Social Media Manipulation Awareness through Deep Learning based Disinformation Generation

Clara Maathuis[1] and Iddo Kerkhof[2]
[1]Open University of the Netherlands, Heerlen, The Netherlands.
[2]NLP Software Engineer, The Netherlands.

clara.maathuis@ou.nl
email@iddo.eu

Abstract: As a digital environment introduced for establishing and enhancing human communication through different social networks and channels, social media continued to develop and spread at an incredible rate making it difficult to find or imagine a concept, technology, or business that does not have or plan to have its social media representation and space. Concurrently, social media became a playground and even a battlefield where different ideas carrying out diverse validity degrees are spread for reaching their target audiences generated by clear and trustable well-known, uncertain, or even evil aimed entities. In the stride carried out for preventing, containing, and limiting the effects of social manipulation of the last two types of entities, proper/effective security awareness is critical and mandatory in the first place. On this behalf, several strategies, policies, methods, and technologies were proposed by research and practitioner communities, but such initiatives take mostly a defender perspective, and this is not enough in cyberspace where the offender is in advantage in attack. Therefore, this research aims to produce social media manipulation security awareness taking the offender stance by generating and analysing disinformation tweets using deep learning. To reach this goal, a Design Science Research methodology is followed in a Data Science approach, and the results obtained are analysed and positioned in the ongoing discourses showing the effectiveness of such approach and its role in building future social media manipulation detection solutions. This research also intends to contribute to the design of further transparent and responsible modelling and gaming solutions for building/enhancing social manipulation awareness and the definition of realistic cyber/information operations scenarios dedicated/engaging large multi-domain (non)expert audiences.

Keywords: information operations, cyber operations, social manipulation, disinformation, misinformation, security awareness, machine learning, deep learning.

1. Introduction

"The success of manipulation depends on the level of conviction and force of the denial." (Tess Binder)

It would be difficult to recall a modern or current societal and technological trend, topic, or event that is not projected in the digital realm through social media discourses and not surrounded by diverse manipulation mechanisms like disinformation and misinformation (Maathuis & Chockalingam, 2022b). In this realm, agents, e.g., public opinion organizations, independent bodies, and civilians carry out different activities and engage (un)intentionally with diverse manipulation forms which impact their behaviour individually and collectively (Bastick,2021; Chockalingam & Maathuis, 2022). Engaging the target audience through social manipulation can be done through vectors, e.g., user profiling being (i) demographic meaning identifying individuals' unique characteristics, beliefs, needs, and vulnerabilities, and (ii) psychometric implying personality-based segmentation, and micro-targeting based on analyzing/altering audience's personal actions (Kertysova, 2018; Fard & Maathuis, 2021). While efforts dedicated to limiting, controlling, and preventing social manipulation using AI-based techniques, e.g., Twitter and Facebook relying on Machine Learning-based solutions for stamping out trolls, finding and removing fake bot accounts, and proactively identifying sensitive content (Perez-Escolar et al. 2021), still these mechanisms are insufficient as the attackers are intelligent in developing adaptive techniques that succeed on bypassing defending mechanisms and reach their goals. Hence, changing the paradigm and treating this phenomenon through the eyes of the attacker while considering building datasets and transparent intelligent solutions (Maathuis, 2022b) that model and draw relevant principles and requirements for building awareness and defending techniques could be (the basis of) a solution in this direction. Hence, this research aims to build social manipulation security awareness through a deep learning model for generating disinformation tweets. To achieve this objective, multidisciplinary approach is conducted by merging studies from social media, cyber security, information/cyber operations, and deep learning domains through a Design Science Research methodology following a Data Science approach having the following contributions:

- Opening a path to scientific and practitioner communities taking offender's stance when building/enhancing existing social manipulation security awareness through content generation.

- Encouraging increasing the variety of social media security awareness solutions that are transparent and responsible being set up by security, social media, and AI experts dedicated to multiple audiences by integrating various disciplines/visions.
- Supporting decision-making processes of policy, security, and social media decision-makers when building strategies and policies for preventing, containing, and/or limiting social manipulation through human and technological-based solutions.

The remainder of this article is structured as follows. In Section 2, the background and related studies of this research are discussed. In Section 3, the research approach taken is elaborated while presenting research activities taken. In Section 4, the deep learning techniques used to build the model are tackled. In Section 5, the implementation process is depicted together with the results obtained. In Section 7, concluding remarks on the findings obtained and future research ideas are addressed.

2. Social Manipulation Research

In this section the multidisciplinary nature of this research is expressed by discussing studies, methods, and strategies relevant to achieving the aim of this research. The EU Commission stresses the importance of protecting values and democratic systems by tackling social manipulation mechanisms like disinformation, defined as "false or misleading content that is spread with an intention to deceive or secure economic or political gain" (EU Commission, 2022a). Among the initiatives established are the Action plan on disinformation, the European Democracy Action Plan, the Covid-19 disinformation monitoring programme, and the Strengthened Code of Practice on Disinformation which calls for values like transparency of political advertising, empowering researchers (e.g., through public data availability), and empowering the fact-checking community (EU Commission, 2018; EU Commission, 2022a; EU Commission, 2022b). For increasing digital resilience and combating disinformation for achieving ideological, political, or financial goals, NATO acknowledges the damage done to "citizens' faith in the institutions of democratic governments and resources of public information and discussion" (NATO, 2021a) and calls for developing fact-based solutions and credible public communications based on (i) understanding and analysing the information environment, (ii) engaging audiences to build resilience, and (iii) communicating proactively and exposing major disinformation cases (NATO, 2021b). As such solutions are currently implemented integrating AI, according to Rathenau (2022), they can help managing credibility; however, can also amplify conspiracy.

It is then important to take a step back and see the whole context. Accordingly, Ki-Aries & Faily (2017) argue for designing user-specific security awareness solutions by incorporating human factors in design through a series of targeted and actionable topics that consider persona characteristics, i.e., (i) goal directed perspective, (ii) role-based approach, (iii) fiction-based perspective through human intuition and assumptions, and (iv) engaging perspective through data use. Reisach (2021) analyzes the role and responsibility that social platforms have on conducting diverse businesses, expressing political discourses and election decisions, integrating AI techniques for different functionalities, and related to the formation, execution, and impact produced by different manipulation incidents. Moreover, the author argues that responsibility is shared and global – of society, social platforms, and users. To support it in a proactive way, the author argues for building education solutions oriented on (i) tech and media literacy, (ii) educational support for vulnerable populations, and (iii) education and awareness to foster an ethical mindset. Chen, Chen & Xia (2022) analyze user's behavior in social networks arguing that social media is becoming an environment for cyber weapons deployment (Maathuis, Pieters & van den Berg, 2016) and a force multiplier in warfare through goals that include public opinion reconnaissance, sentiment analysis, and active intervention. The authors reflect on the actors involved, nature of the environment (warfare versus infosphere), and the arsenal used, e.g., bot, botnet, troll, cyborg, plus behavior information representation, user recognition, and homogeneous community discovery. Shu (2022) conducts a survey on recent advancements and stresses the existing data challenge, i.e., availability of public datasets useful for understanding manipulation styles/methods. While the author acknowledges the significant use of deep learning, the argument for building not just proactive, accurate, and robust disinformation countering solutions is put forward, and as important, that the solutions should be trustworthy, i.e., investigating how to (i) transform meaningful human cognition into knowledge, (ii) incorporate noisy, incomplete, and complex human feedback for better representation learning, and (iii) interpret prediction results with knowledge reasoning and causal discovery. An experiment was conducted by Bastick (2021) with 233 undergraduate students for investigating behavioral effects of fake news. The results reveal that fake news have the real potential of altering individuals' behavior and cluster disinformation effects as empirically observed, refined, and predicted by the malicious agents producing them. Moreover, the author calls for urgent development of solutions that generate

experiments for strengthening users' protection from manipulation mechanisms. Perez-Escolar et al. (2021) conduct a study with 150 students for strengthening fact-checking skills and project-based learning, and mention formulating judgements based on incomplete or limited information as a major competence.

On AI applications, Kertysova (2018) stresses that they are double-edged since they provide useful tools for building powerful and scalable solutions for countering social manipulation while they imply own set of limitations and unintended consequences. Additionally, AI is the backbone since through its techniques facilitates the development and spreading of social manipulation content to large audiences. Particularly, the model of Kula, Kozik & Choras (2021) is based on a BERT architecture named RoBERTa for dealing with Covid-19 fake news using a mix of datasets. The results obtained are promising (accuracy above 90%) and reflect the potential of such techniques. In the survey by Wang et al. (2022), the authors propose as further research developing proper interpretation methods, enhancing robustness to possible security and privacy attacks, properly handing with imbalanced data, and merging GAN with other deep learning techniques. An example of successfully combining GAN with BERT techniques is developed by Yu et al. (2020) for representing medical concepts and discovering annotation inconsistency on medical labels. Moreover, Shu, Li, Ding & Liu (2021) generate synthetic news through the FACTGEN (Fact-Enhanced Synthetic News Generation) method which implements a PSA (Pseudo-Self-Attentive) language model that tackles the challenges on generating news content related to a given claim and ensuring that the generated content contains supplemental fact information. For generating real facts through (journalistic) news, Leppanen et al. (2017) consider that the following criteria should be respected: transparency, accuracy, system modifiability and transferability, output fluency, data availability, and news topicality. These criteria are considered when building the artefact proposed in this research. However, strengthening social media security spanning cyber incidents' life cycle implies building not only AI-based security awareness solutions for educational purposes, but also building intelligent solutions for testing attack resistance against possible cyber attacks. Le, Wang & Lee (2020) generate malicious content for attacking existing security mechanisms embedded in deep learning-based fake news detection models. The solution proposed is based on GAN and contains the attack module, conditional text generator, and style module. Nevertheless, a limited number of disinformation datasets are available and the necessity of proposing new ones is experienced by researchers and practitioners in this domain. Accordingly, Kim et al. (2021) propose the FibVID (Fake news information-broadcasting dataset of Covid-19) with tweets containing Covid-19 and non-Covid-19 fake news for analysis and diffusion of fake news. Another dataset is advanced by Patwa et al. (2021) focusing on Twitter with data labelled as real and fake used for building baseline ML models.

As this extensive literature study reflects, the harm produced by disinformation is experienced by society, by social platforms, and their users. While different incentives and mechanisms for proactively and reactively tackling disinformation incidents exist, social media security awareness to such manipulation mechanisms is still in its infancy due to issues like limited data availability, lack of multidisciplinary/interdisciplinary/transdisciplinary/anti-disciplinary expertise, fast technological advancements, and the existing (un)known vulnerabilities of humans and social platforms. This is the motivation behind our research.

3. Research Methodology

Aiming at building an intelligent model that generates social manipulation disinformation content in a transparent and responsible way for building/enhancing social media security awareness of various audiences, in this research the following research questions are formulated:

- How to build a deep learning model for generating social manipulation disinformation content to support social media security awareness?
- What are the results obtained based on evaluating the model proposed?

To answer these questions and reach the goal formulated, the Design Science Research methodology is followed in a Data Science approach based on multidisciplinary research (Venable, Pries-Heje & Baskerville, 2017; Maathuis, Pieters & van den Berg, 2018a) considering the design requirements for user-oriented security awareness requirements established by Ki-Aries & Faily (2017). Accordingly, a deep learning model is developed and evaluated through the following research activities (Peffers, 2018):

Problem definition and solution aim: the significant increase in number, diversity, and scale of effects of social manipulation incidents shows that social media platforms, users, and governments are at the beginning of a long path concerning proactive security awareness and dealing with such incidents. Moreover, the ongoing Covid-19

pandemic and war in Ukraine are proofs that such incidents do not only influence users believes but also their behaviour towards the topics involved and (in)direct relational concepts, e.g., natural occurrence of the Covid-19 virus and its realistic effects in relation to other viruses. Hence, given existing efforts that mainly adopt defender's vision, in this research offender's vision is taken for building a deep learning model that generates disinformation content for building/enhancing security awareness. Accordingly, multidisciplinary research is conducted starting with extensive literature review on existing research and practitioner strategies, policies, and methods using different combination of keywords like 'social media', 'manipulation', 'disinformation', 'awareness', 'generation', and 'deep learning' in scientific databased like IEEE Digital Library, ACM Digital Library, Scopus, and Google Scholar.

Solution development: to build the model, data is collected surrounding the Covid-19 pandemic tweets concerning the first two waves, i.e., first wave between 01.01.2020 and 21.03.2020 and second wave between 15.10.2020 and 31.12.2020. Accordingly, the data collection, processing, hashtag counting, topic extraction, model implementation using BERT and GAN deep learning techniques, and evaluation processes are implemented in Python, are compliant with the criteria of Leppanen et al. (2017), and are further discussed. Accordingly, the number of mentions per hashtag is depicted in Figure 1 below:

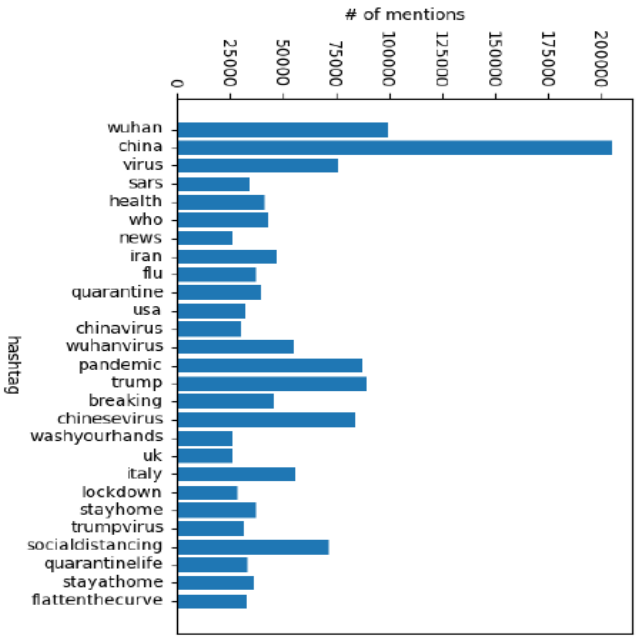

Figure 1: Number of mentions per hashtag

Solution evaluation: the evaluation of the model proposed is carried out continuously through technical and human-based assessment for examining text quality of the content generated and further improving through optimization the solution proposed.

Communication: the approach followed, and the results obtained are herein presented emphasizing the real need for datasets publicly available and arguing for adopting offender's stance when building intelligent solutions for enhancing/building social manipulation security awareness.

4. Modelling Techniques

Nowadays is difficult to point to a field, problem, or system not using or would not benefit from using AI (Maathuis, 2022a). At the core of its current development is Machine Learning which deals with learning from experience through different learning paradigms, e.g., in a predictive approach by mapping input to output given labelled input-output pairs (supervised learning) or in a descriptive approach through knowledge discovery where only the input is provided for learning patterns from data (unsupervised learning) (Murphy, 2012). Due to technological advancements and data availability characterizing, while stepping out of the AI winters, a particular area of ML that amazes through applicability is deep learning (Janiesch, Zschech & Heinrich, 2021). In deep learning are used (complex) neural architectures containing different layers, i.e., input, (one/multiple) hidden, and output where through different operations, the output is computed (Goodfellow, Bengio & Courville, 2016). As this research uses specific deep learning techniques (BERT-Bidirectional Encoder

Representations from Transformers and GAN-Generative Adversarial Networks), they are further discussed.

GAN is a deep learning-based generative technique proposed by Goodfellow et al. (2020) where two distinct neural networks named generator and discriminator built for competing in a game where one should win and the other should lose as in Figure 2. The generator is generating plausible data which should appear real, and the discriminator learns to distinguish between fake data produced by the generator from real data. Hence, the output produced by the generator is fed into the discriminator and is considered a set of negative training data examples. Moreover, the discriminator learns through backpropagation of the loss compared to a fake classification, while the generator learns by comparing the loss to a real classification since it tries to trick the discriminator (Creswell et al., 2018; Aggarwal, Mittal & Battineni, 2021). The training stops when either the generator or the discriminator stops improving, or when the discriminator cannot distinguish fake from real data anymore.

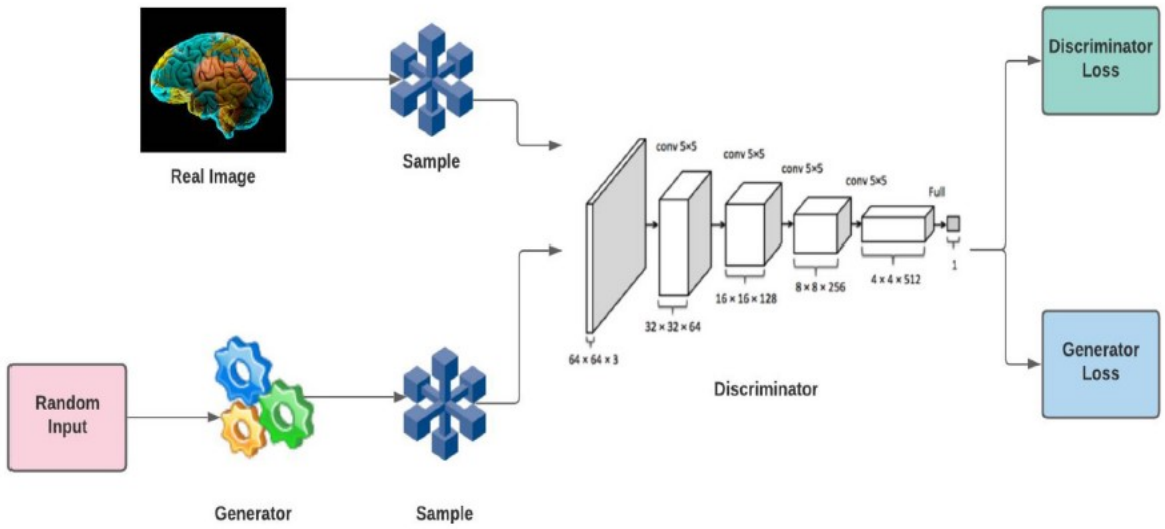

Figure 2. GAN block diagram from (Aggarwal, Mittal & Battineni, 2021)

BERT is a deep learning transformer-based technique proposed by Delvin et al. (2018) at Google for enhancing contextual understanding of unlabelled text across a wide range of activities. BERT is the first deeply bidirectional (left-to-right and right-to-left) unsupervised language architecture containing twelve layers where with each layer more contextual information is learned, see Figure 3. Specifically, the input layer corresponds to tokenized training data and the output layer to the masked language model for which it is trained while aiming to predict the original value of the masked tokens considering the context from both directions (Delvin & Chang, 2018; Qiu et al. 2020) which makes it suitable for text embeddings and text classification tasks.

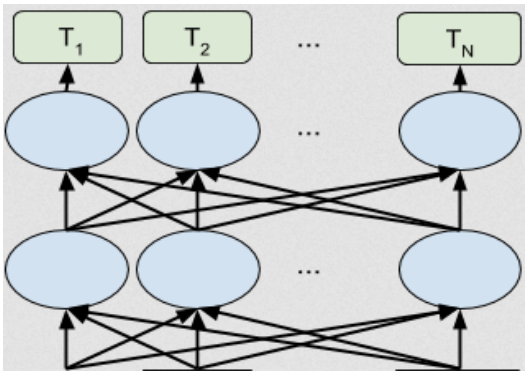

Figure 3: BERT architecture from (Delvin & Chang, 2018)

5. Model Implementation and Results

While datasets useful for tackling disinformation continue to be made available, currently their number is still limited. Hence, for building a proactive approach by expanding the space of available datasets, a merge between GAN and BERT is considered based on previous promising results shown in different domains, e.g., pharmaceutical text classification (Auti et al., 2022), aggressive and violent incidents identification in social

media (Ta et al., 2022), and stock price prediction (Sonkiya, Bajpai & Balsal, 2021). This process is conducted in two steps: generating data without considering whether they imply disinformation or not and generating disinformation data or transforming regular data into disinformation data. While merging these techniques, the hypothesis is that the language understanding of the pretrained BERT model will be transferred to GAN which leads to a better text generation than when only using BERT. Accordingly, the Huggingface transformers library (Wolf et a. 2019) is used, and three experiments are conducted.

In the first experiment, a pre-trained BERT model is used to fill masked words in tweets. When a single word is masked, the model can easily fill the mask in a meaningful way. However, when each word is individually masked with a 15% chance it leads to a loss of context. After further experiments with increasing the mask to different values and finetuning, the model shows reasonable results at 70% masking (looks almost like a real tweet) while from higher than 75% masking, the output deteriorates. Hence, BERT represents a feasible technique for this purpose and a good basis for further research.

Mask %	Output
70	please read the most important thread by @ real _ @ @hfwmington explaining why # socialdistancing is in the name of the the # #securitying the the # coronavirus # covid19 outbreak.
75	# coronavirus # covid19 2 / 3 accelerate the #,,s, therapeutics and and and critical to # to to help all communities. contain misinformation minimize the spread spread.

Figure 4: Results in setting one

In the second experiment, two BERT models are combined into a GAN, see Figure 5. The BertForSequenceClassification model is used for the discriminator and the BertForMaskedLM model for the generator with an extra sigmoid layer added after the last linear layer of the BertForSequenceClassification model. GAN works by generating fake data with the generator which is sent to the discriminator that judges if real or not. Hence, the discriminator is trained with real texts labeled with one and fake text labeled with zero. To allow backpropagation of the loss from the discriminator to the generator, the first BertForSequenceClassification is replaced with a new layer with an extra check if the input is tokenized text, it is embedded as before, else if the input is already embedded, it passes as untouched. Additionally, the last decoder layer of BertForMaskedLM is replaced with an identity layer that passes the input to output without any transformation. Given the existing context, the results obtained are not outperforming the first experimental setting.

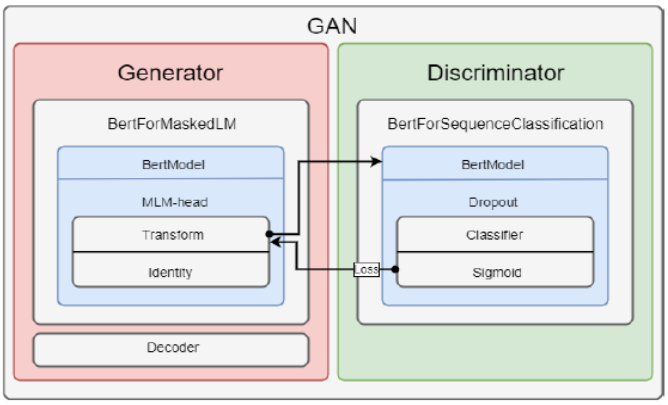

Figure 5: Model architecture setting two

In the third experiment depicted in Figure 6, the sigmoid layer from the first model is removed since BertForSequenceClassification already performs binary classification, the labels are provided in advance and a forward hook is added before the forward pass of the layer. From the experiments conducted, the generator is just as powerful as the discriminator, only slower. Moving this setting to a machine with less memory (6GB instead of 8 GB) shows a direct relation to the fact that although the generator is improving, it is slower than before.

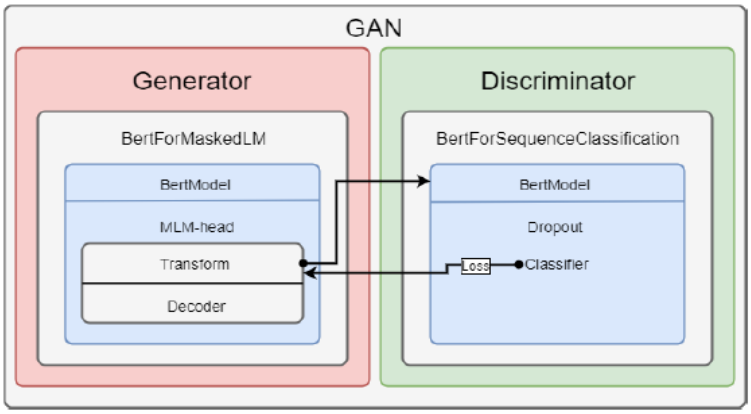

Figure 6: Model architecture setting three

Furthermore, focusing on the third setting, the hyperparameter tunning is applied. This mechanism implies improving the parameters that characterize the learning process, e.g., learning rate, batch size, using optimization techniques for obtaining optimal results. Accordingly, optimizers like SGD (Stochastic Gradient Descent which calculates the gradient of the loss function at random points) and Adam (updates network weights iteratively) are used. Then the model is trained on 10000 randomly selected tweets and validated on 1000 tweets. Hence, the generated tweets are analysed separately, and their metrics are compared with their original considering metrics like diversity for assessing text quality. Herein, while using only one optimizer, Rprop performs the worst, while the others well. Then, different combinations of optimization algorithms are tried which conducts to improved results as in Table 1. To assess which combination is the best, the RMSE (Root Mean Squared Error) of both the losses and the text quality metrics are calculated and shown in Table 2. Hence, the combination with the lowest RMSE for both text quality and the losses are Adam and SGD.

Table 1: Text quality of using combinations of optimizers

Optimizer Gen.	Optimizer Disc.	Word count	Diversity	Stop words	Punctuation count
Adam	AdaDelta	-2.58	-0.14	0.93	1.50
Adam	SGD	-2.36	-0.15	1.33	1.31
AdamW	AdaDelta	-2.23	-0.16	1.32	1.6
AdamW	SGD	-2.41	-0.16	1.49	1.27
RMSprop	AdaDelta	-1.38	-0.16	1.15	2.2
RMSprop	SGD	-1.52	-0.16	1.43	2.42

Table 2: Optimizer combination RMSE

Optimizer Gen.	Optimizer Disc.	Loss RMSE	Text Quality RMSE
Adam	AdaDelta	1.75	0.35
Adam	SGD	1.25	0.30
AdamW	AdaDelta	1.72	0.33
AdamW	SGD	1.25	0.32
RMSprop	AdaDelta	1.83	0.32
RMSprop	SGD	1.29	0.36

Moreover, when ignoring punctuation and stopwords so that they are never masked, the text quality significantly improves having the RSME value of 0.08 as in Table 3 with quite adequate generated tweets depicted in Table 4 and transformation of regular tweets into disinformation tweets shown in Table 5.

Table 3: Text quality after ignoring punctuation and stopwords

Word count	Diversity	Stop words	Punctuation count	Text Quality RMSE	Loss RMSE
-0.20	-0.02	0.03	0.28	0.08	0.90

Table 4: Original and generated tweets

Split	Original	Tweet
50/50	yes	when the #coronavirus first appeared, the natural public response was to demand that the government stop it. the next phase was to blame the government for failing to protect them. the third phase will be attacking the government for taking the steps it took to protect them. - gf
	no	when the # coronavirus outbreak emerged, the publicised reaction was to say that the chinese made it. the chinese reaction was to blame the government for trying to cure them. the real threat will be on the chinese for all the time it takes to cure them. - bsl
55/45	yes	tom hanks and his wife rita confirmed to have contracted #coronavirus in australia?? wow... not sure how reliable the sources are, but if it's true, damn.
	no	jalinhan and his team are supposed to have developed # coronavirus in canada??? ... not sure how real the rumors are, but if it's true, maybe.
60/40	yes	this should be the real worry surrounding the #coronavirus. the economic toll (and the unrest that comes with it) will be far worse than the death toll.
	no	this should be the cover story about the # coronavirus. the virus (and the people that worked with it) will be even worse than the lab lab.

Table 5: Transformation of regular tweets into disinformation tweets

yes	#healthminister: steps taken to prevent spread of #coronavirus in country
no	# coronavirus lab : samples sent to the laboratory of # coronavirus in canada
yes	how can we preserve the valuable networking experience from events as they move online? check out our latest blog post to learn more.
no	how can we keep the latest viruss from china as they come in? check out our own blogging to read more.

Conclusively, both the combination between GAN and BERT as well as BERT itself showed good results in disinformation data generation (hypothesis), and two facts should be considered: (i) with additional relevant data, the solution proposed could be further developed, and (ii) for reducing model's complexity, GAN could be eliminated from the architecture proposed.

6. Conclusions

The first step for building deterrence against social media manipulation mechanisms is security awareness tackling the existence and source plus the impact of akin incidents through corresponding solutions by protecting users' behaviour through cognitive or affective resistance and mitigating mechanisms like labelling (Bastick, 2022; Maathuis, Pieters & van den Berg, 2018b). Whilst such solutions often take defender's stance and are built on a limited set of publicly available social manipulation datasets, attackers improve through adaptivity and autonomy their skills and mechanisms for increasing the spread and intensity of their impact by either reaching quickly specific target audience or scaling it up for hitting a broader range of users when lacking specific target audience. This implies that building corresponding intelligent solutions for tackling social manipulation requires massive publicly available datasets and a multi-angle perspective (Kertysova, 2018; Maathuis, Pieters & van den Berg, 2021). Thus, as these represent major limitations in the existing body of knowledge, this research proposes a deep learning-based social manipulation security awareness approach by building an intelligent model based on GAN and BERT techniques as a disinformation dataset and corresponding generation solution that serves the purpose of strengthening ongoing research and development and reaching multi-domain audiences.

This research continues by combining this solution with a detection approach for building a comprehensive solution for strengthening the space of existing technical solutions dedicated to multi-domain audiences and which implicitly provides the basis for defining values, principles, and requirements that could be used when designing comprehensive social media security awareness solutions, policies, and strategies.

References

Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, *1*(1), 100004.

Auti, T., Sarkar, R., Stearns, B., Ojha, A. K., Paul, A., Comerford, M., ... & McCrae, J. P. (2022). Towards Classification of Legal Pharmaceutical Text using GAN-BERT. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, pp. 52-57.

Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in human behavior*, *116*, 106633.

Chen, L., Chen, J., & Xia, C. (2022). Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, *64*, 103060.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, *35*(1), 53-65.

Chockalingam, S., & Maathuis, C. (2022). An Ontology for Effective Security Incident Management. In *International Conference on Cyber Warfare and Security*. *17(1),* 26-35.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Delvin, J. & Chang, M. W. (2018). Open source BERT: state-of-the art pre-training for Natural Language Processing. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

EU Commission (2018). Communication-Tackling online disinformation: a European approach. https://digital-strategy.ec.europa.eu/en/library/communication-tackling-online-disinformation-european-approach

EU Commission (2022a). Tackling online disinformation. https://digital-strategy.ec.europa.eu/en/policies/online-disinformation.

EU Commission (2022b). The 2022 Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.

Ki-Aries, D., & Faily, S. (2017). Persona-centred information security awareness. *computers & security*, *70*, 663-674.

Kim, J., Aum, J., Lee, S., Jang, Y., Park, E., & Choi, D. (2021). FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics and Informatics*, *64*, 101688.

Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, *29*(1-4), 55-81.

Kula, S., Kozik, R., & Choraś, M. (2021). Implementation of the BERT-derived architectures to tackle disinformation challenges. *Neural Computing and Applications*, 1-13.

Le, T., Wang, S., & Lee, D. (2020). Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, 282-291, IEEE.

Leppänen, L., Munezero, M., Granroth-Wilding, M., & Toivonen, H. (2017). Data-driven news generation for automated journalism. In *Proceedings of the 10th international conference on natural language generation*. 188-197.

Maathuis, C., Pieters, W., & Van Den Berg, J. (2016). Cyber weapons: a profiling framework. In *2016 International Conference on Cyber Conflict (CyCon US)*, pp. 1-8, IEEE.

Maathuis, C., Pieters, W., & van den Berg, J. (2018a). Developing a cyber operations computational ontology. *Journal of Information Warfare*, *17*(3), 32-49.

Maathuis, C., Pieters, W., & Van den Berg, J. (2018b). Assessment methodology for collateral damage and military (Dis) Advantage in cyber operations. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, pp. 1-6, IEEE.

Maathuis, C., Pieters, W., & van den Berg, J. (2021). Decision support model for effects estimation and proportionality assessment for targeting in cyber operations. *Defence Technology*, *17*(2), 352-374.

Maathuis, C. (2022a). On Explainable AI Solutions for Targeting in Cyber Military Operations. In *International Conference on Cyber Warfare and Security*. 17(1), 166-175.

Maathuis, C. (2022b). On the Road to Designing Responsible AI Systems in Military Cyber Operations. In *European Conference on Cyber Warfare and Security*. 21(1), 170-177.

Maathuis, C., & Chockalingam, S. (2022b). Responsible Digital Security Behaviour: Definition and Assessment Model. In *European Conference on Cyber Warfare and Security*. 21(1).

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

NATO (2021a). Countering disinformation: improving the Alliance's digital resilience. https://www.nato.int/docu/review/articles/2021/08/12/countering-disinformation-improving-the-alliances-digital-resilience/index.html

NATO (2021b). How does NATO respond to disinformation? https://www.nato.int/cps/en/natohq/news_184036.htm

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... & Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuition*, 21-29, Springer, Cham.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872-1897.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Pérez-Escolar, M., Ordóñez-Olmedo, E., & Alcaide-Pulido, P. (2021). Fact-Checking Skills and Project-Based Learning about Infodemic and Disinformation. *Thinking Skills and Creativity*, *41*, 100887.

Rathenau (2022). AI and manipulation on social and digital media. https://www.rathenau.nl/en/digital-governance/ai-and-manipulation-social-and-digital-media

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European journal of operational research*, *291*(3), 906-917.

Shu, K. (2022). Combating disinformation on social media: A computational perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 100035.

Shu, K., Li, Y., Ding, K., & Liu, H. (2021). Fact-enhanced synthetic news generation. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 35(15), 13825-13833.

Sonkiya, P., Bajpai, V., & Bansal, A. (2021). Stock price prediction using BERT and GAN. *arXiv preprint arXiv:2107.09055*.

Ta, H. T., Rahman, A. B. S., Najjar, L., & Gelbukh, A. (2022). GAN-BERT: Adversarial Learning for Detection of Aggressive and Violent Incidents from Social Media.

Venable, J. R., Pries-Heje, J., & Baskerville, R. L. (2017). Choosing a design science research methodology.

Wang, X., Guo, H., Hu, S., Chang, M. C., & Lyu, S. (2022). Gan-generated faces detection: A survey and new perspectives. *arXiv preprint arXiv:2202.07145*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yu, G., Yang, Y., Wang, X., Zhen, H., He, G., Li, Z., ... & Shu, L. (2020). Adversarial active learning for the identification of medical concepts and annotation inconsistency. *Journal of Biomedical Informatics*, *108*, 103481.