Design Lessons from Building Deep Learning Disinformation Generation and Detection Solutions

Clara Maathuis[1], Iddo Kerkhof[2], Rik Godschalk[3] and Harrie Passier[4]
[1,4] Open University, Netherlands
[2,3] Independent Researcher, Netherlands

clara.maathuis@ou.nl
email@iddo.eu
ahgodschalk@gmail.com
harrie.passier@ou.nl

Abstract: In its essence, social media is on its way of representing the superposition of all digital representations of human concepts, ideas, believes, attitudes, and experiences. In this realm, the information is not only shared, but also {mis, dis}interpreted either unintentionally or intentionally guided by (some kind of) awareness, uncertainty, or offensive purposes. This can produce implications and consequences such as societal and political polarization, and influence or alter human behaviour and beliefs. To tackle these issues corresponding to social media manipulation mechanisms like disinformation and misinformation, a diverse palette of efforts represented by governmental and social media platforms strategies, policies, and methods plus academic and independent studies and solutions are proposed. However, such solutions are based on a technical standpoint mainly on gaming or AI-based techniques and technologies, but often only consider the defender's perspective and address in a limited way the social perspective of this phenomenon becoming single angled. To address these issues, this research combines the defenders' perspective with the one of the offenders by (i) building a hybrid deep learning disinformation generation and detection model and (ii) capturing and proposing a set of design recommendations that could be considered when establishing patterns, requirements, and features for building future gaming and AI-based solutions for combating social media manipulation mechanisms. This is done using the Design Science Research methodology in Data Science approach aiming at enhancing security awareness and resilience against social media manipulation.

Keywords: social media manipulation, disinformation, misinformation, security awareness, machine learning, deep learning.

1. Introduction

"*We live in a fantasy world, a world of illusion. The great task in life is to find reality*." (Iris Murdoch)

Through their capability of producing confusion, deception, and polarization by influencing human beliefs and behaviour, social media manipulation mechanisms like disinformation and misinformation were recently positioned among the most impactful cyber threats (Caramancion et al., 2022; ENISA, 2022). Such mechanisms have attested records from millennia's ago in the disinformation campaigns conducted in Egypt surrounding the Battle of Kadesh where a wrong location was given to Pharaoh Ramses II, to more recent collection of articles published by The Sun named the Great Moon Hoax in 1835 about the life and culture found on the Moon, and during the World Wars regarding the demonization of parties involved and propaganda (Choras et al., 2021). Due to technological advancements, such mechanisms evolved rapidly reaching global audiences while their counterparts also advance in building combating solutions based on content/data, (creative and participatory) engagement, and a broad range of software and AI-based solutions (Reisach, 2021; Maathuis, 2022b). Nevertheless, social media manipulation is an online phenomenon with offline impact (Shu, 2022). Corresponding combat efforts usually take the defenders' perspective, but a paradigm switch and change of approach to considering attackers' perspective is necessary, but scarcely present in the literature and practitioner efforts.

Hence, this research aims to build a hybrid deep-learning model for generating and detecting disinformation based on previous work, and from there to propose a set of design recommendations that can serve as design patterns, requirements, and features for building responsible software (e.g., gaming) and AI-based solutions for producing or enhancing awareness, resilience, and education on social media manipulation. To achieve this aim, a multidisciplinary stance is taken following the Design Science Research in a Data Science methodological approach (Peffers et al., 2007; Peffers, Tuunanen & Niehaves, 2018; Foroughi & Luksch, 2018) having the following contributions:

- From an implementation perspective, it proposes a hybrid intelligent solution that merges the perspective of the attackers with the defenders' one using advanced deep learning techniques emphasizing the possibility of combining real with synthetic content and obtaining relevant results.

- From a design perspective, it proposes a set of recommendations with socio-technical dimensions that could be integrated when building software-based or intelligent solutions for combating social media manipulation.

This article advances as follows. Section 2 settles the background and presents relevant studies to this research. Section 3 addresses the research methodology pursued to achieve the aim of this research. Section 4 summarizes the deep learning techniques used to build the hybrid model proposed. Section 5 presents important implementation aspects and the evaluation results obtained. Section 6 addresses the recommendations proposed. At the end, Section 7 presents concluding remarks and future ideas.

2. Background and Related Research

In its last report, ENISA (European Union Agency for Cybersecurity) positions social media manipulation among the top eight cyber threats of 2021 and 2022 (ENISA, 2022). To deal with it, Reisach (2021) considers that all agents involved, i.e., social media platforms, organizations, users, society, share responsibility. When aiming at limiting/combating social media manipulation, it is necessary to realize the importance of awareness and education. Hence, corresponding solutions that build a resilient and self-aware social media literacy and mindset should be proposed. Moreover, Interpol (2020) argues for building accurate and trustworthy AI solutions for tackling mis/disinformation mechanisms while accounting existing policy, legal, and societal challenges (Europol, 2022). As a basis for building such solutions whilst protecting human and democratic values and systems, limiting public harm and damage on policy-making processes while contributing to building users' digital security and resilience, the EU Commission proposed initiatives like European Democracy Action Plan and the Strengthened Code of Practice on Disinformation (EU Commission, 2018; EU Commission, 2022a; EU Commission, 2022b).

Thakur, Hayajneh & Tseng (2019) analyze cyber threats in social media together with possible ways to counter them and limit the damage. Among these threats are social engineering attacks and lack of social media policies. Social engineering attacks are based on eliciting individuals' personal information using fake accounts and building trust. Lack of social media policies for organizations points to guidelines that should be deployed for safe social media use by its employees based on proper education and training plus the establishment of a social media coordinator, champion, or manager. Shu et al. (2020) discuss differences and connections between different manipulation mechanisms and emphasize the importance of user engagement, proper design of detecting and mitigating solutions, and the inclusion of socio-technical aspects and technological paradigms like ethics and blockchain, respectively. Moreover, Zhou & Zafarani (2020) conduct a systematic literature review study on fake news theories, detection methods, and opportunities for detection purposes. The authors categorize the studies as follows: the fake knowledge carried, writing style, propagation patterns, and credibility of the source. On this behalf, the importance of {multi, inter, trans}disciplinary research and collaboration between experts from different disciplines which would conduct to explainable results, is highlighted. Choaras et al. (2021) review advanced Machine Learning applications using existing datasets concentrating on scientific and R&D societal projects. The results show the importance of understanding the context, having a clear methodological approach, and being able to explain both the modelling reasoning and the results obtained. Additionally, Pinhanez et al. (2022) argue that disinformation, and particularly, fake news and deep fakes, are an important upcoming cyber security problem that could be tackled building a new way of understanding disinformation by merging generation and detection perspectives while actively engaging users in prevention and treatment processes. On this, Rogers & Niederer (2020) analyze political dimensions of politically deployed social media manipulation and identifies two major waves when dealing with this. The first regarding understanding, i.e., definitional issues and production, and the second dedicated to its consumption which includes persuasion. In both waves, AI-based efforts for detecting and dealing with such mechanisms exist. Alassad, Spann & Agarwal (2021) propose an integrated method for identifying malicious behaviour and actors responsible for propagation based on advanced computational social science methods and graph theoretic techniques. The authors point the importance of user versus group analysis for detecting and investigating malicious intent and provide further decision-making support for applying proper response strategies. Moreover, Reisach (2021) advances a responsibility-based approach for social media platforms with a keen eye on political contexts based on existing socio-ethical principles and requirements developed for safeguarding democratic societies through human rights consideration, education, and transparency of algorithmic decisions. Kertysova (2018) investigates challenges and opportunities posed by AI in the context of producing, disseminating, and countering disinformation. Particularly, the author stresses the importance of understanding and dealing with both human-based and algorithmic-based biases when dealing with disinformation campaigns/operations, e.g., cultural, racial, gender, personality, and political, which improperly treated could

imply disparate impact on certain groups of people. In respect to these dimensions, Maathuis & Chockalingam (2022a) propose a model that incorporates human dimensions for assessing if the digital security behaviour of users could be classified as responsible or not.

Furhermore, Nasir, Khan & Varlamis (2021) propose a hybrid deep learning based model combining CNN (Convolutional Neural Network) with RNN (Recurrent Neural Network) techniques for fake news detection showing promising generalizing results. Stiff & Johansson (2022) propose a RoBERTa-based disinformation detection solution which is robust on both in-distribution and out-of-distribution test data plus realistic in-the-wild data. Kaliyar, Goswami & Narang (2021) build a hybrid deep learning fake news detection model combining BERT (Bidirectional Encoder Representations from Transformers) with CNN techniques for capturing information relevance and reducing ambiguity. Given the approach pursued and results obtained (accuracy of 98.99%), in this research a similar approach is considered for the detection component of the solution proposed. Verma et al. (2022) build the MCred (Message Credibility) framework based on CNN and BERT techniques for disinformation detection through information credibility assessment using the benefits of local and global text semantics. Du, Bosselut & Manning (2022) build a deep learning-based verification framework for sensitivity evaluation of deep learning-based techniques, i.e., KGAT, CorefBERT, and MLA to adversarial generated disinformation done using GROVER. The results obtained show a high sensitivity to automated generated disinformation on different topics. This implies that improvements in automated synthetic content detection is necessary and should be combined with human audits of fact-checker evidence, thus a joint between algorithm-based modelling and expert-based approach. Shu, Li, Ding & Liu (2021) propose a solution for news generation based on FACTGEN (Fact-Enhanced Synthetic News Generation) method that is implementing a PSA (Pseudo-Self-Attentive) model for tackling challenges related to information encountered when generating news on a given claim. Furthermore, it is important to acknowledge the importance of security and privacy threats to deep learning models applied in this domain. Le, Wang & Lee (2020) build a GAN (Generative Adversarial Network) solution for generating malicious content for attacking fake news detection models. Bradshaw & Howard (2018) conduct an extensive literature review, consult field experts, and use content analysis solutions on social media manipulation incidents conducted in 48 countries on Facebook, Twitter, WhatsApp, Telegram, and WeChat. While evidence ranges from highly organized, specialized, and skilled political and/or military cyber troops to simple users' spread of social manipulation, the study emphasize the role of data, communication, behavior change, policy and strategy definition, and collective engagement to enhance trust and security. Specifically, Patwa et al. (2021) propose a labelled dataset with 10.700 tweets containing fake news. The dataset is used for building a series of classification models from which SVM showed the best performance (F1-score is 93.32%). Moreover, Guo & Vosoughi (2022) build a large-scale dataset that contains sub-datasets with tweets and accounts affiliated with 14 countries spanning more than 3 years useful for tasks like analyzing state-sponsored propaganda.

The summary of the academic studies related to this research is provided in Table 1. As the overview suggests, in general, a multidisciplinary perspective is considered in this domain and most of the studies are either qualitative or quantitative, with a small proportion of mixed studies. Furthermore, these studies can be clustered either on detection, generation for detection or resilience to security attacks, or a combination thereof. Taking into consideration the fact that the second and third categories are less represented in the body of knowledge compared to the first category, this research positions itself in the third category by aiming to propose a solution that merges disinformation generation with detection to increase security awareness in social media using a hybrid deep learning approach.

Table 1. Overview of solutions proposed in related studies

No	Type	Method	Approach	Source
1	Analysis	Literature review	Qualitative	Thakur, Hayajneh & Tseng, (2019)
2	Analysis	Literature review	Qualitative	Shu et al. (2020)
3	Detection	Systematic literature review	Qualitative	Zhou & Zafarani (2020)
4	Detection	Systematic literature review	Qualitative	Choras et al. (2021)
5	Generation and Detection	Literature review	Qualitative	Pinhanez et al. (2022)
6	Analysis	Literature review	Qualitative	Rogers & Niederer (2020)
7	User vs group analysis	Computational social science and graph theory	Quantitative	Alassad, Spann & Agarwal (2021)
8	Responsibility based approach	Ethical principles and interdisciplinarity	Qualitative	Reisach (2021)

No	Type	Method	Approach	Source
9	Analysis	Literature review	Qualitative	Kertysova, K. (2018)
10	Responsibility assessment	Bayesian model	Mixed	Maathuis & Chockalingam (2022a)
11	Detection	CNN and RNN	Mixed	Nasir et al. (2021)
12	Detection	RoBERTa	Quantitative	Stiff & Johansson (2022)
13	Detection	BERT and CNN	Quantitative	Kaliyar, Goswami & Narang (2021)
14	Detection	BERT and CNN	Quantitative	Verma et al (2022)
15	Generation	GROVER, KGAT, CorefBERT, and MLA models	Quantitative	Du, Bosselut & Manning (2022)
16	Generation	FACTGEN and PSA	Quantitative	Shu et al (2021)
17	Generation of security attacks	GAN	Quantitative	Le, Wang & Lee (2020)
18	Incident analysis	Literature review, experts' interviews, and content analysis	Mixed	Bradshaw & Howard (2018)
19	Dataset	Dataset collection, preparation, and modelling	Quantitative	Patwa et al. (2021)
20	Large-scale dataset	Dataset collection	Qualitative	Guo & Vosoughi (2022)

3. Research Methodology

This research aims to build a hybrid deep-learning model for generating and detecting disinformation and proposing a set of recommendations for designing social media manipulation combating solutions. Accordingly, the following research questions are formulated:

- How to build a hybrid deep learning model for disinformation generation and detection?
- What are the lessons learned from building the proposed model?
- What recommendations can be drawn for further designing security awareness solutions for combating social media manipulation?

To achieve this goal, a multidisciplinary viewpoint is considered pursuing a Design Science Research in a Data Science methodological approach (Peffers et al., 2007; Venable, Pries-Heje & Baskerville, 2017; Peffers, Tuunanen & Niehaves, 2018; Foroughi & Luksch, 2018). The phases of this research are further described.

Phase I (Definition and solution aim): the number of social media cyber security incidents grounded on threats like disinformation and misinformation continues to increase especially in conjunction with conflicts like the one in Ukraine. While the impact of such incidents is difficult to predict/contain (Chockalingam & Maathuis, (2022), and different mechanisms are developed to prevent/combat them, they often take the defenders' perspective and a detection-based approach. Since it is agreed that the offender has an advantage in attack, a mindset change is necessary to build combating solutions that adopt/include the offenders' perspective by relying not only on detection but combining it with a deep learning-based generation approach, and from there capturing recommendations for designing combating solutions. Hence, the goal and the techniques to be used in this sense are established.

Phase II (Solution development): to build the model proposed, tweets concerning the first two Covid-19 waves are collected, thus for the first wave between 01.01.2020 and 21.03.2020 and second wave between 15.10.2020 and 31.12.2020. The data is processed, topics are extracted, and the generation solution is developed combining BERT with GAN while the detection solution is developed combining BERT with CNN in Python, as proposed in (Maathuis & Kerkhof, 2023; Maathuis & Godschalk, 2023). These solutions are merged in this research. Furthermore, to capture recommendations relevant when designing social media manipulation combating solutions, literature review is conducted using combinations of keywords like 'social media', 'disinformation', 'misinformation', 'detection', 'generation', and 'deep learning' for querying scientific databases like IEEE, ACM, Scopus, and Google Scholar.

Phase III (Solution evaluation): the solution proposed is assessed through technical evaluation using performance metrics and human-based assessment regarding the quality of the results obtained. Moreover, the design recommendations proposed are subject to incorporation, extension, and evaluation in future research.

Phase IV (Communication): the results are communicated through the present article and by means of presentations to relevant communities.

4. Deep Learning Modelling Techniques

Because of ongoing social, economic, and technological developments directly reflected in different directions of digital innovation, the increase and availability of high computing resources and multi-source data support the expansion of AI at theoretical and practical levels (Janiesch, Zschech & Heinrich, 2021; Maathuis, 2022a). Currently, the most innovative sub-field of AI is deep learning which implies in the context of social media building (complex) neural architectures relying on (large) datasets having analytical or predictive purposes for tackling existing or upcoming issues and challenges of this domain. In this research, a solution based on a hybrid deep learning approach is proposed by merging CNN and GAN with BERT techniques.

In a GAN architecture (Goodfellow et al., 2020), two different neural networks called generator and discriminator are competing in a game where the first component is generating data that should look plausible and where the second component must distinguish between the data produced by the generator and real data. In this architecture, the data produced by the generator is fed as input to the discriminator, and the training stops when the discriminator cannot distinguish fake from real data anymore or no improvement can be made. In a CNN architecture (LeCun et al., 1989), features are extracted through convolutional operations carried out in a set of different types of layers: input, convolution, pooling, fully connected, and output, and the learning process is done using backpropagation (Li et al., 2021). In a BERT architecture (Delvin et al., 2018), contextual information is acquired through a set of twelve layers through in a bidirectional perspective (left-to-right and right-to-left). Herein, the first layer is the tokenized training data and the output layer the masked language model containing the context from both directions (Qiu et al., 2020).

5. Implementation and Results

As discussed in Section 2, combining deep learning techniques in hybrid neural architectures showed good results when dealing/tackling social media manipulation. Hence, to address the first research question of this research a hybrid deep learning model that contains two components is developed: (i) a disinformation generation model implemented using GAN and BERT with the hypothesis that the language understanding capacity of the pretrained BERT is transferred to GAN which leads to a proper text generation, and (ii) a disinformation detection model implemented using CNN and BERT with an additional feature captured through sentiment analysis hypnotizing that adding a sentiment analysis feature increases the performance of the model.

The first model is the disinformation generation model which contains two BERT models combined into a GAN as depicted in Figure 1. The BertForMaskedLM is used for the generator, the BertForSequenceClassification is used for the discriminator, and a forward hook is added before the forward pass of the layer. Based on the experiments conducted, both components are effective, but the generator is slower. Furthermore, two optimization algorithms are applied: SGD (Stochastic Gradient Descent) and Adam. Then the model is trained on 10000 tweets, tested on 1000 tweets, and the RMSE (Root Mean Squared Error) of the losses and text quality metrics are computed. Hence, the results obtained after removing the punctuation and stopwords are shown in Table 2 and 3. From the results obtained, the hypothesis formulated is confirmed.

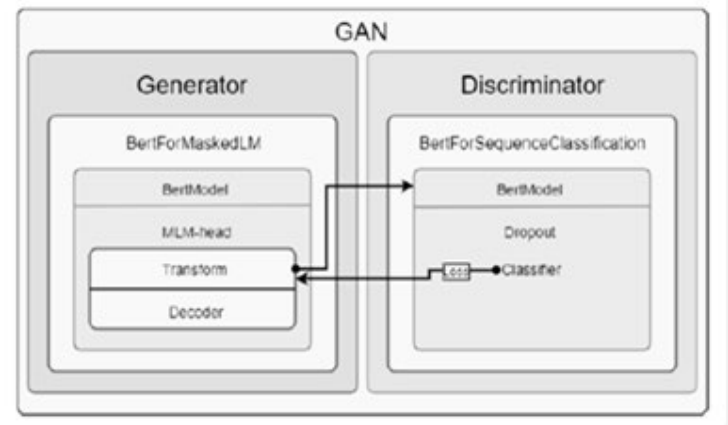

Figure 1. Disinformation generation model architecture

Table 2. Test quality results

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 3. Transformation of regular tweets into disinformation tweets

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.

The second model is the disinformation detection model which combines CNN with or without a sentiment analysis feature with BERT as illustrated in Figure 2 for classifying tweets received as input as being disinformation or not. The CNN model has a ReLu (Rectified linear units) activation function and is added on top of the BERT layers. Furthermore, the model is trained during a set of epochs which represent smaller steps that depend on the batch size and the size of the training set. In each step, data from a batch is passed through the network which gives a prediction that is compared to the actual class for calculating the loss which is used to tune the weights of the network. The model is further tested with and without the feature with sentiment values and the results obtained are presented in Figure 3 and 4. From the results obtained, the hypothesis is confirmed.

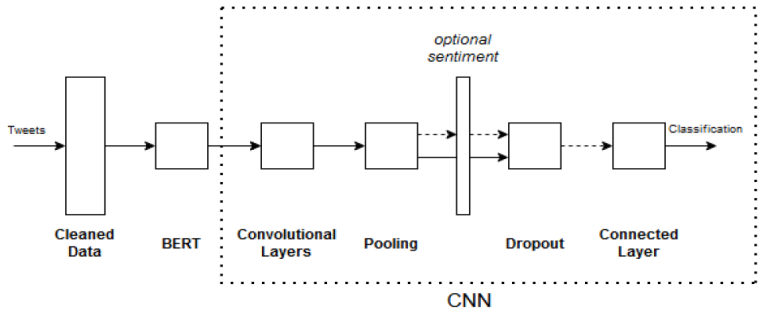

Figure 2. Disinformation detection model architecture

Type	Loss	Accuracy	Precision	Recall	F1
Validation #1	0.165	0.958	0.957	0.958	0.951
Test #1	0.108	0.979	0.963	0.979	0.970
Validation #2	0.076	0.969	0.968	0.969	0.964
Test #2	0.099	0.969	0.977	0.969	0.965
Validation #3	0.141	0.967	0.975	0.969	0.967
Test #3	0.088	0.979	0.971	0.979	0.973

Type	Loss	Accuracy	Precision	Recall	F1
Validation #1	0.050	0.990	0.981	0.990	0.985
Test #1	0.080	0.979	0.981	0.979	0.971
Validation #2	0.036	0.979	0.990	0.979	0.998
Test #2	0.080	0.979	0.963	0.979	0.970
Validation #3	0.07	0.989	0.980	0.990	0.984
Test #3	0.154	0.958	0.972	0.958	0.962

Figure 3. Test results: with sentiment feature (left) and without sentiment feature (right)

Tweet
Leak originated from a bio weapons lab in Wuhan which I believe is in China.
The truth behind delayed and broken #coronavirus CDC testkits tied to contamination due to poor mishandling in Atlanta lab, says federal scientist.

Figure 4. Disinformation tweets

Furthermore, to evaluate the system proposed and analyse its performance, the two models presented above are merged. Accordingly, this process is executed in two settings: in the first setting, the detection model receives half of the data from the generation model (1000) and half of the data being real tweets (1000); in the second setting, the detection model receives data only from the generation model (2000).

For the first setting, to determine the number of epochs needed to train the model and prevent it from overfitting, the model is trained first on many epochs. While the training loss keeps decreasing every epoch, the validation loss stabilizes or becomes worse after a certain point. This is the ideal point for the model to stop training and further make sure that overfitting does not occur. Accordingly, the number of epochs is five, and the results obtained are depicted in Figure 5. These results show a high accuracy of 0.97 which implies that the model performs well in separating and detecting between fake and real messages, thus the first setting represents a viable system for combating disinformation.

Loss	Accuracy	Precision	Recall	F1
0.12	0.97	0.98	0.97	0.97

Figure 5. First setting evaluation results

For the second setting, a balanced set of 2000 messages is used, thus 1000 labelled as disinformation and 1000 as non-disinformation. The model correctly classified 899 out of 1000 disinformation messages as being disinformation and falsely classified 50 out of 1000 non-disinformation messages as being disinformation. The results are depicted in Figure 6. These results show a high accuracy of 0.92, thus less than the first setting, but it can still be considered as a viable setting for combating disinformation.

Loss	Accuracy	Precision	Recall	F1
0.36	0.92	0.94	0.92	0.92

Figure 5. Second setting evaluation results

6. Recommendations

Based on the literature review conducted, the lessons learned from building the hybrid deep learning solution proposed in this research and considering the need for adopting a socio-technical approach when dealing or combating social media manipulation, a series of design recommendations are gathered and structured in Table 4. These recommendations are meant to represent design patterns, requirements, and features for building reliable, transparent, accountable, and enjoyable software based (e.g., gaming solutions) or intelligent systems (e.g., AI modelling and simulation).

Table 4. Design recommendations overview

Design Recommendation	Description Recommendation	Source Recommendation
1	Consider a proactive adversary/offensive thinking mindset having an ethical conduct. This implies building solutions that do not consider only defenders' perspective, but they mix it with the one of the offenders taking into consideration and inclusion of social and ethical norms and values.	Bradshaw & Howard (2018); Pinhanez et al. (2022); Maathuis & Chockalingam (2022a); Du, Bosselut & Manning (2022)
2	Consider a multidisciplinary approach when designing social media manipulation countering solutions by merging social media analytics, cyber/digital security, behavioural science, and ethical principles, methods, techniques, and technologies. This relates to conducting literature review on both academic and practitioner resources in the domains involved taking a socio-technical stance and involving experts from these domains in all the phases of the life cycle of the solutions proposed in order to reflect their trustworthiness.	Kertysova, K. (2018); Bradshaw & Howard (2018); Thakur, Hayajneh & Tseng, (2019); Le, Wang & Lee (2020); Reisach (2021); Maathuis & Chockalingam (2022a); Pinhanez et al. (2022)
3	Think, draw, and propose offensive/malicious scenarios that would facilitate a critical thinking approach to design, develop, and deploy effective and efficient social media manipulation countering solutions. This implies combining the data uses for building and evaluating the solutions proposed from offensive scenarios and incidents from both real and synthetical data.	Bradshaw & Howard (2018); Shu et al (2021); Du, Bosselut & Manning (2022); Pinhanez et al. (2022)
4	Consider using multi-source data by combining with synthetical data and expert-based analysis and advise in both generation and detection-based solutions. Additionally, build and make publicly available datasets for research purposes. This implies considering the use of the at least from all the domains involved when building and evaluating the solutions proposed and further support the transparency and accountability which are such important values in this domain.	Bradshaw & Howard (2018); Choras et al. (2021); Patwa et al. (2021); Alassad, Spann & Agarwal (2021); Guo & Vosoughi (2022); Pinhanez et al. (2022); Du, Bosselut & Manning (2022)
5	Consider a mixed approach, i.e., merge qualitative with quantitative research and development methods. This implies benefiting from the advantages characterizing both qualitative and quantitative approaches when designing, developing, deploying, and using the solutions proposed.	Bradshaw & Howard (2018); Alassad, Spann & Agarwal (2021); Reisach (2021); Nasir et al. (2021); Maathuis & Chockalingam (2022a)
6	Consider a human centric and trustworthy approach while reducing human and algorithm biases. This implies and can be reduced to building solutions that support the are created and used in the benefit of humanity.	Bradshaw & Howard (2018); Le, Wang & Lee (2020); Alassad, Spann & Agarwal (2021); Reisach (2021); Maathuis & Chockalingam (2022a)

7. Discussion and Conclusions

The efforts of social media platforms, specialized organizations, practitioners, and academic researchers in understanding the motivation, strategies, techniques, and technologies behind social media manipulation used by attackers continue to increase and materialize concerning several spatial and temporal dimensions through developing and deploying tailored, advanced, and intelligent containing and/or combating solutions. Nevertheless, they often rely on a reactive and passive approach (Shu, 2022) implying detecting specific behaviour or directly triggering corresponding events. Regardless, the effectiveness of this approach could be supplemented by another one due to the dynamism and uncertainty surrounding the content (e.g., authenticity and reliability), the users (e.g., cognitive, and emotional biases), but also the solutions deployed (e.g., logical fallacies and insufficient data), as the social media content surrounding the Covid-19 pandemic and the war in Ukraine shows. A shift in this direction is adopting a human-centred approach wearing attacker's glasses and investing in designing and developing awareness, resilience, and educational solutions (Maathuis, Pieters & van den Berg, 2018a; Reisach, 2021).

Moreover, while an upcoming plethora of AI-based generation or mixing generation with detection solutions for combating social media manipulation exist, to the best of our knowledge a limited number of studies merge these approaches with literature review and existing expertise to capture and propose lessons that could be further considered as requirements or patterns for building future social media manipulation preventing and/or combating solutions. Therefore, it is the aim of this research to build a solution that merges the defenders' perspectives with the one of the attackers and propose a set of design recommendations for building further social media manipulation combating solutions. On this behalf, three research questions are formulated and addressed considering a Design Science Research applied in Data Science approach as follows. Firstly, a hybrid deep learning based generation and detection solution is proposed consisting of two models: the generation one implemented using BERT and GAN, and the detection one implemented using BERT and CNN with sentiment analysis. Secondly, a set of design recommendations are captured and presented as design considerations for building future gaming and/or AI-based solutions developed for preventing and/or combating social media manipulation.

This research advances in two main directions. The first one by addressing other disinformation topics that are related to ongoing conflicts and from there gathering socio-technical supportive design lessons. And the second one by using the proposed design recommendations and further extending them to another social media manipulation mechanism, i.e., deep fakes, for further building awareness, resilience, and education solutions since one should not forget that not only computations and computable systems are hackable, but humans too.

References

Alassad, M., Spann, B., & Agarwal, N. (2021). Combining advanced computational social science and graph theoretic techniques to reveal adversarial information operations. Information Processing & Management, 58(1), 102385.

Bradshaw, S., & Howard, P. N. (2018). Challenging truth and trust: A global inventory of organized social media manipulation. The computational propaganda *project*, *1*, 1-26.

Caramancion, K. M., Li, Y., Dubois, E., & Jung, E. S. (2022). The Missing Case of Disinformation from the Cybersecurity Risk Continuum: A Comparative Assessment of Disinformation with Other Cyber Threats. *Data*, *7*(4), 49)

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

Choraś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., ... & Woźniak, M. (2021). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing*, *101*, 107050.

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*.

Du, Y., Bosselut, A., & Manning, C. D. (2022). Synthetic Disinformation Attacks on Automated Fact Verification Systems. *arXiv preprint arXiv:2202.09381*.

ENISA (2022). ENISA threat landscape 2022. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

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)

Europol (2022), Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg.

Foroughi, F., & Luksch, P. (2018). Data science methodology for cybersecurity projects. *arXiv preprint arXiv:1803.04219*.
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.
Guo, X., & Vosoughi, S. (2022). A Large-Scale Longitudinal Multimodal Dataset of State-Backed Information Operations on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 1245-1250.
Interpol & UNICRI (2020). Towards Responsible AI Innovation. Second Interpol-UNICRI Report on AI for Law Enforcement.
Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.
Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, *80*(8), 11765-11788.
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.
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.
LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*.
Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.
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. (2022a). 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. (2022b). An Outlook of Digital Twins in Offensive Military Cyber Operations. In *European Conference on the Impact of Artificial Intelligence and Robotics*, 4(1), 45-53.
Maathuis, C. & Chockalingam, S. (2022a). Responsible Digital Security Behaviour: Definition and Assessment Model. In *European Conference on Cyber Warfare and Security*, 21(1).
Maathuis, C. & Godschalk, R. (2023). Social Media Manipulation Deep Learning based Disinformation Detection. In *International Conference on Cyber Warfare and Security*.
Maathuis, C. & Kerkhof, I. (2023). Social Media Manipulation Awareness through Deep Learning based Disinformation Generation. In *International Conference on Cyber Warfare and Security*.
Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007.
Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... & Chakraborty, T. (2021, February). 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.
Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.
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.
Pinhanez, C. S., Flores, G. H., Vasconcelos, M. A., Qiao, M., Linck, N., de Paula, R., & Ong, Y. J. (2022). Towards a New Science of Disinformation. *arXiv preprint arXiv:2204.01489*.
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.
Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European journal of operational research*, *291*(3), 906-917.
Rogers, R., & Niederer, S. (2020). *The politics of social media manipulation* (p. 257). Amsterdam University Press.
Shu, K. (2022). Combating disinformation on social media: A computational perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 100035.
Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. In *Disinformation, misinformation, and fake news in social media* (pp. 1-19). Springer, Cham.
Stiff, H., & Johansson, F. (2022). Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, *13*(4), 363-383.
Thakur, K., Hayajneh, T., & Tseng, J. (2019). Cyber security in social media: challenges and the way forward. *IT Professional*, *21*(2), 41-49.
Venable, J. R., Pries-Heje, J., & Baskerville, R. L. (2017). Choosing a design science research methodology.
Verma, P. K., Agrawal, P., Madaan, V., & Prodan, R. (2022). MCred: multi-modal message credibility for fake news detection using BERT and CNN. *Journal of Ambient Intelligence and Humanized Computing*, 1-13.
Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1-40.