experience', In *AI in marketing, sales and service* (pp. 81-125). Palgrave Macmillan, Cham.

Gearhart, S., and W. Zhang (2014). 'Gay bullying and online opinion expression: Testing spiral of silence in the social media environment', *Social science computer review*, **32**(1), 18-36.

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., and J. Crowcroft (2017, July). 'Of bots and humans (on twitter)', In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 349-354).

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto,F., Ganguly, N., and K. P. Gummadi, (2012). 'Understanding and combating link farming in the twitter social networ', In *Proceedings of the 21st international conference on World Wide Web*, WWW '12.

Ghanem, B., Rosso, P., and F. Rangel (2020). 'An emotional analysis of false information in social media and news articles', *ACM Transactions on Internet Technology (TOIT)*, **20**(2), pp. 1-18.

Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H. and J. Schmidhuber (2009). 'A Novel Connectionist System for Improved Unconstrained Handwriting Recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(5), pp. 855–868.

Guzman, A. L., and S. C. Lewis (2020). 'Artificial intelligence and communication: A Human–Machine Communication research agenda', *New Media and Society*, **22**(1), pp. 70-86.

Hegelich, S., and D. Janetzko (2016). 'Are social bots on Twitter political actors? Empirical evidence from a Ukrainian social botnet', In *Proceedings of the Tenth International Conference on Weblogs and Social Media* (ICWSM-2016), Cologne, Germany (pp. 579–582). Palo Alto, CA: The AAAI Press.

Holtgraves, T. M., Ross, S. J., Weywadt, C. R., and T. L. Han (2007). 'Perceiving artificial social agents', *Computers in human behavior*, **23**(5), pp. 2163-2174.

Huang, M.-H., R. Rust and V. Maksimovic (2019). 'The Feeling Economy: Managing in the Next Generation of Artificial Intelligence (AI)', *California Management Review*, **61,** pp. 43-65.

Huang, M.-H. and R. T. Rust (2018). 'Artificial intelligence in service', *Journal of Service Research,* **21,** pp. 155-172.

Huang, M.-H. and R. T. Rust (2020). 'Engaged to a Robot? The Role of AI in Service', *Journal of Service Research***,** p. 1094670520902266.

Huang, Z., Xu, W., and K. Yu (2015). 'Bidirectional LSTM-CRF models for sequence tagging', arXiv preprint arXiv:1508.01991.

Hu, X., Tang, J., Zhang, Y., H. Liu (2013). 'Social spammer detection in microblogging', In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*.

Jang, B., Kim, M., Harerimana, G., Kang, S. U., and J. W. Kim (2020). 'Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism', *Applied Sciences*, **10**(17), 5841.

Khatri, C., Venkatesh, A., Hedayatnia, B., Gabriel, R., Ram, A., and R. Prasad (2018). 'Alexa Prize—State of the Art in Conversational AI', *AI Magazine*, **39**(3), pp. 40-55.

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Kešelj, V., Peng, F., Cercone, N., and C. Thomas (2003, August). N-gram-based author profiles for authorship attribution. In Proceedings of the conference pacific association for computational linguistics, *PACLING* (Vol. 3, pp. 255-264). sn.

Kohavi, R., and J. R. Quinlan (2002). 'Data mining tasks and methods: Classification: decision-tree discovery', In Handbook of data mining and knowledge discovery (pp. 267-276).

Kudugunta, S. and Ferrara, E., 2018. Deep neural networks for bot detection. *Information Sciences, 467*, pp.312-322.

Larsen, K. R., D. Hovorka, A. Dennis and J. D. West (2019). 'Understanding the Elephant: The Discourse Approach to Boundary Identification and Corpus Construction for Theory Review Articles', *Journal of the Association for Information Systems*, **20,** 15.

Latour, B. (1996). 'On actor-network theory: A few clarifications', *Soziale welt*, pp. 369-381.

Latour, B. (1998). 'A Relativistic Account of Einstein's Relativity', *Social Studies of Science*, **18**(1), pp. 3-44.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford, UK: Oxford University Press.

Law, J. (2008). 'Actor Network Theory and Material Semiotics', In B.S. Turner (ed.), *The New Blackwell Companion to Social Theory*, pp. 141–158. Oxford: Wiley-Blackwell.

Leonardi, P. M. (2017). 'The social media revolution: Sharing and learning in the age of leaky knowledge', *Information and Organization*, **27,** pp. 47-59.

Lee, K., Eoff, B. D., & Caverlee, J. (2011, July). 'Seven months with the devils: A long-term study of content polluters on twitter', In *Fifth international AAAI conference on weblogs and social media*.

Lin, S. W., and S. C. Chen (2009). 'PSOLDA: A particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis', *Applied Soft Computing*, **9**(3), pp. 1008-1015.

Liu, H., Mi, X., and Y. Li (2018). 'Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM', *Energy Conversion and Management*, **159**, pp. 54-64.

Lomas, N. (2017). 'Lyrebird is a voice mimic for the fake news era', *Obtenido de Techcrunch:* https://techcrunch. com/2017/04/25/lyrebird-is-a-voice-mimic-for-the-fake-news-era

Lukyanenko, R., J. Parsons, Y. Wiersma, G. Wachinger, B. Huber and R. Meldt (2017). 'Representing crowd knowledge: Guidelines for conceptual modeling of user-generated content', *Journal of the Association for Information Systems*, **18**, p. 2.

Mattozzi, A. (2019). What can ANT still learn from semiotics. *The Routledge Companion to Actor-Network Theory*. pp. 87-100. *Abingdon, Oxon*.

Mol, A. (2010). 'Actor-Network Theory: Sensitive Terms and Enduring Tensions', *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, **50**, pp. 253–269.

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., ... & A. Pettigrue (2018). 'Conversational ai: The science behind the alexa prize' *arXiv preprint arXiv:1801.03604*.

Russell, S and P. Norvig (2016). *Artificial Intelligence: A Modern Approach*, 3rd Ed., *Pearson Education Limited, Essex*, England.

Rust, R. T. and M.-H. Huang (2014). 'The service revolution and the transformation of marketing science', *Marketing Science*, **33,** pp. 206-221.

Rust, R. T. and M.-H. Huang (2012). 'Optimising service productivity', *Journal of Marketing*, **76,** pp. 47-66.

Le, T., Vo, M. T., Vo, B., Hwang, E., Rho, S., and S. W. Baik (2019). 'Improving electric energy consumption prediction using CNN and Bi-LSTM', *Applied Sciences*, **9**(20), 4237.

Maillo, J., Ramírez, S., Triguero, I., and F. Herrera (2017). 'kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data', *Knowledge-Based Systems*, **117**, pp. 3-15.

Mangold, W. G. and D. J. Faulds (2009). 'Social media: The new hybrid element of the promotion mix', *Business horizons*, **52,** pp. 357-365.

Matthes, J. (2015). 'Observing the "spiral" in the spiral of silence', *International Journal of Public Opinion Research*, **27**, pp. 155-176.

Matthes, J., Knoll, J., and C. von Sikorski (2018). 'The "spiral of silence" revisited: A meta-analysis on the relationship between perceptions of opinion support and political opinion expression', *Communication Research*, **45**(1), pp. 3-33.

Mende, M., M. L. Scott, J. van Doorn, D. Grewal and I. Shanks (2019). 'Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses', *Journal of Marketing Research*, **56,** pp. 535-556.

Metzger, M. J. (2009). The study of media effects in the era of Internet communication. In R. L. Nabi & M. B. Oliver (Eds.), *The Sage handbook of media processes and effects* (pp. 561–576). Los Angeles, CA: Sage.

Minaee, S., Azimi, E., and A. Abdolrashidi (2019). 'Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models', arXiv preprint arXiv:1904.04206.

Miranda, S. M., A. Young and E. Yetgin (2016). 'Are social media emancipatory or hegemonic? Societal effects of mass media digitisation in the case of the SOPA discourse', *MIS quarterly*, **40,** pp. 303-329.

MIT Media Lab (2018). The Spread of True and False Information Online. Available from: https://www.media.mit.edu/projects/the-spread-of-false-and-true-info-online/overview/

Mingers, J., and G. Walsham (2010). 'Toward Ethical Information Systems: The Contribution of Discourse Ethics', *MIS Quarterly*, **34** (4), pp. 833-854.

Moon, S. and W. A. Kamakura (2017). 'A picture is worth a thousand words: Translating product reviews into a product positioning map', *International Journal of Research in Marketing*, **34,** pp. 265-285.

Mostafa, M. M. (2013). 'More than words: Social networks' text mining for consumer brand sentiments', *Expert Systems with Applications*, **40** (2013), pp. 4241-4251.

Munger, K. (2017). 'Tweetment effects on the tweeted: Experimentally reducing racist harassment', *Political Behavior*, **39**(3), pp. 629–649.

Nadeem, W., Juntunen, M., Hajli, N. and M. Tajvidi (2019). 'The Role of Ethical Perceptions in Consumers' Participation and Value Co-creation on Sharing Economy Platforms', *Journal of Business Ethics*, **169**(3), pp. 421-441.

Rangel, F., and P. Rosso (2019). 'Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling', In Cappellato L., Ferro N., Müller H, Losada D. (Eds.) *CLEF 2019 Labs and Workshops*, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org.

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G. and S. Stieglitz (2019). 'Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks', *European Journal of Information Systems*, **28**, pp. 394-412.

Oh, O., M. Agrawal and H. R. Rao (2013). 'Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises', *MIS Quarterly***,** pp. 407-426.

PAN (2020). Cross-Domain Authorship Attribution 2019. Available at: https://pan.webis.de/clef20/pan20-web/author-profiling.html

Pew Research Center (April, 2018). Q&A: How Pew Research Center identified bots on Twitter. https://www.pewresearch.org/fact-tank/2018/04/19/qa-how-pew-research-center-identified-bots-on-twitter.

Petrič, G., A. Pinter (2002). 'From social perception to public expression of opinion: A structural equation modeling approach to the spiral of silence', *International Journal of Public Opinion Research*, **14**, pp. 37-53

Pham, B. T., Nguyen, M. D., Bui, K. T. T., Prakash, I., Chapi, K., and D. T. Bui (2019). 'A novel artificial intelligence approach based on Multi-layer Perceptron Neural Network and Biogeography-based Optimisation for predicting coefficient of consolidation of soil', *Catena*, **173**, pp. 302-311.

Potthast, M., Gollub, T., Wiegmann, M., B. Stein (2019). TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer.

Ptaszynski, M., P. Lempa, F. Masui, Y. Kimura, R. Rzepka, K. Araki, M. Wroczynski and G. Leliwa (2019). 'Brute-Force Sentence Pattern Extortion from Harmful Messages for Cyberbullying Detection', *Journal of the Association for Information Systems*, **20,** 4.

Ross, B., L. Pilz, B. Cabrera, F. Brachten, G. Neubaum and S. Stieglitz (2019). 'Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks', *European Journal of Information Systems*, **28,** pp. 394-412.

Sak, H., Senior, A. W., and F. Beaufays (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

Satariano, A. (June 14, 2019). Russia Sought to Use Social Media to Influence E.U. Vote, Report Finds, The New York Times. Available from: https://www.nytimes.com/2019/06/14/business/eu-elections-russia-misinformation.html

Schwenke, C., and A. G. Schering (2007). True positives, true negatives, false positives, false negatives. Wiley Encyclopedia of Clinical Trials.

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., and F. Menczer (2017). 'The spread of fake news by social bots', arXiv preprint arXiv:1707.07592, **96**, 104.

Sharda, R., Delan, D., and E. Turban (2013). *Business Intelligence, Analytics, and Data Science: A managerial Perspective (4th edition)*, Pearson Publication, New York.

Scheufele, D. A., Shanahan, J., E. Lee (2001). 'Manipulating the dependent variable in the spiral of silence research', *Communication Research*, **28**, pp. 304-324.

Shahid, F., Zameer, A., and M. Muneeb (2020). 'Predictions for COVID-19 with deep learning models of LSTM, GRU, and Bi-LSTM', *Chaos, Solitons & Fractals*, **140**, 110212.

Sheng, J., J. Amankwah-Amoah, X. Wang and Z. Khan (2019). 'Managerial Responses to Online Reviews: A Text Analytics Approach', *British Journal of Management,* **30,** pp**.** 315-327.

Shi, P., Zhang, Z., and K. K. R. Choo (2019). 'Detecting malicious social bots based on clickstream sequences', *IEEE Access*, **7**, 28855-28862.

Shi, W., Liu, D., Yang, J., Zhang, J., Wen, S., and J. Su (2020). 'Social Bots' Sentiment Engagement in Health Emergencies: A Topic-Based Analysis of the COVID-19 Pandemic Discussions on Twitter', *International Journal of Environmental Research and Public Health*, **17**(22), 8701

Shu, K., Sliva, A., Wang, S., Tang, J., and H. Liu (2017). 'Fake news detection on social media: A data mining perspective', *ACM SIGKDD explorations newsletter*, **19**(1), pp. 22-36.

Sigfusson, T. and S. Chetty (2013). 'Building international entrepreneurial virtual networks in cyberspace', *Journal of World Business*, **48**, **pp.** 260-270.

Somerville, I. (1999). 'Agency versus identity: actor-network theory meets public relations', *Corporate Communications: An International Journal*, **4**(1), pp. 6-13.

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Stieglitz, S., Brachten, F., Berthelé, D., Schlaus, M., Venetopoulou, C., and D. Veutgen (2017a, July). 'Do social bots (still) act different to humans?–Comparing metrics of social bots with those of humans', In *International conference on social computing and social media* (pp. 379-395). Springer, Cham.

Stieglitz, S., Brachten, F., Ross, B., and A.-K. Jung, (2017b). 'Do social bots dream of electric sheep? A categorisation of social media bot accounts', In *Proceedings of the Australasian Conference on Information Systems*, Hobart, Tasmania, Australia.

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., ... and F. Menczer (2016). 'The DARPA twitter bot challenge', *Computer*, **49**(6), pp. 38–46.

Su, J., Shirab, J. S., and S. Matwin (2011). 'Large scale text classification using semi-supervised multinomial naive bayes', In *Proceedings of the 28th international conference on machine learning (ICML-11)*, (pp. 97-104).

Tang, Q., B. Gu and A. B. Whinston (2012). 'Content contribution for revenue sharing and reputation in social media: A dynamic structural model', *Journal of Management Information Systems*, **29,** pp. 41-76.

Thorpe, A., R. Craig, D. Tourish, G. Hadikin and S. Batistic (2018). ''Environment'Submissions in the UK's Research Excellence Framework 2014', *British Journal of Management*, **29,** pp. 571-587.

Tong, S., and D. Koller (2001). 'Support vector machine active learning with applications to text classification', *Journal of machine learning research*, **2**, pp. 45-66.

Ullah, S., Ahmad, S., Akbar, S., and D. Kodwani (2019). 'International evidence on the determinants of organisational ethical vulnerability', *British Journal of Management*, **30**(3), pp. 668-691.

Van Doorn, J., M. Mende, S. M. Noble, J. Hulland, A. L. Ostrom, D. Grewal and J. A. Petersen (2017). 'Domo arigato Mr. Roboto: Emergence of automated social presence in organisational frontlines and customers' service experiences', *Journal of service research*, **20,** pp. 43-58.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and A. Flammini (2017).' Online human-bot interactions: Detection, estimation, and characterisation', In *Proceedings of the Eleventh International Conference on Web and Social Media (ICWSM-2017*), Montréal, Québec, Canada (pp. 280–289). Palo Alto, CA:The AAAI Press.

Vermeer, S. A. M., T. Araujo, S. F. Bernritter and G. van Noort (2019). 'Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media', *International Journal of Research in Marketing*, **36,** pp. 492-508.

Walsham, G. (1997). Actor-network theory and IS research: current status and future prospects. In *Information systems and qualitative research* (pp. 466-480). Springer, Boston, MA.

Wang, A. H. (2010). 'Detecting spam bots in online social networking websites: A machine learning approach', In *24th Annual IFIP WG 11.3 Working Conference on Data and Applications Security*.

Wang, X., Tajvidi, M., Lin, X. and N. Hajli (2019). 'Towards an ethical and trustworthy social commerce community for brand value co-creation: A trust-commitment perspective', *Journal of Business Ethics*, **167**(1), pp. 137-152.

Washington Post (2020). Analysis of millions of coronavirus tweets shows 'the whole world is sad. Available from: https://www.washingtonpost.com/science/2020/03/17/analysis-millions-coronavirus-tweets-shows-whole-world-is-sad/

Wedel, M. and P. Kannan (2016). 'Marketing analytics for data-rich environments', *Journal of Marketing*, **80, pp.** 97-121.

Wiegmann, M., Stein, B., & M. Potthast (2019, September). 'Overview of the Celebrity Profiling Task at PAN 2019', In *CLEF (Working Notes)*.

This paper has been accepted at **the British Journal of Management**. It is on the production stage and will go online soon.

Williams, C., J. Du and H. Zhang (2020). 'International orientation of Chinese internet SMEs: Direct and indirect effects of foreign and indigenous social networking site use', *Journal of World Business*, **55,** 101051.

Willnat, L., Lee, W., and B. H. Detenber (2002). 'Individual-level predictors of public outspokenness: A test of the spiral of silence theory in Singapore', *International Journal of Public Opinion Research*, **14**(4), pp. 391-412.

Wirtz, J., P. G. Patterson, W. H. Kunz, T. Gruber, V. N. Lu, S. Paluch and A. Martins (2018). 'Brave new world: service robots in the frontline', *Journal of Service Management*.

Woolley, S. C. (2016). 'Automating power: Social bot interference in global politics', *First Monday*.

Wuenderlich, N. V., and S. Paluch (2017). A nice and friendly chat with a bot: User perceptions of AI-based service agents.

Xia, W., Cao, M., and K. H. Johansson (2015). 'Structural balance and opinion separation in trust–mistrust social networks', *IEEE Transactions on Control of Network Systems*, **3**(1), pp. 46-56.

Xu, B., Shi, X., Zhao, Z., and W. Zheng (2018). 'Leveraging biomedical resources in bi-lstm for drug-drug interaction extraction', *IEEE Access*, **6**, pp. 33432-33439.

Yang, K. C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., and F. Menczer (2019). 'Arming the public with artificial intelligence to counter social bots', *Human Behavior and Emerging Technologies*, **1** (1), pp. 48–61.

Yin, C., Zhu, Y., Fei, J., and X. He (2017). 'A deep learning approach for intrusion detection using recurrent neural networks', *IEEE Access*, **5**, pp. 21954-21961.

Zeng, D., Chen, H., Lusch, R., and S. H. Li (2010). 'Social media analytics and intelligence', *IEEE Intelligent Systems*, **25**(6), pp. 13-16.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and B. Xu (2016, August). 'Attention-based bidirectional long short-term memory networks for relation classification', In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (pp. 207-212).

Zhou, X. and R. Zafarani (2018). 'Fake news: A survey of research, detection methods, and opportunities', arXiv preprint arXiv:1812.00315.

Zuckerman, M., DePaulo, B.M., R. Rosenthal (1981). 'Verbal and nonverbal communication of deception. In: Advances in experimental social psychology', **14**, pp. 1-59.