Social Media Manipulation Deep Learning based Disinformation Detection

Clara Maathuis[1], Rik Godschalk[2]
[1]Open University of the Netherlands, Heerlen, Netherlands.
[2]Independent Researcher, Netherlands.

clara.maathuis@ou.nl
ahgodschalk@gmail.com

Abstract: The rapid grow and use of different social platforms enhanced communication between different entities and their audiences plus the transformation through digitalization of existing, e.g., ideas and businesses, or the creation of new ones fully existing or depending on this digital environment. Nevertheless, next to these promising aspects, social media is a vulnerable digital environment where a diverse plethora of cyber incidents are planned and executed engaging a diverse range of targets. Among these, social media manipulation through threats like disinformation and misinformation produce a broad span of effects that cross digital borders into the human realm by influencing and altering human believes, behaviour, and attitudes towards specific ideas, institutions, or people. To tackle these issues, existing academic, social platforms, dedicated organizations, and institutions efforts exist for building specific advanced and intelligent solutions for detecting and preventing them. Regardless, these efforts embed defender's perspective and are focused locally, at target level, without being designed to fit a broader agenda of producing and/or strengthening social media security awareness. On this behalf, this research proposes a deep learning-based disinformation detection solution for facilitating and/or enhancing social media security awareness in respect to offender's perspective. To achieve this objective, a Data Science approach is taken based on the Design Science Research methodology, and the results obtained are discussed with a keen on further field developments regarding intelligent, transparent, and responsible solutions countering social manipulation through realistic participation and contribution of different stakeholders from different disciplines.

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

1. Introduction

"I do respect people's faith, but I don't respect their manipulation of that faith in order to create fear and control." (Javier Bardem)

The efforts that agents like social platforms and dedicated outlets, institutions, researchers, and practitioners invest for dealing with social media manipulation mechanisms imply firstly understanding what they mean, why and how are built, and what is the scale and intensity of their force projection at individual and collective levels (Bastick, 2021; Fard & Maathuis, 2021). For instance, wars like the ones in Ukraine and Syria reflect how social media is used for communication purposes next to, e.g., intelligence information gathering, influencing public perception through techniques like propaganda, attacking the opposition or launching aggressive online campaigns, driving division and polarization, and building psychological deterrence through trolling, harassment, and personal attacks (Chen, Chen & Xia, 2022). Another example is the ongoing Covid-19 pandemic which through its digital presence and impact scale received its corresponding digital name Infodemic seeing the superabundance of information regarding concepts and events surrounding it, e.g., nature of the virus, its effects, vaccination types and campaigns, unknown and unforeseen effects of vaccines, and relation between the virus, vaccination, and general effects affecting people (WHO, 2020; Pérez-Escolar, 2021; Europol, 2021; Ivendi, 2022; Chockalingam & Maathuis, 2022). Hence, while defending mechanisms should be properly designed, developed, and deployed, they should consider a security awareness goal on existing and possible future threats plus the necessity of protecting (socio-ethical) values of humans and systems (EU Commission, 2022b; Maathuis & Chockalingam, 2022c).

Accordingly, this research aims to build a social manipulation detection solution that takes offender's perspective through a deep learning model for building or strengthening social media manipulation security awareness through a detection approach (Kaliyar, Goswami & Narang (2021), i.e., based on both context and content. Hence, a multidisciplinary perspective is adopted by integrating relevant studies from cyber/information security and operations, social media, and deep learning following a Design Science Research methodological approach (Peffers, 2007; Peffers, 2018) through the Data Science framework having the following contributions:

- To research and practitioner communities that could adapt/change their approach to thinking through the eyes of the attacker/offender when building social manipulation countering initiatives and solutions.
- To security, social media, and AI experts as the solution proposed is implemented based on deep learning techniques by encouraging them to consider creative and advanced approaches when dealing with social manipulation, and if applicable extend the approach taken in this research, or further release useful social manipulation datasets that could be used in upcoming studies.
- To decision makers involved in designing, developing, and deploying strategies and solutions that produce/enhance social media awareness on manipulation mechanisms and incidents based on intelligent techniques that should be further understood by a multi-domain audience and approach.

The reminder of this article is structured as follows. Section 2 discusses the background of this research together with relevant studies. Section 3 describes the research approach taken. Section 4 presents the deep learning techniques considered for building the model proposed. Section 5 discusses model implementation reflecting on choices and the results obtained. Section 6 discusses the findings and future research lines.

2. Social Manipulation Detection Research

The UN stresses that "information pollution is a hard-to-fix problem, for which there is no easy solution" and that disinformation is an "existential risk to humanity" as it damages trust in democratic societies, governments, institutions, and processes further calling to division and conflict (UN, 2021a). Specifically, the Common Agenda (UN, 2021b) contains 12 commitments from where the third (Promote peace and prevent conflicts), fourth (Abide by international law and ensure justice), sixth (Build trust), and seventh (Improve digital cooperation) are the most relevant measures when dealing with disinformation (UN, 2021c). The EU Commission highlights how important is to protect democratic values by tackling social manipulation methods like disinformation which is misleading or false information spread through social platforms for deceiving or protecting specific economic or political gain (EU Commission, 2022a). The institution proposed a set of initiatives on combating disinformation like the European Democracy Action Plan and the Strengthened Code of Practice on Disinformation which argues for, e.g., strengthening political transparency and empowering users and researchers (EU Commission, 2018; EU Commission, 2022a; EU Commission, 2022b). Moreover, through the risk assessment on Covid-19 disinformation conducted by Europol, resulted that disinformation threats are represented by several agents, i.e., individuals like criminals that want to get profit, states and state-backed actors that hope to achieve specific geopolitical aims, and opportunists who want to discredit official (re)sources.

Reisach (2021) argues that social platforms allow carrying out diverse businesses and political discourses, but also play a role in the creation and execution of social manipulation campaigns while integrating AI techniques to perform different functions. Furthermore, the author argues that society, social media platforms, and their users share their responsibility for users' (in)actions. To support this, the author suggests developing effective education solutions that are geared toward helping users become more responsible and tech-savvy. These include providing educational support to vulnerable populations, developing an ethical mindset, and promoting tech and media literacy. Chen, Chen, and Xia (2022) investigate user behavior and regulation processes in social networks and argue that social media has become an environment for the use of cyber weapons (Maathuis, Pieters & van den Berg, 2016), thus a new battlefield environment (Maathuis, Pieters & van den Berg, 2018c). Herein, power is augmented by goals that include analyzing relationships between people and actively intervening with deceptive techniques. Additionally, the authors are concerned since social media is used for warfare purposes or just as infosphere of cyber weapons, e.g., bots, botnets, trolls, and cyborgs with their functioning modules that include behaviour representation, user recognition, and discovery of homogeneous communities. Moreover, Shu (2022) performs a survey on existing recent detection advancements while acknowledging still one of the biggest issues in this field: the data challenge, i.e., the limited or lack of publicly open datasets which could be helpful for developing solutions for understanding existing manipulation topics, methods, and tactics utilized. Furthermore, the author acknowledges the role and increase in spread of deep learning-based solutions, but argues that when building and deploying them, they should be trustworthy, i.e., based on human knowledge, take into consideration aspects like noise or incomplete human feedback for improving representation learning, and broadly interpret the results obtained.

Critical to understanding and dealing with such social manipulation mechanisms is the existence and availability of relevant datasets as this represents the basis for building tackling/countering solutions. Patwa et al. (2021) built a Twitter fake news dataset with real and fake labels and a set of corresponding baseline ML (Machine

Learning) models. Furthermore, Kim et al. (2021) proposed the FibVID (Fake News Distribution Database for Covid-19) containing tweets with both Covid-19 and non-Covid-19 fake news content to support further analysis of the phenomenon and its spreading mechanisms. Sharma, D. K., & Garg et al (2021) proposed the IFND (Indian fake news dataset) dataset which has text and images for fake news identification based on fact-checking events from India between 2013-2021.

On detecting disinformation, Nasir, Khan & Varlamis (2021) propose a fake news solution based on a hybrid deep learning approach embedding CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks) techniques obtaining good results in comparison with other recent models. Moreover, Kaliyar, Goswami & Narang (2021) argue for capturing semantic and long-distance dependencies in sentences and build a hybrid deep learning model combining BERT (Bidirectional Encoder Representations from Transformers) with CNN techniques in a model named FakeBERT for capturing fake news relevant information for improving the classification performance. The authors stress that such an architecture helps dealing with ambiguity which represents one of the main challenges in this field. Seeing the high accuracy obtained (98.90%) and the approach followed, this research considers a similar combination. As sentiment analysis proved useful when detecting social manipulation mechanisms, Drus & Khalid (2019) execute a systematic review on different methods and conclude that both lexicon-based or ML-based techniques (baseline techniques like Naïve Bayes and SVM (Support Vector Machine), and deep learning) plus a combination thereof are used for social media analytics tasks ranging from consumer trends, disaster impact assessment, feelings and perceptions surrounding the ISIS group, to governmental elections. Furthermore, focusing on the common points of the applicability of diverse deep learning techniques, e.g., CNN and LSTM for multilingual sentiment analysis, Torales, Salas & Herrera (2021) militate for focus changing from single language to multilingual analysis, and the upcoming trend in building transformer-based architectures like the ones we implement.

The term Infodemic is defined by WHO (World Health Organization) as "too much information including false or misleading information in digital and physical environments during a disease outbreak" (WHO, 2020). Ivendi et al. (2022) conduct a sentiment analysis-based research considering data fusion from news broadcasting, health, and governments websites on different issues and events surrounding the Covid-19 pandemic. The authors use features like positive, negative, or neutral categorizations, text language and character count. For implementation, a deep learning approach is taken based on RNN and two-word clouds for real and real news with words like vaccine, people, and health, and fake news with China, pandemic, and Wuhan exist. Based on the results obtained the anger, fear, surprise, and negative emotions have higher frequency values in the fake news content when compared to the real news content. Moreover, to investigate sentiments in social media discussions on Covid-19 vaccination, a sentiment analysis and LDA (Latent Dirichlet Allocation) topic analysis is conducted by (Melton et al., 2021) based on Reddit data from December 2020 to May 2021. There the results show that in general a more positive sentiment is present in vaccine-related discussions over time although as the topic modelling shows, hesitation on vaccination implementation, digital interventions, and building new policies exist.

From the literature review conducted, this research acknowledges the limited amount of open publicly available datasets plus the risks associated with their availability, and the incipient state of the existing body of knowledge on advanced/ hybrid deep learning approaches for tackling social manipulation.

3. Research Methodology

As this research intends to build an intelligent model that detects social media manipulation mechanisms like disinformation for further contributing to building/enhancing social media security awareness of various communities, the following research questions are considered:

- How to build a deep learning model for detecting disinformation considering producing social media security awareness?
- How to evaluate the proposed model?

To reach this goal, a Design Science Research approach (Peffers, 2007; Peffers, 2018) is followed based on the Data Science framework through multidisciplinary research (Venable, Pries-Heje & Baskerville, 2017; Maathuis, Pieters & van den Berg, 2018b). Hence, a deep learning-based model is developed, evaluated, and proposed considering the following research activities:

Problem definition and solution aim: while the existing strategies, policies, and solutions consider, define, and implement different mechanisms for preventing and detecting social manipulation, they usually (i) take

defender's perspective which in this field proved weak as the offender has advantage in attack, and (ii) are still reliant on limited data and/or human experts which could be incomplete or might imply unexpected (propagated)errors/biases. Furthermore, seeing the large-scale impact and discourses that the ongoing war in Ukraine and Covid-19 pandemic continue to produce, there is an increasing need for building advanced and transparent intelligent solutions (Maathuis, Pieters & van den Berg, 2021) for proactively detecting social manipulation through offensive lenses. Thus, this research adopts offender's perspective proposing a deep learning-based social manipulation detection solution. Accordingly, literature review is conducted in scientific databases like IEEE Digital Library, ACM Digital Library, Scopus, and Google Scholar using combinations of keywords like 'social media', 'manipulation', 'disinformation', 'detect', and 'deep learning'.

Solution development: to develop the model, Covid-19 pandemic tweets are collected focusing on the first two waves, from 01.01.2020 to 21.03.2020 (first wave) and from 15.10.2020 to 31.12.2020 (second wave). Furthermore, the processes of data collection, processing, hashtag analysis, topic extraction, development using CNN and BERT, and the evaluation mechanism are implemented in Python and are further addressed. Specifically, the number of times used per hashtag is depicted in Figure 1:

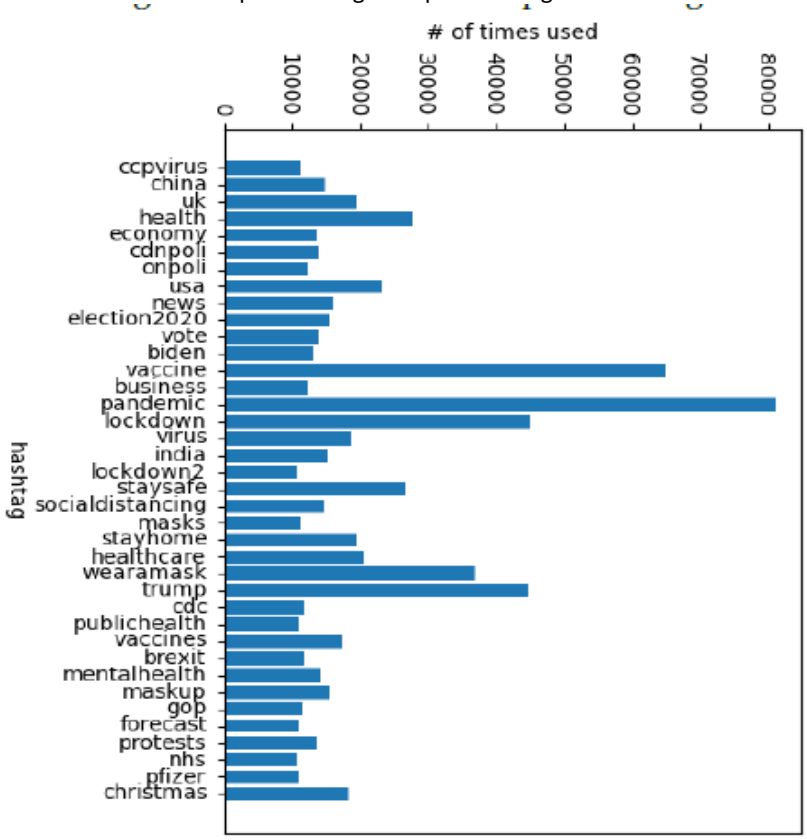

Figure 1: Number of times used per hashtag

Solution evaluation: the model is assessed evaluation metrics like accuracy and is improved through optimization while the results obtained are discussed.

Communication: the results obtained are discussed here and integrated in ongoing research in this domain while addressing the need for (i) realising publicly available social manipulation datasets for research and practitioner communities, and (ii) merging human with technological assessment in modelling and simulation settings for proactively countering social manipulation campaigns/operations, mechanisms, and incidents.

4. CNN and BERT Modelling Techniques

Due to the increase in innovation, computation, and (multi-source) data availability, the developments and solutions proposed based on AI techniques that match and even go further beyond human cognition and decision-making through different ML approaches, either have completely changed some societal domains or are in the process of doing that (Janiesch, Zschech & Heinrich, 2021; Maathuis, 2022b). Currently, at the core of development in the field of ML is building intelligent (complex) neural architectures, i.e., deep learning, which

while is being perceived as a novelty, "is around for decades" (Ng, 2017). A deep learning architecture contains three or more layers, meaning input, hidden, and output, that learns from a series of mathematical representations and operations from (large amounts of) data (Goodfellow, Bengio & Courville, 2016). Since this research uses specific neural architectures, i.e., BERT – Bidirectional Encoder Representations from Transformers and CNN – Convolutional Neural Networks, they are further introduced for background purposes.

CNN is among the most used deep learning techniques, was introduced by LeCun et al. (1989), and has a hierarchical structure implying extracting features from data through convolutional operations, see Figure 2. A CNN contains the following layers: input, convolution, pooling, fully connected, and output. The operational flow is as follows: convolutional layers learn features from the input data depending on the kernel used and pass results to pooling layers where data is condensed to a lower dimensional data without losing much information and further provided to fully connected layers which are normal fully connected neural networks. Shortly, the convolutional and pooling layers are used for extracting feature information from the dataset before the data is used to train the network for a task (O'Shea, K., & Nash, 2015; Gu et a., 2018; Li et al., 2021). Additionally, multiple pairs of convolutional and pooling layers could exist and learning is done through backpropagation.

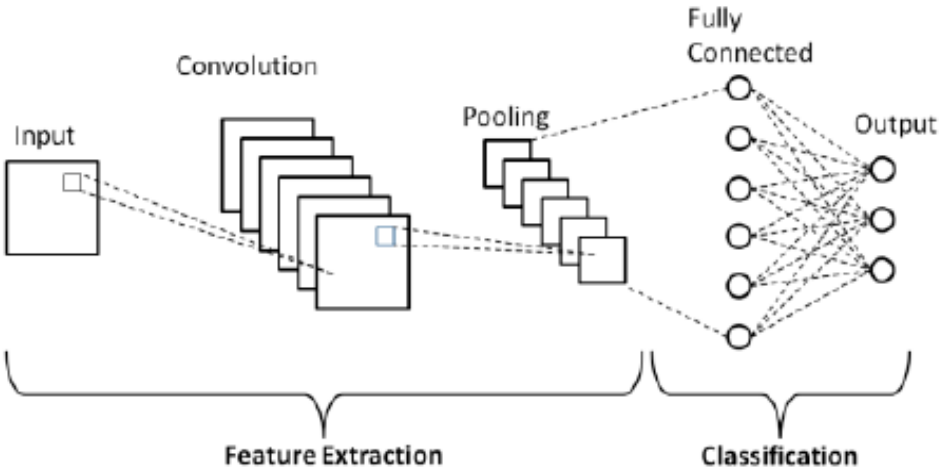

Figure 2: CNN architecture from (Phung & Rhee, 2019)

BERT is a transformer-based deep learning technique proposed by Delvin et al. (2018) and embeds twelve layers where through each layer more contextual information is acquired. Through its nature, BERT is the first left-to-right and right-to-left language architecture (Figure 3), i.e., deeply bidirectional. The input layer is the tokenized data, and the output layer is the masked language model which makes it a good choice when building text contextual embedding and text classification solutions (Delvin & Chang, 2018; Qiu et al. 2020).

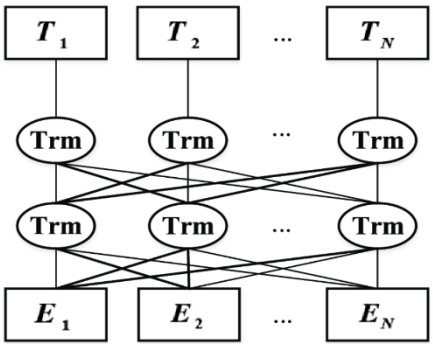

Figure 3. BERT model from (Lu et al. 2020)

5. Model Implementation and Evaluation

Previously, solutions that mix BERT with CNN show good results for tasks like depression detection (Rodrigues Makiuchi et al., 2019), message credibility assessment for fake news detection (Verna et al., 2022), and offensive speech identification in social media (Safaya, Abdullatif & Yuret, 2020). In this research BERT is combined with CNN with an additional feature from sentiment analysis that captures emotions within text. Moreover, for

implementation the Huggingface (Wolf et al., 2019) and VADER (Hutto & Gilbert, 2014) libraries are used in Python for BERT and generating polarity values for the twitter messages that are used as sentiment values, respectively. The evaluation is made using Covid-19 tweets.

The techniques are combined based on the architecture depicted in Figure 3 where the data flow and reasoning go from the input data(tweets) to output which represents classifying the input data as disinformation or not. Herein, a CNN model is added on top of the BERT layers with the architecture of the CNN model laid out in with ReLu (rectified linear units) as activation function (Chen, 2015) where the following changes are applied: (i) the number and size of convolutional layers is made a hyperparameter for allowing additional testing of size combinations to see which one is suitable for this goal, (ii) the softmax function is not used as the model only predicts two classes, and (iii) providing additional data into the CNN during the training phase as a way to add sentiment values as feature. This results in a vector with an additional dimension, step which could be skipped if desired which implies that no extra sentiment information will be added and later can be assessed if its addition improves the results obtained.

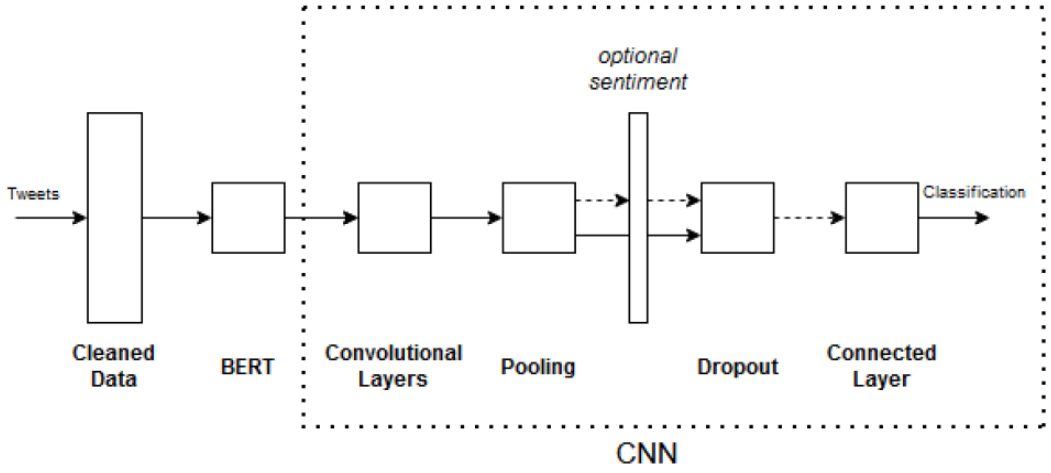

Figure 3`; Model architecture

Moreover, the hyperparameters of the model are: dropout (probability that a feature is dropped by the input layer), epochs (number of epochs to train the model), learning rate (rate at which the internal weights get adjusted each training step), sentiment (if the model should use the sentiment feature), convolution windows (list of sizes for the different convolutional windows), and bert layers. Accordingly, if the hyperparameter sentiment is set to true then extra sentiment values are unpacked from the batch and added to the call to the CNN network; otherwise, if the value is false then only the hidden layers from the BERT network are passed. Shortly, the model is trained during a given number of epochs that each consist out of smaller steps where the amount of these depends on the batch size and the size of the training set. During each step data from a batch is passed through both the BERT and CNN network which gives a class prediction of a sample. This sample is then compared to the actual class to calculate the loss which is in turn used to tune the weights in both networks. Once all epochs are completed the model is fully trained. Since there are many combinations of hyparameters that could be used for tunning, a random selection is adopted to reduce training and evaluation time. Hence, the results are in Table 1 and show that the model already reaches a high score after completing the second epoch and continues training for two more where it improves. Even though the model has a high accuracy, this does not necessarily mean that is the best option since it could overfit to training data and could not perform well on test data.

Table 1: Results overview

Epoch #	Training Loss	Training F1	Training Time	Validation Accuracy	Validation F1
1	0.189	0.938	0:00:30	0.934	0.920
2	0.033	0.988	0:00:34	0.989	0.993
3	0.01	1.00	0:00:33	0.989	0.993
4	0.01	0.998	0:00:36	0.89	0.993

For testing the model, two perspectives are considered: the first with sentiment values as extra feature and the second without. Accordingly, in Table 2 and Table 3 the results for three sets for both perspectives are presented. Hence, both perform at a similar level, but it appears that using sentiment values might improve the

generalization towards unseen data for the model reflecting the importance of human-based contributions.

Table 2: Test results with sentiment feature

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

Table 3: Test results without sentiment feature

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

To further test the quality of the model, a randomly selected sample of 10000 tweets is taken with 9874 non-empty tweets new to the model. Out of them, 4 were labelled as disinformation as in Table 4: while the first and fourth are clearly disinformation, the second and third tweet could be more nuanced seeing the training set. After a manual check on random tweets (thus not all due to its size), we conclude that do not contain disinformation. Of course, further updates can be considered when more data are available, and new forms of disinformation forms are identified.

Table 4: Disinformation tweets

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.
Hackers posing as exec from #China based Haier Biomedical target #COVID19 vaccine cold-chain, hoping to disrupt distribution. Potentially the work of a government backed entity, given targeting methods. Report from IBM
Corona Virus is a biological distraction created by the USA To disguise the financial crash that has been triggered by the FED to burst to equity bubble. This is a Black Swan event hidden from the public by the Elite.

6. Conclusions

As long as the humankind exists, conflict and manipulation are present, thus building a broad range of skills transferred/transformed to strategies, tactics, and solutions for tackling threats like manipulation inside or through social media platforms, and facilitating the development of intelligent, autonomous, and adaptive threats for producing new or enhancing existing social manipulation mechanisms. Accordingly, for developing next generation or enhancing existing solutions that tackle social manipulation mechanisms like disinformation, a high number of diverse datasets and a multi-faceted approach are necessary (Tucker et al., 2018; Maathuis, Pieters & van den Berg, 2018) while keeping humans in the loop. To tackle these issues, this research proposes a disinformation detection solution based on combining deep learning techniques, BERT and CNN with an additional feature that captures the sentiment values taking the offender's perspective and embedding lessons learned from previous social media manipulation campaigns. To build this solution, a Design Science Research is conducted through a Data Science approach using a built-in dataset of Covid-19 tweets.

As the solution proposed provides valuable results to the existing body of knowledge and considering existing practitioner approaches, this research advances by merging this solution with a generative approach for building a social media security awareness system that incorporates both defender and offender perspectives, and from there proposing design principles and requirements useful when building systematic corresponding strategies, policies, mechanisms, and solutions.

References

Agüero-Torales, M. M., Salas, J. I. A., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, *107*, 107373.

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, Y. (2015). *Convolutional neural network for sentence classification* (Master's thesis, University of Waterloo).

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

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

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, *161*, 707-714.

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 (2021). Covid-19 fake news. https://www.europol.europa.eu/covid-19/covid-19-fake-news

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

Ferrara, E. (2020). # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv preprint arXiv: 2004.09531*.

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

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, *77*, 354-377.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media.* 8(1), 216-225.

Iwendi, C., Mohan, S., Ibeke, E., Ahmadian, A., & Ciano, T. (2022). Covid-19 fake news sentiment analysis. *Computers and electrical engineering*, *101*, 107967.

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.

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.

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*.

Lu, Q., Zhu, Z., Xu, F., Zhang, D., Wu, W., & Guo, Q. (2020). Bi-gru sentiment classification for chinese based on grammar rules and bert. *International Journal of Computational Intelligence Systems*, *13*(1), 538-548.

Maathuis, C., Pieters, W., & van den Berg, J. (2018a). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C., Pieters, W., & Van Den Berg, J. (2018b). A computational ontology for cyber operations. In *Proceedings of the 17th European Conference on Cyber Warfare and Security*, pp. 278-288.

Maathuis, C., Pieters, W., & Van den Berg, J. (2018c). Assessment methodology for collateral damage and military (Dis) Advantage in cyber operations. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, 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. (2022c). Responsible Digital Security Behaviour: Definition and Assessment Model. In *European Conference on Cyber Warfare and Security.* 21(1).

Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, *14*(10), 1505-1512.

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.

Ng, A. (2017). Machine learning yearning. *URL: http://www. mlyearning. org/(96)*, *139*.

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.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Rodrigues Makiuchi, M., Warnita, T., Uto, K., & Shinoda, K. (2019, October). Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (pp. 55-63).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.

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.

Phung, V. H., & Rhee, E. J. (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, *9*(21), 4500.

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

Safaya, A., Abdullatif, M., & Yuret, D. (2020). Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation,* 2054-2059.

Sharma, D. K., & Garg, S. (2021). IFND: a benchmark dataset for fake news detection. *Complex & Intelligent Systems*, 1-21.

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

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., ... & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature*.

UN (2021a). Countering disinformation and promoting data transparency. https://www.undp.org/speeches/countering-disinformation-and-promoting-data-transparency?utm_source=EN&utm_medium=GSR&utm_content=US_UNDP_PaidSearch_Brand_English&utm_campaign=CENTRAL&c_src=CENTRAL&c_src2=GSR&gclid=EAIaIQobChMI7t2Di8KC-gIVRJBoCR38vQajEAAYAyAAEgLBJfD_BwE

UN (2021b). Common Agenda. https://www.un.org/en/common-agenda

UN (2021c). UN Common Agenda. https://www.un.org/en/content/common-agenda-report/

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.

WHO (2020). Infodemic. World Health Organization. https://www.who.int/health-topics/infodemic#tab=tab_1

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*.