ERINIA: Evaluating the Robustness of Non-Credible Text Identification by Anticipating Adversarial Actions

Piotr Przybyła[1,2,*], Horacio Saggion[1]

[1]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain
[2]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract
The ERINIA project is aimed to address the challenges posed by the increasing importance of automatic assessment of text credibility. Text classifiers are commonly used by platforms hosting user-generated content, including social media, to aid or replace human moderation in filtering out text that is undesirable for some reason – bullying, hate speech, fake news, etc. Unfortunately, deep neural networks are known for their vulnerability to adversarial examples, i.e. data instances with small modifications that preserve the original meaning, yet change the prediction of the target classifier. Here we describe the research actions of the ERINIA project, planned to tackle this challenge by assessing the robustness of currently used classifiers in the misinformation context, creating better methods for discovering adversarial examples and detecting machine-generated content.

Keywords
robustness, credibility assessment, adversarial examples

1. Introduction

Herein we summarise the ongoing project *Evaluating the Robustness of Non-Credible Text Identification by Anticipating Adversarial Actions* (ERINIA), carried out with the TALN group[1] at the Universitat Pompeu Fabra in Barcelona, Spain. The goal of this article is to briefly summarise the work being performed in the project. ERINIA is a an effort undertaken in response to clear societal and computational challenges and aims to provide solutions that could be implemented in practice. Thus, we hope our project will stimulate the discussion on these challenges.

In the following, we outline the motivation that led to the project (section 2), the work planned to address them (section 3) and, finally, the expected connections between our outputs and necessary future work (section 4). The project is funded as a Marie Skłodowska-Curie Postdoctoral Fellowship (MSCA PF) grant no 101060930 and takes place between November 2022 and October 2024. Note that apart from the research component, the ERINIA project also includes several training activities, which are not covered here. All the current information in the project is available at its website[2].

NLP-MisInfo 2023: SEPLN 2023 Workshop on NLP applied to Misinformation, held as part of SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing, September 26th, 2023, Jaen, Spain
piotr.przybyla@upf.edu (P. Przybyła); horacio.saggion@upf.edu (H. Saggion)
0000-0001-9043-6817 (P. Przybyła); 0000-0003-0016-7807 (H. Saggion)

[1]https://www.upf.edu/web/taln
[2]https://www.upf.edu/web/erinia

2. Motivation

The challenges that misinformation poses, both at the level of individuals and societies, are widely known [1, 2]. The abundance of misleading or wrong information on platforms hosting user-generated content[3] has naturally led to attempts to limit its prevalence through a variety of means, motivated not only by avoiding harm for users, but also by legal requirements in many countries [3].

Using the tools available in the field of artificial intelligence (AI), especially machine learning (ML), has long been recognised as a possible help in these efforts [4]. More specifically, assessing *credibility* of content can be seen as as a binary text classification: differentiating between credible and unreliable (*fake*) instances. Many tasks have been investigated through this lens, including detection of fake news [5], social media bots [6], biased journalism [7], rumours [8], propaganda [9] or inaccurate statements [10]. The vast majority of the proposed solutions are based on deep neural networks, either trained from scratch or based on pretrained large language models.

To deal with misinformation, but also other undesirable types of content (e.g. hate speech or bullying), the major platforms have implemented content filtering systems involving a mixture of AI- and human-based elements [11]. In some situations, automatic classifiers can even play a dominant role[4]. Unfortunately, the neural network architectures usually applied to such tasks show susceptibility to *adversarial examples* [12], i.e. data instances that have been maliciously modified in order to fool a classifier. The misinformation spreaders are likely to try to use such techniques to circumvent the filters in place. Additionally, text generators capable of producing realistic content in response to a prompt make it easy to massively produce content with deceptive purposes [13] and this possibility has become even more likely with the release of the newest models, such as ChatGPT [14].

For example, consider a hypothetical scenario, where a malicious actor causes confusion and distress by spreading entirely made-up rumours, e.g. about a hazardous fallout, using alarming headings such as *Radioactive dust approaching from the Mediterranean!*. The platforms used for hosting such content, e.g. social media and search engines, should detect all mentions of this information and act accordingly: recommending debunking articles, discouraging from sharing or banning the spreaders altogether. But if the ML model used for detection is not robust enough, the actor might rephrase the heading, e.g. as *Radioactive dust coming from the south!*, avoid the detection and continue the misinformation campaign.

The reasons outlined above lead to a need to assess the *robustness* of the credibility assessment models, i.e. their ability to maintain the expected accuracy level even in adversarial setting, where the content creator attempts to mislead the classifier. This is precisely the goal of the ERINIA project, implemented by developing methods and resources that help the discovery of the adversarial examples, improving our understanding of how easy it is to attack the common text classification methods. This could inform the discussion on whether such algorithms are fit to shape the content of the media channels of great importance for the modern societies.

[3] Note that we are only discussing textual content here, as the most wide-spread carrier of misinformation.

[4] https://www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/

3. Action plan

The investigation of adversarial examples (AEs) in the context of NLP is a relatively new effort [15], with many potentially interesting research directions widely open. Thus, within ERINIA, we plan to make contributions towards various areas: firstly, assessing the current situation regarding vulnerability of existing models (section 3.1), secondly, improving the search for adversarial examples through reinforcement learning (section 3.2) or meaning preservation (section 3.3). Finally, we are also experimenting with detecting machine-generated text (section 3.4).

3.1. Assessment of robustness against current attacks

The first goal of this action is to improve understanding of the current situation. This is necessary because content management platforms are already using machine learning-based tools [11] and many attack methods, albeit simple, exist [15]. The second goal is to establish a foundation for measuring the effectiveness of adversarial attacks, that could then be used to evaluate emerging methods, both in terms of new attacks and robust classifiers.

The results of this action is the BODEGA (*Benchmark for Adversarial Example Generation in Credibility Assessment*) framework, published recently [16]. It is based on four misinformation detection tasks (news bias assessment, propaganda detection, fact checking and rumour detection), used to train two types of general-purpose text classifiers (BiLSTM and a fine-tuned language model). Their robustness is then assessed by running eight AE generation techniques and checking if the classifiers indeed change their output after small modifications. The evaluation is based on a custom-designed measure, taking into account two aspects of similarity: surface forms and meaning.

The results of the experiments show that the fine-tuned language model is more robust than BiLSTM, but the success of an attack depends on the scenario, and in some cases the AEs based on character replacements (DeepWordBug) perform better, while in others more complex solutions (BERTT-ATTACK) are appropriate. It also appears that tasks with a longer text (i.e. news bias assessment) are more vulnerable to attacks than those with shorter input (i.e. fact checking). BODEGA is based on the *OpenAttack* framework [17] and openly available for download and use.[5]

3.2. Reinforcement learning for adversarial examples

As explained above, discovering AE is paramount to understand the vulnerability of a classifier before it is deployed in a sensitive application. The task can be seen as performing a search in a vast space – we have $(V * N)^k$ possibilities to perform k word replacements in a text of length N using a dictionary of size V. Most of the currently used methods rely on a human-designed heuristic, iterated until a classifier's decision is flipped.

However, it might be more efficient to train an ML model to design a procedure for a AE generation automatically. The ML framework used to train agents that can learn a behaviour in a given environment is *reinforcement learning* (RL). In this approach, an agent can perform

[5] https://github.com/piotrmp/BODEGA

actions that change the *state* of its environment and receive *rewards*, indicating the success of the current strategy. During training, the agent tries many different strategies and gradually learns to choose those that provide the highest rewards. The application of this framework for discovering AEs is quite natural: state corresponds to the current text of the example, actions to making modifications (e.g. word replacements) and rewards to AE quality measurement.

Essentially, this would mean training an ML model to find weaknesses in another ML model. Reinforcement learning has already been shown to lead to solutions that humans find surprising, e.g. in computer games [18]. Some initial experiments on similar solutions have shown promising results [19, 20] and we hope to apply this approach to misinformation detection within the ERINIA project.

3.3. Meaning preservation for adversarial examples

The modifications of the original text, turning it into an AE, are not unlimited: if too many are made, they change the meaning of the text, which fails to fulfil its original role. For example, a credible news piece might become unreliable after enough word replacements are performed in its text. Thus, another direction we aim to explore is *meaning preservation*, i.e. finding ways in which a text might be modified into an AE that preserve its original semantics.

This problem is not entirely novel, as it can be seen as *paraphrasing* [21], but including additional constraints on the modified text, namely that it changes the classifier's decision. Other tasks of paraphrasing with constraints include text simplification, i.e. making a document easier to read and understand [22], and style transfer, i.e. rewriting the text in a different style [23]. We hope to draw the inspiration from these areas to improve the AE generation methods in order to test the robustness of credibility assessment against such examples.

3.4. Detection of machine-generated text

Finding solutions that perform well in adversarial scenario requires understanding the perspective of the adversary, i.e. the author of the misinformation content. Since fake news articles are carefully designed to maximise their emotional appeal [24], writing them manually is likely to take significant effort. Thus, it seems plausible that the misinformation spreaders might turn to automatic text generators to increase their speed, especially when faced with the need to prepare many version of the same article to bypass content filtering. The NLP community was aware of such possibilities when the first models generating human-like text emerged [13], but recent advancements in the field, e.g. ChatGPT [14] have increased these concerns.

Detecting machine-generated content might be helpful in recognising attempts to circumvent content filtering procedures. While the current text generators produce content that untrained humans cannot detect [25], this task might be performed by automatic classifiers instead. Therefore, automatic text detection has been added to ERINIA goals to address this emerging challenge.

Our contribution [26] has been realised within the framework of the shared task *AuTexTification: Automated Text Identification* [27], a part of *5th Workshop on Iberian Languages Evaluation Forum (IberLEF 2023)*. We propose a collection of sequential features that measure the *predictability* of tokens, i.e. how likely they are according to language models. The underlying

assumption is that model-generated text is formulaic and repetitive, while human authors can compose surprising and creative writing. The predictability is supplemented with features describing grammatical correctness, word frequency and linguistic patterns. Finally, a neural network combining LSTM [28] and RoBERTa [29] is trained on the examples of human- and machine-generated content provided through the shared task.

Our solution has achieved the best performance in the binary classification task, both for English and Spanish input [27]. We hope this contribution will aid further development in the field, including in context of misinformation detection.

4. Impact and future work

Given the seriousness of the challenges posed by adversarial character of the misinformation, we don't expect them to be solved by a single project of limited size, such as ERINIA. Instead, we hope to deliver impact by enabling and encouraging further work in this area by NLP researchers and other stakeholders.

In particular, we expect the following new directions to follow our project and similar efforts:

- **Improved understanding of text classification robustness.** We hope that our analysis of the vulnerability of the models used in content filtering will contribute to a wider discussion on whether these models are fit for the role they play in the current online ecosystem.
- **Better ways to discover adversarial examples.** We know for sure the AEs exist, as even the simple methods available now uncover plenty of them. Thus, we expect the development of new methods for finding AEs, including our contributions based on reinforcement learning and meaning preservation, will allow these weaknesses to be spotted and fixed before the classifiers are deployed in publicly accessible systems.
- **Established procedures for verifying robustness of newly introduced models.** Thanks to the provision of BODEGA, there is a fast and easy way to test a given text classification algorithm against common attacks. Continuing this effort by adding new misinformation tasks and updated algorithms (both for attack and defense) is essential for establishing clear workflows to verify if a given model is ready for its deployment in an adversarial scenario.
- **More robust text classification models.** The currently common architectures based on deep neural networks appear to be particularly susceptible to adversarial examples. We expect the exploration made within ERINIA and similar efforts will be followed by the design of text classifiers that are more robust, without sacrificing the classification accuracy.
- **High-certainty text authorship assessment.** Crucially, platforms hosting user-generated content need reliable methods to uncover if given text was indeed authored by a user, or an automatic generator. While the performance of our submission was better than of other approaches, we do not believe it is yet sufficient for a high-stakes application scenario. Hopefully, more accurate methods will follow, delivering higher certainty in practical usecases.

Finally, we need to emphasise that while many of the issues tackled in this project may appear technical and hard to understand, they are relevant to the society as a whole. ML models have an increasingly important role on shaping the public debate and media landscape and we consider it paramount that their role is transparently demonstrated by the content platforms, understood by their users and controlled by policymakers.

Acknowledgments

The work is a part of ERINIA project that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. We acknowledge partial support from grant number MCIN/AEI/10.13039/501100011033 under the Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M) and by *Google Cloud* through *Research Credits*[6]

References

[1] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, B. Nyhan, Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature, Technical Report, Hewlett Foundation, 2018. URL: https://hewlett.org/library/social-media-political-polarization-political-disinformation-review-scientific-literature/.

[2] S. van der Linden, Misinformation: susceptibility, spread, and interventions to immunize the public, Nature Medicine 2022 28:3 28 (2022) 460–467. URL: https://www.nature.com/articles/s41591-022-01713-6. doi:10.1038/s41591-022-01713-6.

[3] F. Durach, A. Bargaoanu, C. Nastasiu, Tackling Disinformation: EU Regulation of the Digital Space, Romanian Journal of European Affairs 20 (2020) 5–20.

[4] G. L. Ciampaglia, A. Mantzarlis, G. Maus, F. Menczer, Research Challenges of Digital Misinformation: Toward a Trustworthy Web, AI Magazine 39 (2018) 65. URL: https://144.208.67.177/ojs/index.php/aimagazine/article/view/2783. doi:10.1609/aimag.v39i1.2783.

[5] P. Przybyła, Capturing the Style of Fake News, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), volume 34, AAAI Press, New York, USA, 2020, pp. 490–497. URL: https://aaai.org/ojs/index.php/AAAI/article/view/5386. doi:10.1609/aaai.v34i01.5386.

[6] K. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, F. Menczer, Arming the public with artificial intelligence to counter social bots, Human Behavior and Emerging Technologies 1 (2019) 48–61. doi:10.1002/hbe2.115.

[7] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A Stylometric Inquiry into Hyperpartisan and Fake News, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2018, pp. 231–240. URL: https://www.aclweb.org/anthology/P18-1022.

[6] https://edu.google.com/programs/credits/research/

[8] S. Han, J. Gao, F. Ciravegna, Neural language model based training data augmentation for weakly supervised early rumor detection, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019, Association for Computing Machinery, Inc, 2019, pp. 105–112. URL: https://dl.acm.org/doi/10.1145/3341161.3342892. doi:10.1145/3341161.3342892. arXiv:1907.07033.

[9] G. da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020), 2020, pp. 1377–1414. URL: http://propaganda.qcri.org/annotations/definitions.htmlhttp://arxiv.org/abs/2009.02696. arXiv:2009.02696.

[10] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The FEVER2.0 Shared Task, in: Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), 2018.

[11] M. Singhal, C. Ling, P. Paudel, P. Thota, N. Kumarswamy, G. Stringhini, S. Nilizadeh, SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice, in: The 8th IEEE European Symposium on Security and Privacy (EuroS&P 2023), IEEE, 2022. URL: https://arxiv.org/abs/2206.14855v2. doi:10.48550/arxiv.2206.14855. arXiv:2206.14855.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks (2013). URL: https://arxiv.org/abs/1312.6199v4. arXiv:1312.6199.

[13] I. Solaiman, M. Brundage, O. Jack, C. Openai, A. A. Openai, A. Herbert-Voss, J. W. Openai, A. R. Openai, G. K. Openai, J. Wook, K. Openai, S. Kreps, M. M. Politiwatch, A. Newhouse, J. Blazakis, K. Mcguffie, J. Wang, Release Strategies and the Social Impacts of Language Models, Technical Report, OpenAI, 2019. URL: https://arxiv.org/abs/1908.09203v2. arXiv:1908.09203.

[14] OpenAI, GPT-4 Technical Report, Technical Report, OpenAI, 2023.

[15] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, C. Li, Adversarial Attacks on Deep-learning Models in Natural Language Processing, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (2020). URL: https://dl.acm.org/doi/10.1145/3374217. doi:10.1145/3374217.

[16] P. Przybyła, A. Shvets, H. Saggion, BODEGA: Benchmark for Adversarial Example Generation in Credibility Assessment, arXiv preprint (2023). URL: https://arxiv.org/abs/2303.08032v1. arXiv:2303.08032.

[17] G. Zeng, F. Qi, Q. Zhou, T. Zhang, Z. Ma, B. Hou, Y. Zang, Z. Liu, M. Sun, OpenAttack: An Open-source Textual Adversarial Attack Toolkit, in: ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the System Demonstrations, Association for Computational Linguistics (ACL), 2021, pp. 363–371. URL: https://aclanthology.org/2021.acl-demo.43. doi:10.18653/V1/2021.ACL-DEMO.43. arXiv:2009.09191.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533. URL: https://www.nature.com/

articles/nature14236. doi:10.1038/nature14236.

[19] P. Vijayaraghavan, D. Roy, Generating Black-Box Adversarial Examples for Text Classifiers Using a Deep Reinforced Model, in: U. Brefeld, É. Fromont, A. Hotho, A. J. Knobbe, M. H. Maathuis, C. Robardet (Eds.), Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II, volume 11907 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 711–726. URL: https://doi.org/10.1007/978-3-030-46147-8_43. doi:10.1007/978-3-030-46147-8_43.

[20] Y. Li, P. Xu, Q. Ruan, W. Xu, Text Adversarial Examples Generation and Defense Based on Reinforcement Learning, Tehnički vjesnik 28 (2021) 1306–1314. URL: https://doi.org/10.17559/TV-20200801053744. doi:10.17559/TV-20200801053744.

[21] J. Zhou, S. Bhat, Paraphrase Generation: A Survey of the State of the Art, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5075–5086. URL: https://aclanthology.org/2021.emnlp-main.414. doi:10.18653/v1/2021.emnlp-main.414.

[22] M. Shardlow, A Survey of Automated Text Simplification, International Journal of Advanced Computer Science and Applications 4 (2014). doi:10.14569/SpecialIssue.2014.040109.

[23] R. Y. Pang, The Daunting Task of Real-World Textual Style Transfer Auto-Evaluation, in: EMNLP Workshop on Neural Generation and Translation (WNGT 2019), 2019. URL: http://arxiv.org/abs/1910.03747. arXiv:1910.03747.

[24] V. Bakir, A. McStay, Fake News and The Economy of Emotions: Problems, causes, solutions, Digital Journalism 6 (2017) 154–175. doi:10.1080/21670811.2017.1345645.

[25] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N. A. Smith, All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7282–7296. URL: https://aclanthology.org/2021.acl-long.565. doi:10.18653/v1/2021.acl-long.565.

[26] P. Przybyła, N. Duran-Silva, S. Egea-Gómez, I've Seen Things You Machines Wouldn't Believe: Measuring Content Predictability to Identify Automatically-Generated Text, in: Proceedings of the 5th Workshop on Iberian Languages Evaluation Forum (IberLEF 2023), CEUR Workshop Proceedings, Jaén, Spain, 2023.

[27] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of the AuTexTification 2023 Shared Task: Detection and Attribution of Machine-Generated Text in Multiple Domains, in: Procesamiento del Lenguaje Natural, Jaén, Spain, 2023.

[28] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, P. G. Allen, RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). URL: https://arxiv.org/abs/1907.11692v1. doi:10.48550/arxiv.1907.11692. arXiv:1907.11692.