Using Argumentation Theory to Fight Misinformation

Blanca Calvo Figueras

HiTZ Center - Ixa, University of the Basque Country UPV/EHU

Abstract

To fight misinformation, it is important to verify it, but also to ensure that true information reaches all those spaces that are likely to continue spreading hoaxes. To do this, it is necessary to generate reasoned, persuasive explanations adapted to each space and context. This project is the first to propose fighting misinformation through the automatic generation of explanatory arguments. To achieve this objective, it is necessary to create a tool that is capable of detecting where already debunked hoaxes are being replicated, deciding which is the most appropriate argumentative form to counteract it and, finally, generating the argument that can explain and persuade the spreader of the hoax or the users. For all this, an exhaustive study of the theory of argumentation is necessary, as well as the use of Natural Language Processing (NLP) techniques. Although argumentation has been studied in various contexts, there is no work to date that focuses on the analysis of the argumentation necessary for news verification. Taking advantage of the fact that much of the misinformation that spreads the most is usually verified by humans, in this work, we propose (i) generating a corpus of verifications paired with the disseminating messages associated to these, (ii) using existing theoretical frameworks to analyze the argumentation necessary to fight against misinformation, (iii) using cutting-edge techniques of language generation to explore this application, and finally (iv) exploring the impact that argumentative explanations can have in mitigating misinformation. In addition, we also want to study the use of NLP techniques as pre-bunking tools, which teach users how to detect misinformation.

Keywords

argumentation theory, counterargumentation, misinformation, large language models

1. Introduction

Fact-checkers dedicate great efforts to verifying misinformation. In addition, the work of fact-checkers also includes manually publishing the results of the data verification process to explain why a message spread on social networks (text message, video, image) is misinformation.

Recent studies suggest that even short rebuttals (280 characters) are preferable to no rebuttal at all. In fact, it has been empirically established that argumentative refutations are more effective than a simple indicator of possible misinformation, something that most social networks already do and which is considered could even be harmful [1]. Taking this into account, the goal of technology combating misinformation could be to help scale both explainable detection and the generation of counterarguments to the large amount of misinformation that is continually spread within social networks.

However, although current AI technology works reasonably well for automatic detection [2], most existing works explaining AI technology predictions are still in their infancy. Previous research has mainly focused on highlighting fragments in the message input or generating

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.

blanca.calvo@ehu.eus (B. C. Figueras)

simple summaries of the evidence used to make the prediction. With respect to mitigation of misinformation, the situation is bleaker, as there is no previous research on possible automatic AI-based strategies to generate automatic responses on social networks explaining why a given message is considered to be spreading misinformation.

To date, much of the fight against misinformation has focused primarily on two directions: 1) the creation of human hoax verification teams and 2) the creation of automated systems for news verification [3]. Human teams, although essential to combat the problem, are often overwhelmed by the speed of the spread of hoaxes. On the other hand, automated verification systems have several problems that are difficult to overcome, such as the impossibility of defining the limits of true knowledge in an automated and immutable manner, or the lack of transparency and explainability of these systems [4, 5].

2. Related Work

For this reason, in recent years solutions have been proposed that put verification teams at the center of the solution, but use technology as a catalyst for their work [6], helping in tasks such as evidence retrieval [7] or filtering the information that must be verified. In addition, it has also been proposed to use technology to spread truthful information and counteract misinformation proactively. Thus, cases in which an already-debunked hoax is being spread, these can be detected and attempts can be made to prevent its further dissemination [4, 8, 9].

With the aim of reducing the spread of debunked hoaxes, some researchers have studied the most effective methods informing citizens about the falsehood of certain information. These studies have observed that a good misinformation mitigation technique should include an explanation that reasons why that information is false [1]. This has already been applied in the fight against hate speech on social networks [10]. Other works have focused on analyzing the quality of the arguments in a debate [11], evaluating their persuasive power [12], identifying the winner of a debate [13], or predicting the usefulness of different arguments in a dialogue [14]. It has been observed that the persuasiveness of an argument can vary depending on the topic being discussed [15].

To identify the evidence and arguments used by verifiers, it is necessary to focus on the task of argument mining. This task is divided into three subtasks: argument segmentation, argument classification, and relationship classification between arguments [16]. Currently, there are no models to automate these tasks in Spanish, although the data that would allow training this type of systems does exist [17].

[18] proposed the generation of arguments to prevent privacy violations on social networks by developing templates that take into account the context and content of the publication. Based on this information, their system chooses one of the predefined arguments. In the field of misinformation, [19] proposed generating textual explanations about the decisions of a verification model. To do this, they formulated a task similar to automatic summarization, in which the model selects the most relevant parts of the verifiers' explanation and generates a justification. However, recent advances in the field of NLP have shown very satisfactory results in the task of automatic text generation, which allows us to go beyond the extraction of information and create the arguments automatically from a prompt [20, 21]. A prompt is a

linguistic template that provides the information necessary for a language model to generate text. In this case, contextual information about the publication on social networks and the hoax will be provided, with the aim of generating a reasoned explanation about the veracity of the information.

To this end, new methods based on automatic argumentation will be investigated to provide explanations of the verification process and generate automatic counterargumentation to counteract misinformation on social networks. This vision constitutes a disruptive approach regarding current research: (i) with respect to explainability, most previous research focuses on post-hoc or simple signaling methods and, (ii) with respect to counterargumentation to refute misinformation, no previous work has been done in the field of AI, although some psychological aspects have been studied and there is some work in the field of communication [1]. This is possible thanks to the enormous advances in performance in NLU and NLG provided by the development of large generative language models such as mT5 [21] and GPT3 [22].

3. Research Proposal

Automated text generation has undergone great advances in recent years. However, it has hardly been used to generate argumentative language. In this work, the main objective is to provide methods for the generation of automated argumentation to combat the spread of misinformation. More specifically, we want to explore whether the use of reasoned explanations can serve to reduce the spread of false information on social networks. In this sense, an important aspect of this thesis will be to investigate whether to respond with counterargumentation to messages that contain information already verified can reduce its spread.

Our hypothesis is that an explanation argued and adapted for each context of misinformation dissemination can have positive effects in terms of persuasion of disseminators and misinformation receivers.

The proposed objective raises a novel and ambitious line of research for several reasons: (i) there is no previous research in NLP on the automatic generation of counterarguments to mitigate the spread of misinformation; (ii) the first labeled corpus will be generated with explanatory counterarguments to refute the misinformation; (iii) while most of the work on fact-checking has been carried out in English, this project will promote a multilingual and cross-lingual approach; (iv) evaluation methods will be studied and proposed to measure the effect of counterargument against misinformation on social networks, something that has not been investigated to date. Achieving the main objective of the thesis will have two main benefits. First of all, the results of the thesis will allow us to better understand the various argumentative mechanisms underlying an automatic generation that is effective against misinformation in multiple languages. Second, the development of NLP systems for the mitigation of misinformation will have beneficial effects on the work of professional fact-checkers, since they will be able to count on these systems as assistants in a human-in-the-loop AI development environment. To achieve this end goal, a series of intermediate objectives are proposed that will contribute to the general vision of the thesis.

1. Study the persuasive effects of different types of arguments. A comprehensive study of research in argumentation theory is necessary to identify the most effective argumentative

structures against misinformation in different contexts. It will be based on the works on argumentation by Ecker et al., El Baff et al., and Donadello et al..

2. Collect data from both already verified claims and from social media posts that spread these claims for the languages of interest to the project.
3. Develop a computational scheme that allows us to note the types of arguments used by fact-checkers, as well as generate new argumentative explanations that respond to the requirements identified in the first part of the project. The labeling process of this data will be carried out between several annotators, quantitatively evaluating the agreement between annotators to alleviate the subjectivity inherent to the task.
4. Investigate different techniques for taking advantage of language models for text generation, such as mT5 or GPT3 [21, 22], for their application in the generation of arguments adapted to the context of the publication. The use of multilingual language models such as mT5 will allow us to obtain systems that generate these argumentative explanations automatically in several languages, something essential given the multilingual nature of most social networks such as Twitter.
5. Evaluate the generation of counterarguments using common metrics for text generation such as BERTscore or ROUGE. However, we will also adopt other metrics such as Novelty and Diversity [23], which go beyond simple distance-based evaluation.
6. Evaluate the effect that this counterargument can have on the spreaders and receivers of misinformation on social networks. We will look into relevant research from psychology [1] to propose an evaluation of the impact of counterargumentation to combat misinformation.

From a data point of view, the development of the project requires the compilation and annotation of corpora composed of verified data (fact-checks). This data will include the necessary levels of annotation for the project, explanations associated with the news verification label, and instances of counterargument of various types [1] to learn how to maximize the effect of automatic counterargument. Furthermore, the argumentative structure must transcend the support/attack paradigm, establishing relevant argumentative relationships for the counterargument of misinformation.

With respect to previously developed methods and algorithms, the current state-of-the-art of NLP is marked by deep learning neural systems. However, the reference data needed to train argument detection and generation models is generally very topic- and domain-specific, meaning that each topic requires its own data to obtain competitive models. Furthermore, the vast majority of previous works deal with the detection of misinformation in English. In this context, our methodological proposal will consist of the following innovations: (i) research in prompting and few-shot learning [24] will be proposed for the detection and generation of argumentation to counteract misinformation, thus facilitating the generation of high-performance models based on linguistic templates built by domain experts, and avoiding costly manual annotation of training data; (ii) the development of multilingual and cross-lingual approaches for the generation of counterarguments to mitigate the impact of misinformation will be investigated; (iii) an evaluation framework will be proposed to evaluate the impact of the argument in an objective way.

4. Methodology and Experiments

This thesis will span four years. The tasks and expected results for each of these years are described as follows.

4.1. Year 1: Theoretical framework, experimental design and data collection

In the first year, the theory of argumentation will be studied in order to determine what types of arguments can be most effective in combating misinformation. To do this, it will be necessary to review the works that study the persuasive power of counterargumentation and, specifically, the impact of counterargumentation on social networks. A method for evaluating the social impact of counterargument in misinformation will also be proposed.

At the same time, a database of fact-checks (verified news) will be created that will be extracted from the news verification media Maldita.es, Newtral, and EFEVerifica. We will focus on health and migration verifications. Additionally, an automatic system will be built to collect publications on Twitter and Telegram that are spreading these hoaxes.

Finally, it will be important to track the evolution of language generation models, since it is a field that is changing rapidly today. We hope to obtain the following results:

1. A theoretical framework to establish what types of argumentation are the most relevant to combat misinformation.
2. A methodology for evaluating the impact of counterargumentation.
3. A database of fact-checks on health and migration, accompanied with social media posts that talk about these hoaxes.
4. A baseline system to learn to detect already verified misinformation.
5. A publication based on the fact-check database and baseline system for detecting already verified news.

4.2. Year 2: Data Annotation

In the second year, an annotation scheme for the arguments contained in the fact-checks will be defined. This annotation scheme should allow the evidence and arguments contained in the explanations of the fact-checkers to be extracted. Next, the arguments will be noted in the database collected in the first year and this annotation will be evaluated.

Finally, linguistic templates (prompts) will be created that will be used as input to the language generation models. These templates will contain the arguments extracted from the fact-checkers' explanations as well as the context in which the fake news has been redistributed.

We hope to obtain the following results:

1. An argument annotation scheme.
2. A database annotated by domain experts. This first version will be used as a test bed for the extraction and generation of arguments.
3. Linguistic templates for the generation of counterargumentation.
4. A model for for multilingual argument extraction.
5. A publication applying few-shot learning and templates for the extraction of arguments and for the explainability of already verified news.

4.3. Year 3: Generation of arguments

The third year will focus on the generation of arguments. To do this, the different methods of language generation will be studied, the creation of different types of arguments will be experimented with, and the arguments generated will be evaluated in terms of coherence, factuality, and linguistic correctness. Experimentation will continue in multilingual scenarios, given that misinformation on social networks occurs in different languages.

We hope to obtain the following results:

1. Experiments with different language generation techniques based on seq2seq models.
2. Experimentation with techniques for knowledge grounding that allow mitigating the so-called “hallucination” of generative language models.
3. A dataset of automatically generated arguments, associated with a specific context and fragment of misinformation.
4. The evaluation of these arguments will be based on task-specific metrics, beyond evaluation metrics based on text generation.
5. A publication on generation of multilingual counterargumentation.

4.4. Year 4: Evaluation of the social impact of counterargumentation

Finally, the fourth year will focus on evaluating the impact of automatically generated arguments in the fight against misinformation. To this end, social experiments will be carried out based on two scenarios: (i) when the counterargumentation is manually generated by users of social networks; (ii) using counterargumentation models to respond automatically to social media posts that spread false information. In these experiments, the behavior of the misinformation spreaders will be observed before and after receiving the counterargument in both scenarios, as well as the interactions of other users that this publication attracts.

5. Conclusion

In this thesis, we will be studying the best techniques to fight misinformation using mitigation strategies, such as counterargumentation. We will be focusing on studying how argumentation theory can contribute to these strategies, the best NLP techniques to automate these strategies, and the social impact such tools might have.

Acknowledgments

We are thankful to the project DeepKnowledge (PID2021-127777OB-C21) funded by MCIN/AEI/10.13039/501100011033 and ERDF A way of making Europe. Blanca Calvo is funded by a PIF UPV/EHU 2023 doctoral grant.

References

[1] U. K. H. Ecker, Z. O'Reilly, J. S. Reid, E. P. Chang, The effectiveness of short-format refutational fact-checks, British Journal of Psychology (London, England: 1953) 111 (2020) 36–54. doi:10.1111/bjop.12383.

[2] I. Augenstein, Towards Explainable Fact Checking, 2021. URL: https://arxiv.org/abs/2108.10274. doi:10.48550/ARXIV.2108.10274.

[3] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a Large-scale Dataset for Fact Extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://www.aclweb.org/anthology/N18-1074. doi:10.18653/v1/N18-1074.

[4] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media, 2019. URL: http://arxiv.org/abs/1809.01286. doi:10.48550/arXiv.1809.01286, arXiv:1809.01286 [cs].

[5] F. Yang, M. Du, E. D. Ragan, S. K. Pentyala, H. Yuan, S. Ji, S. Mohseni, R. Linder, X. Hu, XFake: Explainable fake news detector with visualizations, in: The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019, Association for Computing Machinery, Inc, 2019, pp. 3600–3604. URL: https://researchwith.njit.edu/en/publications/xfake-explainable-fake-news-detector-with-visualizations. doi:10.1145/3308558.3314119.

[6] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. D. S. Martino, Automated Fact-Checking for Assisting Human Fact-Checkers, volume 5, 2021, pp. 4551–4558. URL: https://www.ijcai.org/proceedings/2021/619. doi:10.24963/ijcai.2021/619, iSSN: 1045-0823.

[7] C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, N. Collier, STANDER: An Expert-Annotated Dataset for News Stance Detection and Evidence Retrieval, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4086–4101. URL: https://aclanthology.org/2020.findings-emnlp.365. doi:10.18653/v1/2020.findings-emnlp.365.

[8] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a Known Lie: Detecting Previously Fact-Checked Claims, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3607–3618. URL: https://aclanthology.org/2020.acl-main.332. doi:10.18653/v1/2020.acl-main.332.

[9] A. Martín, J. Huertas-Tato, Huertas-García, G. Villar-Rodríguez, D. Camacho, FacTeR-Check: Semi-automated fact-checking through Semantic Similarity and Natural Language Inference, Technical Report arXiv:2110.14532, arXiv, 2022. URL: http://arxiv.org/abs/2110.14532. doi:10.48550/arXiv.2110.14532, arXiv:2110.14532 [cs] type: article.

[10] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini, CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech,

in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2819–2829. URL: https://aclanthology.org/P19-1271. doi:10.18653/v1/P19-1271.

[11] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, B. Stein, Computational Argumentation Quality Assessment in Natural Language, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 176–187. URL: https://aclanthology.org/E17-1017.

[12] R. El Baff, H. Wachsmuth, K. Al Khatib, B. Stein, Analyzing the Persuasive Effect of Style in News Editorial Argumentation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3154–3160. URL: https://aclanthology.org/2020.acl-main.287. doi:10.18653/v1/2020.acl-main.287.

[13] R. Ruiz-Dolz, S. Heras, A. García-Fornes, Automatic Debate Evaluation with Argumentation Semantics and Natural Language Argument Graph Networks (2022). URL: https://arxiv.org/abs/2203.14647v1.

[14] I. Donadello, A. Hunter, S. Teso, M. Dragoni, Machine Learning for Utility Prediction in Argument-Based Computational Persuasion, arXiv:2112.04953 [cs, stat] (2021). URL: http://arxiv.org/abs/2112.04953, arXiv: 2112.04953.

[15] R. J. Thomas, J. Masthoff, N. Oren, Can I Influence You? Development of a Scale to Measure Perceived Persuasiveness and Two Studies Showing the Use of the Scale, Frontiers in Artificial Intelligence 2 (2019). URL: https://www.frontiersin.org/article/10.3389/frai.2019.00024.

[16] J. Lawrence, C. Reed, Argument Mining: A Survey, Computational Linguistics 45 (2020) 765–818. URL: https://doi.org/10.1162/coli_a_00364. doi:10.1162/coli_a_00364.

[17] R. Ruiz-Dolz, J. Alemany, S. M. H. Barberá, A. García-Fornes, Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation, IEEE Intelligent Systems 36 (2021) 62–70. doi:10.1109/MIS.2021.3073993, conference Name: IEEE Intelligent Systems.

[18] R. Ruiz-Dolz, J. Alemany, H. SLA, A. Garcia-Fornes, Automatic Generation of Explanations to Prevent Privacy Violations (2019).

[19] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, Generating Fact Checking Explanations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7352–7364. URL: https://www.aclweb.org/anthology/2020.acl-main.656. doi:10.18653/v1/2020.acl-main.656.

[20] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual Denoising Pre-training for Neural Machine Translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742. URL: https://aclanthology.org/2020.tacl-1.47. doi:10.1162/tacl_a_00343, place: Cambridge, MA Publisher: MIT Press.

[21] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, in: Proceedings

of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[22] L. Floridi, M. Chiriatti, GPT-3: Its Nature, Scope, Limits, and Consequences, Minds and Machines 30 (2020) 681–694. URL: https://doi.org/10.1007/s11023-020-09548-1. doi:10.1007/s11023-020-09548-1.

[23] K. Wang, X. Wan, SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks (2018) 4446–4452. URL: https://www.ijcai.org/proceedings/2018/618.

[24] S. Wang, H. Fang, M. Khabsa, H. Mao, H. Ma, Entailment as Few-Shot Learner, arXiv:2104.14690 [cs] (2021). URL: http://arxiv.org/abs/2104.14690, arXiv: 2104.14690.