Social Media and Fake News Detection using Adversarial Collaboration

Karen M. DSouza
Information Systems
Kennesaw State University
kdsouza@students.kennesaw.edu

Aaron M. French
Information Systems
Kennesaw State University
afrenc20@kennesaw.edu

Abstract

The diffusion of fake information on social media networks obscures public perception of events, news, and relevant content. Intentional misleading news may promote negative online experiences and influence societal behavioral changes such as increased anxiety, loneliness, and inadequacy. Adversarial attacks target creating misinformation in online information systems. This behavior can be viewed as an instrument to manipulate the online social media networks for cultural, social, economic, and political gains. A method to test a deep learning model- long short-term memory (LSTM) using adversarial examples generated from a transformer model has been presented. The paper attempts to examine features in machine learning algorithms that propagate fake news. Another goal is to evaluate and compare the usefulness of generative adversarial networks with long-term short-term recurrent neural network algorithms in identifying fake news. A closer look at the mechanisms of implementing adversarial attacks in social media systems helps build robust intelligent systems that can withstand future vulnerabilities.

1. Introduction

Social media platforms have become a beacon for information reaching billions of users worldwide. Eight out of ten U.S. adults consume news from digital platforms with over half of the turning to social media as a source for news [1]. News stories containing falsehoods have been shown to spread faster and broader than truthful information [2]. Fake news spreads on social media through engagement behaviors (i.e., sharing, liking, or commenting on news stories), which are significantly influenced by confirmation bias [3]. During a crisis, such as the COVID-19 pandemic, the impact of misinformation, and fake news shared on social media can have potentially damaging effects on society and the ability to manage the crisis.

Social networking platforms have turned to machined learning to identify and flag fake news. However, adversarial attacks on machine learning systems attempt to falsify information by inserting false inputs and misleading public opinion. Neural networks such as long short-term memory (LSTM) networks have been used in speech recognition and machine translation. More recently, in 2019, a deep learning neural network called the generative pre-trained transformer (GPT2) model has achieved success in its ability to synthesize natural language. A special kind of neural network called the generative adversarial network (GAN) has been gaining popularity for image manipulations since its introduction in 2014. GANs consist of two machine learning systems – the generator and the discriminator that train each other to produce powerful unique and often fake results. GANs are used to generate extensive fake images that are indistinguishable from real images, thus aiding in the propagation of fake news.

Fake news stories can be generated from seemingly official sources. Fake content created by adversarial algorithms mimics the pattern of real news. The errors in text are so slight that they can be overlooked by an untrained eye. The results appear real and have the power to mislead people and thus influence human opinion or behavior. While GANs have been widely implemented to create fake images, a pitfall of GANs is that it is hard to use them in detection models to identify discrete data in fake online reviews, comments, and opinions. This is because the structure of the Text GANs makes it difficult to pass gradients from the discriminator to generator modules [4].

Several challenges exist in analyzing fake news in social media. The dynamic nature, complexity, and diversity of fake news generated by adversaries poses a challenge in detecting the threat rapidly. Past work has not been successful in examining fake news created through adversarial examples. The absence of high-quality fake news training sets also adds to the challenge of developing fake news detection models.

URI: https://hdl.handle.net/10125/79344
978-0-9981331-5-7

Verifying the source of the fake news is another issue. Since fake news is generated to mislead readers, detection algorithms are unable to classify the news as fake only due to the content which could be semantically and visually correct.

The research examines machine learning algorithms to determine the feasibility of generating adversarial examples and detecting fake news in social media. The objectives of this research are to 1) examine adversarial behavior in propagating fake news in social media and 2) evaluate and propose methods of generating and detecting adversarial examples through machine learning in fake news experiments. The contributions lie in the deep learning LSTM model which was able to train and successfully detect real and fake news articles that were chosen from social media. Also, the research used an existing GPT2 model to create fake adversarial examples. These fake examples were again input back to the LSTM model and successfully classified as fake. Using results from the GPT2 model back into the LSTM model has significance becauseit shows the feasibility of future collaborative research between deep learning neural networks and specificallytransformer models for detecting fake news.

To carry out this research, we first review background information on fake news, detection algorithms, adversarial behaviors, GANs, SeqGAN, LSTM and GPT2 networks. Section 3 describes the method followed by our analysis of the results in section 4. Section 5 consists of findings and contributions, and section 6 describes the conclusion and future work.

2. Background

The number of social media users worldwide is projected to be almost 4 billion users in 2022. Adversarial agents take advantage of social media networks by infiltrating an existing narrative with fake news and then amplifying the messages through social bots. This fake news spreads quickly through social media to millions of people in just milliseconds. The increasing high number of social media users worldwide raises concerns about the effects of fake news as it propagates through millions of people in just milliseconds. This section is divided into four parts to provide a background to understand fake news and how adversarial agents apply manipulations to influence behaviors.

2.1. Fake News and Social Media

Fake news was popularized in 2016 during the presidential election as disinformation spread across social media seeking to influence election results [5]. There are several definitions of fake news in existence. One definition makes the distinction between fake and genuine news. Fake news is counterfeit news, while genuine news is comprised of news that has gone through fact checkers and editors [6]. Another definition of fake news refers to news articles that contain verifiably false information intentionally created to mislead others [7]. In our research, we will use this definition of fake news.

The deceptions through false news stories are exacerbated due to the speed at which information travels through social networking services, such as Facebook, Twitter, and various other popular mediums. Based on a worldwide survey conducted from 2011-2020, the percentage of adults that trust news via social media has dropped from 45% to 35% during the past decade. These results are displayed in Fig 1.

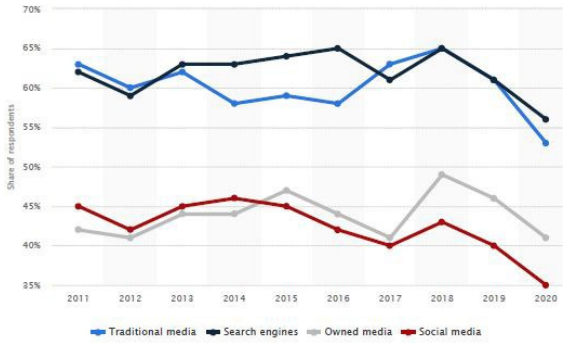

Figure 1. Most Trusted Sources of News from 2011to 2020 Surveyed Worldwide [8]

There is an urgent need to combat fake news in social media by professionals both in the industry and academia. Due to the detrimental outcomes resulting from the consumption and continued spread of fake news, many social media platforms are implementing fake news detection algorithms and alerts to mitigate the adverse effects. In 2020, Twitter updated the company's approach to fake news to include three categories – misleading information, disputed claims, and unverified claims [9]. Misleading information comprises of falsified statements that are confirmed to be misleading by experts in the subject area. Disputed claims are statements that lack veracity of facts. The credibility of the information is unknown and has not been verified. Information that cannot be verified are labeled unverified claims. In each of these categories, Twitter has introduced labels that indicate a propensity for harm. This alerts the online community of potential warnings and removal notices. The social media ecosystem has become an important tool that companies use to influence consumers by integrating marketing efforts

online and connecting with potential target audiences [10].

While social media is the least trusted source of news at 35%, this number still poses a significant threat as one third of the population is susceptible to fake news manipulations. An online survey of over 6000 respondents conducted in March 2019 showed that 67% of people experienced a great deal of confusion from misleading fake news sources. This is shown in Fig. 2 below.

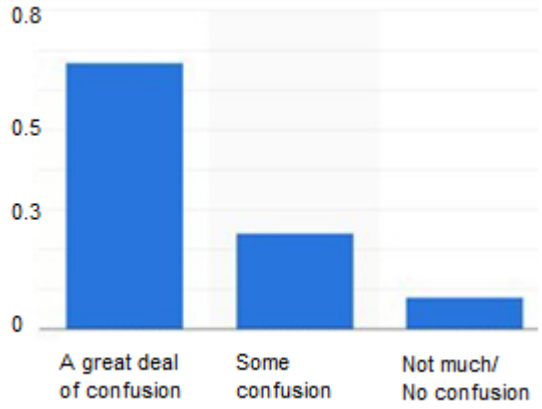

Figure 2. Statistic of the number of respondents versus their level of confusion in an online survey in March 2019 [11]

Traditional methods of using filtering algorithms to detect malicious content fail to recognize the presence of adversaries [12]. Adversarial nodes have been proven successful in preventing consensus in social media [13]. In the past, Twitter had introduced a "Get the facts" alert link to tweets flagged misleading by their algorithms. One example of this proactive approach and reaction taken by Twitter officials is the Twitter alert link that was tagged to alleviate U.S. sentiment surrounding the then U.S. President's tweets about voter fraud [14].

The success and applicability of machine learning algorithms depend on the generality of the algorithms [15]. Future research directions involving adversarial attacks and defense mechanisms are on the forefront today [16]. Facebook uses an active journalism project that enables its fact-checking partners to provide several ratings to published content [17]. The ratings were introduced with the intent to provide additional context to readers about misinformation, fake news and manipulated posts. The first rating – Altered, applies to images, and videos that have failed the Facebook community standards. Another rating - Missing Context alerts users to misleading articles that are not substantiated with matching content. Published content with a rating either False or Altered indicates misinformation. This content is subject to aggressive reduction in distribution. Content that contains partial inaccuracies is labeled Partly False and is subject to less aggressive blocking that a False or Altered article [17].

2.2. Fake News Detection Algorithms

According to Su et al, fake news detection research in social media can be viewed from four general perspectives - data-oriented, feature-oriented, model-oriented, and application-oriented approach [18]. This is displayed in fig. 3 below. In the data-oriented approach, the properties of the dataset, temporal and psychological aspects are considered in the detection algorithms. However, there is no standard or guideline that has been established to evaluate the dataset itself. Temporal challenges include the rapid speed at which social media information propagates and dynamic nature of news. Also, psychological approaches are difficult to quantify.

Figure 3. Fake news detection in social media approaches [18]

In the feature-oriented approach, the news content and social context play a primal role. Natural language processing (NLP) detection algorithms extract textual features to detect fake news. However, NLP based algorithms do not rely on fact checking and have become the subject of adversarial attacks. To combat the attacks, linguistic characteristics must be checked together with fact checking [19].

The model-oriented approach utilizes supervised machine learning techniques such as decision trees, k-nearest neighbor, support vector machines and logistic regression. Unsupervised or semi-supervised models are difficult to build but they are more practical for fake news detection. Research on natural language tasks using machine learning techniques has progressed in reducing exposure bias on training sets by reducing the temperature parameter. Language GANs fall short and underperform when compared to maximum likelihood estimation models [20].

LSTMs have been used to detect fake news in various methods. By adding speaker profiles such as party affiliation, speaker title, location, and credit

history, Long et al have achieved 14.5% higher accuracy compared to traditional methods [21]. Other researchers have resorted to adding part of Speech tags to speaker profiles in bi-directional LSTMs in conjunction with convolutional neural networks improve accuracy in a hybrid architecture [22]. However, these modified algorithms work only when the added attributes are available. In our research motivation, we are tackling the issue when additional details of the news article is not provided on social media channels.

The application-oriented approach consists of fake news diffusion and fake news intervention [18]. Fake information diffusion follows patterns in social media. A GAN style approach that utilizes information campaigns for rumor detection achieved 86% accuracy on publicly available Twitter dataset [23]. Fake news Intervention involves immunization techniques such as Hawkes process algorithms to isolate news from directed consumers [19].

2.3. Adversarial Behaviors

Adversarial behaviors stem from the existence of adversarial roles and motivations. From a definitional standpoint, it refers to two entities that oppose each other. Within information systems research, adversarial behaviors have been linked to cybersecurity to describe the adversarial roles of nefarious actors and security experts who protect resources. The MITRE ATT&CK Matrix presents a taxonomy of adversarial behaviors consisting of tactics and techniques used to compromise networks, systems, and people (see Appendix A). This consists of 14 different tactics comprised of 215 techniques and 498 sub-techniques used to compromise networks, systems, and data [24].

An example of adversarial behaviors is the use of phishing attacks to obtain data. Successful phishing attached consist of the effective exploitation of human weakness through social engineering tactics [25]. The adversarial behavior consists of the bad actor seeking to compromise the targets data through psychological manipulations. While prior research has focused on adversarial behaviors to manipulate, interrupt, or destroy systems and data, there are new impacts that result from current trends such as fake news.

The low cost of social media accounts gives rise to spam bots. Spam bots inflict harm by closely following social media trends to plan and organized collaborative spam attacks to sway public opinion. Adversarial behaviors would also include social engineering through the creation of fake news stories that seek to influence perceptions and behaviors. Certain adversarial outcomes could be achieved such as influencing outcomes of elections or interfering with the handling of the COVID-19 pandemic. While social networking sites such as Facebook and Twitter have implemented fake news detection algorithms using machine learning, the refinement of adversarial attacks has also increased in sophistication. A class of machine learning known as Generative Adversarial Networks (GAN) have been created to manipulate fake news identifiers resulting in the misclassification of fake news and its continued spread. In text classification applications such as detecting fake opinions which influence behavior, GANs have found some success [26].

2.4. Generative Adversarial Networks (GAN)

GANs are deep learning based generative models based on the adversarial mini-max game theory. Introduced in 2014, GANs demonstrated the viability of using two neural networks with competing adversarial objectives to create generative models [4].

The system consists of generator and discriminator neural networks. The generator neural network was proposed as an unsupervised training algorithm that generates outputs based on an original dataset and noise. The discriminator neural network is a classifier that distinguishes between real and fake results from the generator as shown in Fig. 4. The generative network produces samples that attempt to confuse the discriminator, while the discriminator attempts to discern real and fake images.

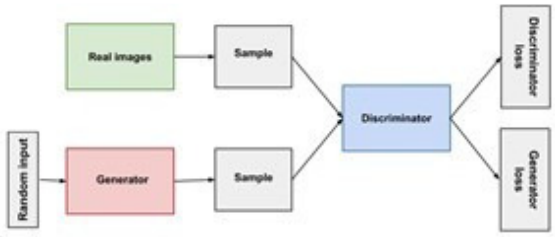

Figure 4. Generative Adversarial Network (GAN) Structure [27]

A type of GAN called the Sequence Generative Adversarial Networks or SeqGAN made improvements in text generation using a sequence generator in the decision-making process for text generation [28]. A policy gradient is applied to the output from the discriminator. The training in the generator continues with rewards assigned to each Monte Carlo search as displayed in Fig. 5 below.

The examples of fake and true news sentences generated from the discriminator in an experiment

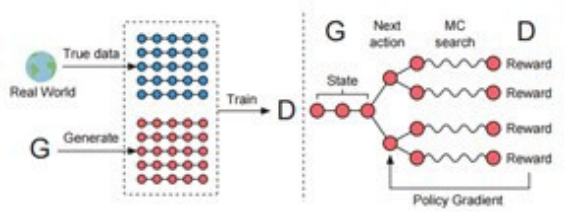

Figure 5. SeqGAN – Sequence Generative Adversarial Network Structure [28]

performed by Yu et al are displayed in the fig. 6. below [28]. The figure shows better performance to Obama's fake speech generation from SeqGAN than MLE algorithm. From the real-life data, the experiment was able to prove the effectiveness of SeqGAN in creating adversarial political speech by training on prior political speeches.

Obama political speech text generation

Algorithm	BLEU-3	*p*-value	BLEU-4	*p*-value
MLE	0.519	$< 10^{-6}$	0.416	0.00014
SeqGAN	**0.556**		**0.427**	

Figure 6. Experiment results from an implementation of SeqGAN [28]

2.5. Long-term Short-term Memory Recurrent Neural Network

A deep learning method using recurrent neural networks with long-term short-term memory (LSTM) can be implemented with improved results for classification of fake news in natural language processing [29]. The architecture of the network is displayed below in Fig. 7. The set up referred to as a semantically controlled LSTM introduced by Wen et al was successful in producing natural responses to colloquial language [29].

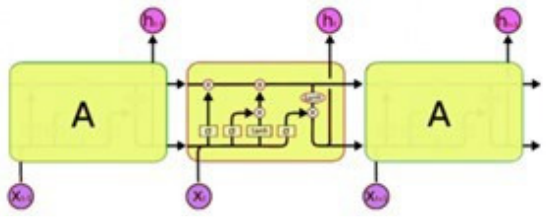

Figure 7. LSTM long-term short-term memory recurrent neural network [30]

2.6. Generative Pre-trained Transformer2 (GPT-2) model

The Generative Pre-Trained Transformer2 (GPT-2) model is a general-purpose learner launched in February, 2019. It is a special kind of deep learning network. The model is unique in its ability to generate an entirely fake story complete with quotations from a single input sentence. What sets the GPT-2 model apart from other deep learning models is its characteristic attention modeled around cognitive attention. 1.5 billion parameters are included in the GPT-2 package so this makes its outputs highly convincing and believable to the human eye [31].

3. Method

First, an attempt to classify real and fake news using a recurrent neural network called Long-term Short-term memory (LSTM) on a public dataset is made. Second, recent news feeds from websites are tested on the LSTM model to determine if the model can identify fake or real news. Third, the GPT-2 model is used to generate fake news. This fake news is tested on the LSTM model for fake news classification. Finally, the SeqGAN model is used as a comparison to the LSTM model.

3.1. Fake news classification using LSTM Model

The LSTM model was chosen for this task because this is a deep learning model that has feedback connections, making it possible to not only process single data points, but also entire sequences of data. Since the dataset has sequences of data and a high diversity ratio in fake news, LSTM is an ideal choice. The algorithm inherently enables backpropagation of error through time and layers. This preserves the structure. For the LSTM method, the data used is from a publicly available Kaggle dataset which comprises of a collection of news articles - both real and fake news [32]. The dataset consists of four fields - title, text, subject, and date. A grouping of the subjects reveals five main subjects- news, politics, government news, left-news, U.S. news and Middle East.

An initial comparison of the word frequency from fake news and real news reveal that that "Donald Trump" and "said" were the top words in the fake news dataset. The "U.S." and "said" featured in the top words in the real dataset. The word clouds are good indicators of the diversity of words in the datasets. The lexical diversity ratio of the dataset is calculated as the number of unique words in the target subject over number of words in target categories. The

fake news set contained two times more lexical diversity and more punctuation such as exclamation marks. The total number of sentences were 44,897. Fake texts comprised of 23,481 sentences and only 21,416 real texts. The number of fake news texts was higher than the real news texts, so the algorithm had more fake data to train and classify fake news with a high accuracy rate. The datasets are subject to preprocessing of data using Python. Special characters and stop words are eliminated. Real news is assumed to have verified publishers. Fake news can have missing or anonymous publishers. The sentences are then broken down into individual words and tokenized. A weight matrix is created for each token in the dataset.

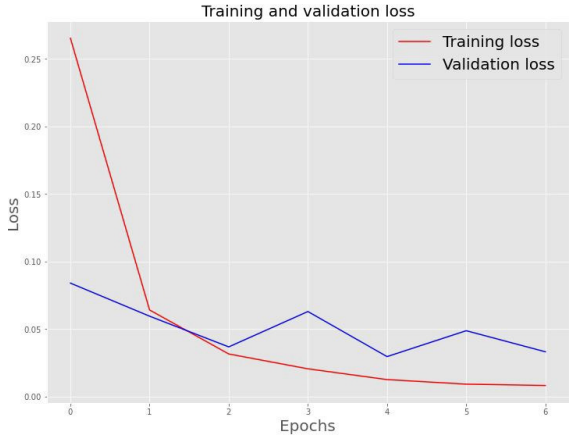

Figure 8. LSTM model - training and validation lossover six epochs

The LSTM is designed using Python's TensorFlow and Keras package. TensorFlow is an open-source machine learning application. Keras is a deep learning API that runs on TensorFlow. After the classification, a time series analysis of fake news indicates the periods when real and fake news were in circulation. This classification reveals indications of the time period when fake news was being shared more often than others. Other insights from the plot of the time series can reveal interference of adversaries in influencing public opinion. Fig. 8 shows the plot of training and validation loss over six epochs of the training cycle in the model. With an optimum choice epoch, the LSTM model achieved the highest accuracy of 98.9% for fake texts.

3.2. Fake news generation using GPT-2 Model

GPT-2 is a publicly available large language model with 1.5 billion parameters. The diversity of GPT-2 model can be explained further as it has been trained on a dataset of 8 million web pages. As a result, it is capable of generating text samples of exceptional quality. This makes it an ideal choice for generation of fake news examples in our research.

In this step, the generative pre-trained transformer2 (GPT-2) model is used solely to generate fake news with the intent to test this data on the LSTM model. Several rounds of data collection experiments are carried out to collect fake data from GPT-2 model. The transformers library from Python provided a seamless API to install GPT-2 model.

3.3. Adversarial Text Generation using SeqGAN

In the implementation of the SeqGAN model, the LSTM model was chosen as the generator, while discriminator was a continuous neural network. Using the SeqGAN model, the generator begins creative adversarial samples to trick the discriminator. The algorithm did not generate accurate classification results and was inconsistent. Additional preprocessing of data and a continuous refinement of the parameters in the algorithm was insufficient to accurately classify the fake news.

4. Analysis

The analysis on the LSTM model is discussed further conducting several live experiments on real news currently circulating in the media. Randomly selected news articles from social media sites such as CNN's twitter feed were copied to a text file and then imported to the LSTM model. For example, a news headline reporting the daily number of Covid-19 cases in Boston and further reporting. The LSTM model correctly classified the news as real. Several such iterations of recent news articles were tested. One sweeping observation was the fact that when the input text was considerably large and consisted of several sentences, the model was able to correctly classify it as real or fake.

Further analysis involved testing grammatically incorrect sentences from a user generated input field created in Python. The model correctly classified the text as fake. One possible reason for the success could be that the deep learning network was trained well and exposed to similar grammatically correct and incorrect sentences from the news articles in the training phase. The results from LSTM method show that it is easier to compute the percentage of fakeness without the need for a policy gradient algorithm such as SeqGAN.

5. Discussion

The results from the LSTM model are promising because they can be used to classify fake news from social media especially in cases when news incidents go viral, and it is required to ascertain reliable news sources quickly. Future comparisons can be made on Covid-19 or travel related news articles and additional public datasets to monitor the spread of fake news in these realms. In the past year, instances when fake health news concerning Covid-19 were propagating fear in individuals, the presence of a fake news classifier would have increased confidence in media outlets supporting positive claims. The LSTM method can be used to train a deep learning neural network from multiple social media news sources. For instance, the real time data feed from Twitter can be compared to the historical tweets to classify fake news in real time.

Additional research on the capability of the SeqGAN to generate fake texts through the random noise input can provide insights into the future of adversarial attacks. The unsupervised learning algorithm produces outputs that can be trained well in the Monte Carlo chain sequences. This allows the SeqGAN to be manipulated by adversaries intending to perfect the technique of generating fake news.

A solution to determining which news is fake can be made public through the construction of an online database where individuals can submit their fake news request. The underlying algorithm that drives the database can be a LSTM or SeqGAN. The proposed database can produce a classification or confusion matrix based on the probability of true or fake news. The value of a public news classification database can be further enhanced to include a cost matrix. Here, individuals can assign a cost value to each output of the confusion matrix to view the effects of fake news to their organization or society. The scalability and stability of the project are areas that need to be further explored in future work.

6. Conclusion

The spread of fake news through social media is a challenge to curtail. In the industry, social media companies are intensifying their efforts to combat fake news. There are no guidelines that can be enforced to prevent the malicious content from spreading. However, fake news classifiers and generators can prove useful in estimating the veracity of news. In conclusion, the LSTM method is preferred in situations when data can be sequenced into a network. This makes LSTM recurrent networks powerful solutions for fake news classifications. As people start using the GPT-2 algorithm, more research should be done in this area to improve the methods of defense against adversarial examples. GANs have been designed to perform well with continuous data sources such as images, music, and voice generation. On the other hand, realistic human language generation in GANs has been a challenge due to the discrete nature of the text structure. Since fake text generation like human language can be produced using the SeqGAN algorithm, this can be used as a source of adversarial behavior. In the future, social media networks need to adopt advanced intelligent systems capable of detecting adversarial examples.

7. References

[1] E. Shearer, "More than eight-in-ten americans get news from digital devices," *Pew Research Center*, vol. 12, 2021.

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[3] A. Kim, P. L. Moravec, and A. R. Dennis, "Combating fake news on social media with source ratings: The effects of user and expert reputation ratings," *Journal of Management Information Systems*, vol. 36, no. 3, pp. 931–968, 2019.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[5] T. Lee, "The global rise of "fake news" and the threat to democratic elections in the usa," *Public Administration and Policy*, 2019.

[6] D. Fallis and K. Mathiesen, "Fake news is counterfeit news," *Inquiry*, pp. 1–20, 2019.

[7] P. L. Moravec, A. Kim, and A. R. Dennis, "Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media," *Information Systems Research*, vol. 31, no. 3, pp. 987–1006, 2020.

[8] A. Watson, "Most trusted sources of general news and information worldwide from 2011 to 2020," 2021.

[9] Y. Roth and N. Pickles, "Updating our approach to misleading information," 2020.

[10] R. Hanna, A. Rohm, and V. L. Crittenden, "We're all connected: The power of the social media ecosystem," *Business horizons*, vol. 54, no. 3, pp. 265–273, 2011.

[11] A. Watson, "Level of confusion caused by fake news about the basic facts of current issues and events in the United States as of March 2019," 2019.

[12] S. Yu, Y. Vorobeychik, and S. Alfeld, "Adversarial classification on social networks," *arXiv preprint arXiv:1801.08159*, 2018.

[13] C. Hajaj, S. Yu, Z. Joveski, and Y. Vorobeychik, "Adversarial coordination on social networks," *arXiv preprint arXiv:1808.01173*, 2018.

[14] T. Hatmaker, "Twitter adds a warning label fact-checking Trump's false voting claims," 2020.

[15] B. Kuchipudi, R. T. Nannapaneni, and Q. Liao, "Adversarial machine learning for spam filters," in *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pp. 1–6, 2020.

[16] B. Guo, Y. Ding, L. Yao, Y. Liang, and Z. Yu, "The future of false information detection on social media: New perspectives and trends," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–36, 2020.

[17] K. Goldshlager and A. Berman, "New Ratings for Fact-Checking Partners," 2020.

[18] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[19] Z. Zhou, H. Guan, M. M. Bhat, and J. Hsu, "Fake news detection via nlp is vulnerable to adversarial attacks," *arXiv preprint arXiv:1901.09657*, 2019.

[20] M. Caccia, L. Caccia, W. Fedus, H. Larochelle, J. Pineau, and L. Charlin, "Language gans falling short," *arXiv preprint arXiv:1811.02549*, 2018.

[21] Y. Long, "Fake news detection through multi-perspective speaker profiles," Association for Computational Linguistics, 2017.

[22] M. K. Balwant, "Bidirectional lstm based on pos tags and cnn architecture for fake news detection," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, IEEE, 2019.

[23] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors on twitter by promoting information campaigns with generative adversarial learning," in *The World Wide Web Conference*, pp. 3049–3055, 2019.

[24] Mitre, "Enterprise Matrix," 2021. accessed 2021-01-31.

[25] P. Rajivan and C. Gonzalez, "Creative persuasion: a study on adversarial behaviors and strategies in phishing attacks," *Frontiers in psychology*, vol. 9, p. 135, 2018.

[26] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting deceptive reviews using generative adversarial networks," in *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 89–95, IEEE, 2018.

[27] G. D. Site, "Overview of GAN Structure," 2019.

[28] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[29] T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," *arXiv preprint arXiv:1508.01745*, 2015.

[30] C. Olah, "Understanding LSTM Networks," 2021.

[31] J. C. Irene Solaiman and M. Brundage, "GPT-2: 1.5B Release," 2019.

[32] Kaggle, "Fake and real news dataset," 2020.

Appendix A: Adversarial behaviors enterprise matrix (MITRE ATT&CK Matrix)

Tactic	**Techniques (Sub-techniques)**	**Description**
Reconnaissance	10 (31)	Gather information for future operation
Resource Development	7 (31)	Establish resources for future operation
Initial Access	9 (10)	Gain access to the network
Execution	12 (22)	Run malicious code
Persistence	19 (82)	Maintain a present on the network
Privilege Escalation	13 (82)	Gain higher level permission and access
Defense Evasion	39 (116)	Avoid detection
Credential Access	15 (40)	Obtain count and access credentials
Discovery	27 (12)	Learn about your environment
Lateral Movement	9 (12)	Navigate the environment
Collection	17 (18)	Obtain and collect data of interest for goals
Command and Control	16 (22)	Communicate with and control systems
Exfiltration	9 (8)	Steal data
Impact	13 (12)	Manipulate, interrupt, or destroy systems & data.

Source: https://attack.mitre.org/matrices/enterprise/# [22]