HiTZ@Disargue: Few-shot Learning and Argumentation to Detect and Fight Misinformation in Social Media

Rodrigo Agerri, Jeremy Barnes, Jaione Bengoetxea, Blanca Calvo Figueras, Joseba Fernandez de Landa, Iker García-Ferrero, Olia Toporkov and Irune Zubiaga

HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Donostia-San Sebastián, Spain

Abstract

DISARGUE opens a new and exciting avenue of research in AI-based explanatory argumentation to fight misinformation. This project will investigate and develop new methods based on automatic argumentation to provide explanations of misinformation detection systems and to generate automatic counterspeech to counteract misinformation in social media. This vision constitutes a disruptive approach with respect to current research: (i) with respect to explainability, most previous research has been focused on post-hoc or simple flagging methods and, (ii) with respect to counter-argumentation to refute misinformation in real time, no previous work has been done in the AI field, although some psychological and communication studies exist. Furthermore, DISARGUE's vision is made possible by the huge leaps in performance in Natural Language Understanding and Generation provided by the Transformer-based Large Language Models on which DISARGUE will investigate new methods to exploit them in few-shot learning settings. Additionally, the project aims to follow recent trends on human-centric AI where humans are by design in the loop. Being aligned with many of the hot topics in AI research (argumentation, few-shot learning, explainability) DISARGUE will benefit from the advances being achieved on those disciplines. Apart from the project description, we also provide an overview of the project's first contributions.

Keywords

Argumentation, Text Generation, Social Media, Automated Journalism, Media Discourse, Online Communication, Misinformation, Hate Speech, Counter Narratives, Natural Language Processing

1. Introduction

DISARGUE (TED2021-130810B-C21) is a project funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR within the call on *Proyectos Estratégicos Orientados a la Transición Ecológica y a la Transición Digital* (TED 2021), a program run by the Spanish Ministry of Science and Innovation. DISARGUE is a coordinated project, in which consortium is composed by the HiTZ Center - Ixa[1], and GureIker[2], both from the University of the Basque Country UPV/EHU. In this paper we will focus on the description of the first subproject which, lead by HiTZ, constitutes the Natural Language Processing (NLP) branch of DISARGUE.

The spread of misinformation[3] and hate speech in online social media and networks has become one of the greatest problems in the past decade [1, 2]. In fact, current spreading of misinformation is so massive that not even the largest journalistic and fact-checker teams can cope using manual methods only, making it an obvious task to automate [3]. While fact-checking organizations have started to combine their manual efforts with some Artificial Intelligence (AI) technology with the aim of semi-automatizing the misinformation detection step, current strategies and initiatives to address or mitigate the spread of misinformation remain mostly based on manual methods or on simple flagging mechanisms to point out that a given message (or thread) may be suspicious. Examples of the latter include the strategy followed by Twitter[4] (e.g., for the last 2020 US elections), TikTok, YouTube and Facebook, among others, which explicitly (and sometimes automatically) flagged some messages stating simply that they may be misinformation items but, crucially, without providing any explanations to justify such decisions. In this sense, the main mitigation strategy aiming to explain the reasons to flag as misinformation a given message relies on social media users (which could be fact-checkers, journalists, etc.) to publish themselves in social media the fact-checking results of highly shared messages, in the hope that it may help to mitigate the network noise generated by the misinformation [4, 1].

While there is no overall agreement among social researchers, journalists, and fact-checkers about which is the most appropriate response to a perceived misinfor-

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain

rodrigo.agerri@ehu.eus (R. Agerri)

[1] https://hitz.eus

[2] https://www.ehu.eus/es/web/gureiker/home

[3] For the sake of brevity, we will use the term "misinformation" to refer to "misbehaviour", namely, both "misinformation" (spreading fake news) and "disinformation" (intention to do harm by spreading fake news). Most of the time we will refer also to hate speech, another kind of "misbehaviour" in social media.

[4] https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

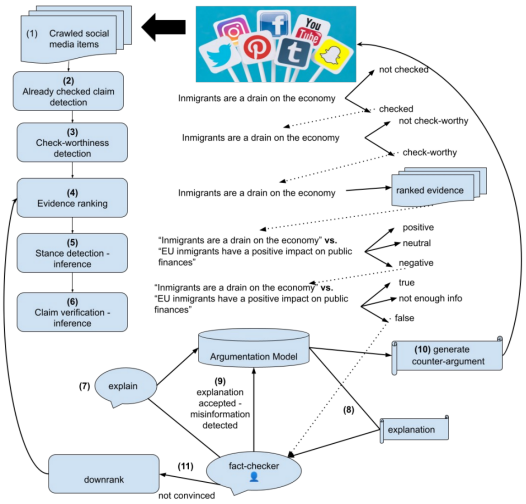

Figure 1: DISARGUE use-case scenario.

mation, most of the recommendations regarding the type of response that may be adequate refer, in one way or the other, to the fact that an appropriate mitigation strategy should include an explanation or argument providing reasons of various possible types (factual, rhetoric...) [4, 1]. Another important aspect is to adapt to the language of the message spreading misinformation. The aim of such explanations would be to convince or at least sow doubts on the person sharing the message and, perhaps most importantly, on the large number of users reading the interaction in social media.

Taking these considerations into account, DISARGUE's vision is to develop new techniques based on automatic argumentation to address both aspects of explainability thereby improving current techniques on misinformation detection and mitigation. By including argumentation-based explanations, DISARGUE will advance the state of the art in misinformation detection and mitigation by: (i) improving the interpretability of the predictions given by misinformation detection systems, (ii) automatically fighting misinformation by providing high-quality argumentative-based explanations and, (iii) using automatic natural language argumentation to provide a more interactive experience for the fact-checker using AI technology as an assistance. Thus, in the detection step, argumentation-based explanations would help domain-experts to better understand the decisions of the system. After detection, argumentation would focus on providing appropriate explanatory responses to counter items suspected of spreading misinformation thereby mitigating their overall effect on the public.

Figure 1 depicts the use-case scenario envisioned for DISARGUE. Steps 1 to 6 in the figure are originally from Augenstein [3], and describe the process that misinformation detection and mitigation pipelines tend to follow: claims crawled from social media are examined for check-worthiness. If deemed worthy, evidence is retrieved and ranked. Then, a stance detection process assesses agreement or disagreement with the claim. Finally, claim verification determines the claim's truthfulness based on obtained evidence. This basis of the use case scenario envisaged by our project is already implemented in many professional fact-checking teams, as human fact-checkers use AI technology as an assistant to detect misinformation in social media.

However, the next steps illustrated in Figure 1 is where DISARGUE's novelty lies, as (7) human fact-checkers request explanatory arguments about the automatic detection results. The system then (8) provides arguments based on input and evidence, leading to two outcomes: (9) acceptance, with the fact-checker flagging input as misinformation, followed by the (10) optional publication of an automated response or, if unconvinced, (11) rejection, with the fact-checker downranking the message, thus having to repeat the claim verification (6) and argumentative explanation steps when new evidence becomes available.

DISARGUE's vision faces several scientific and interdisciplinary challenges which are related with misinformation and argumentation theory, explainability and few-shot learning. First, how to leverage and produce new research from domain-experts (fact-checkers, social scientists, journalists) to guide the argumentation-based counteracting of misinformation in social media. Second, misinformation is spread nowadays in a variety of modalities (video, audio, images, text) and DISARGUE will face the challenge of offering explanations of attribution working in a multimodal environment and for different social media. Third, by the nature of the problem itself and of the current AI technology, misinformation detection and mitigation suffers perennially from a lack of annotated data. Thus, DISARGUE will need to research new methods of leveraging large pre-trained transformer-based language models to apply few-shot learning (learning with few examples from a specific topic or domain) for multimodal and multilingual misinformation detection, including the generation of argumentation-based explanations.

Although the DISARGUE's vision explained above may be applicable to any topic of misbehaviour in social media, this project will focus on tackling misinformation about: (i) public health and vaccines, (ii) immigration and, (iii) climate change, in a number of social media (Twitter, YouTube, Tiktok, etc.) and for Spanish, Catalan, Basque and English. The choice of topics is based on their perceived universality and cross-cultural character, namely, on the fact that misinformation on these three

topics follow a number of common themes independently of the specific countries, languages and local policies.

2. Related Work

In this section we review the most relevant previous work focusing on explainable misinformation detection and generation for misbehaviour mitigation, as well as few-shot learning and evaluation challenges in Natural Language Generation (NLG) tasks.

2.1. Explainable Misinformation Detection

A commonly accepted trend in Natural Language Processing (NLP) is to consider fact-checking as a multi-step automatic process usually performed sequentially, in a pipeline architecture, as depicted in Figure 1 steps 1-6. Thus, in the last step, claim verification, misinformation detection is essentially modelled as a pairwise classification task where the objective is to infer a label from a given claim with respect to a piece of evidence or a predefined topic, in what is usually also known as a Natural Language Inference (NLI) or Textual Entailment task [5].

Nowadays, as it is the case with many NLP tasks, the large majority of the best performing approaches address the task by considering only the textual content [6, 7] and, from 2018 onwards, by applying (in one way or another) large pre-trained language models [8, 9, 10]. This trend is recently changing by incorporating user-based interaction information from social media to improve the performance of the textual-based classifiers [11, 12, 13]. In any case, most approaches simply provide a prediction label, without aiming to provide any explanation to justify the classifier's decision. In an effort to make the decisions of the detection models more transparent, explainability has been addressed by post-hoc and by generation methods. Post-hoc methods focus on finding specific regions of the input that may explain the predicted label [14], while generation methods aim to generate a summary of the evidence used to predict the label in a simplified setting [3].

DISARGUE will develop unified vector-based representations for both textual and interaction data with the aim of providing a common approach to misinformation detection which exploits not only the text but also any network-based information characteristic from social media. Furthermore, it will integrate argument mining and explanatory argument generation in the decision making addressing both positive and negative evidence supporting the prediction. This would provide domain-experts with argumentation-based explanations, also using evidence from external knowledge, to support the decision taken by the misinformation detection system.

2.2. Argument Mining and Generation

Automatic techniques to counteract and mitigate the effects of misinformation in social media are mostly based on explicitly flagging a given message as being suspicious (without any specific explanation to justify the decision). Other approaches include the chatbot service created by the WHO and Facebook to combat misinformation regarding COVID-19[5]. However, the chatbot allows users to get factual and accurate information about the pandemic, it is not a service to counteract misinformation being spread in social media. Therefore, there is a clear lack of AI-based automated approaches to mitigate misinformation by generating appropriate counter-arguments in real time. The closest to this is the work undertaken within the HATEMETER project[6], where they propose using text generation techniques to generate counter-narratives to tackle anti-muslim hate speech. However, the aim of generating counter-narratives is substantially different from generating arguments to address misinformation [15] and it should work under different domain-experts' informed guidelines.

Natural Language Generation (NLG) has become one of the most important yet challenging tasks in NLP which is currently being addressed by the intense development and release of many Large Language Models (LLMs) [16, 17, 18]. One of the advantages of these neural models is that they enable end-to-end learning of semantic mappings from input to output in text generation. Transformer models such T5 [19] or a single Transformer decoder blocks like Llama 2 or Mistral [16, 17, 18] are currently the standard architectures for generating high quality text.

DISARGUE will provide novel AI technology by leveraging the latest advances in NLG to automatically generate counter-arguments guided by Retrieval Augmented Generation (RAG) [20] with the aim of counteracting the spread of misinformation in social media. This endeavour requires multidisciplinary work between domain-experts on misinformation (fact-checkers, journalists, policy makers, etc.) and AI researchers to generate arguments that fulfil a number of task-specific objectives related to fact-checking and reason-checking . In this sense, legitimate objectives could be to provide arguments based on factual, rhetoric (assessing the quality of premises and reasoning in persuasive or explanatory texts) or simply by alerting other users of the social media that a particular message might be spreading misinformation (and arguing the justification to do so).

[5] https://www.facebook.com/WHO/
[6] http://hatemeter.eu/

2.3. Few-shot Learning

The currently available data for misinformation tasks is highly compartmentalized and topic-specific, meaning that each topic requires its own data in order to learn relevant classifiers for fact-checking. This results in a general lack of data for the misinformation detection task, as many of the available data is also small in size, or has incompatible labelling schemes [3].

Recent work has shown that pre-trained language models can robustly perform classification tasks in a few-shot or even in zero-shot fashion, when given an adequate task description in its natural language prompt [16]. Unlike traditional supervised learning, which trains a model to take in an input and predict an output, prompt-based learning is based on exploiting pre-trained language models to solve a task using text directly [9]. Thus, some NLP tasks can be solved in an almost unsupervised fashion by providing a pre-trained language model with task descriptions in natural language [19, 21]. Surprisingly, fine-tuning pre-trained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks [22, 23].

2.4. Evaluation of Generated Text

NLG tasks such as the one proposed in DISARGUE present a considerable evaluation challenge. Thus, while it is possible to use usual distance-based metrics to evaluate the generated text such as ROUGE, BLEU or Bertscore [24], other works have proposed to use quality-based metrics such as Diversity and Novelty to evaluate the capacity of the model to generate diverse responses and the ability to generate sequences different from the data seeing during training or fine-tuning [25, 26].

However, a proper evaluation of the explanatory arguments generated in DISARGUE to explain the label prediction (in the detection phase) or to counteract misinformation (in the mitigation phase), requires to consider task-specific issues not taken into account in previous NLG or argumentation work. This implies evaluating the quality of the generated counter arguments regarding the supporting evidence found in trusted resources. A new promising avenue is that represented by JudgeLM, a scalable language model judge, designed for evaluating LLMs in open-ended scenarios [27].

3. Methodology and Work Plan

DISARGUE will focus on two novel models in the misinformation detection and mitigation pipeline, as depicted in Figure 1: (i) the Argumentation Model, which provides arguments based on both the input message and the evidence available to justify the prediction; (ii) the Generation model, which focuses on automatically generating arguments to counteract a perceived misinformation.

3.1. Work Plan

The Work Plan is structured in six Work Packages of which four are focused on the scientific contributions of the project.

WP2: Methodology. The aim is to define, adapt and integrate the modules, resources, data structures, data formats, and module APIs within the DISARGUE architecture. Additionally, focus will be given to the development of evaluation datasets and corpora to train argumentation-based explainable AI systems.

WP3: Explainable Misinformation Detection. The purpose of this WP is to work on joint and multitask models for explainable misinformation detection beyond post-hoc explainable methods. Novel approaches to exploit the full potential of LLMs will be developed, including prompting, generation and multimodal training, in order to make these models usable for the various tasks and languages of DISARGUE with minimal preparation effort, through zero-shot and, especially, few-shot learning.

WP4: Argument Generation. WP4 focuses on (i) defining and analyzing counter-argumentative patterns, creating natural language counter-arguments against detected misinformation and (ii) improving counter-argument generation by mining textual arguments from reliable sources via RAG. In summary, this task aims to prompt and train generative language models to enhance their text generation abilities for producing clear and understandable argumentation.

WP5: Evaluation of misinformation. WP5 aims to improve qualitative and quantitative evaluation of text generation-based tasks such for argument generation. More specifically, the objective will be to evaluate: (i) the effectiveness and quality of the prediction; (ii) the quality of the generated arguments for explanation and counter-argumentation, (iii) the effect of the counter-argumentation strategy via user-based evaluation guided by domain-experts.

4. Ongoing Work

There are a number of tasks currently being undertaken within the project. In this section we provide details of the most important ones with respect to the objectives and motivation provided in the introduction.

4.1. CONAN-EUS

CONAN-EUS[7] is a new parallel Basque and Spanish dataset for CN generation consisting of automatic trans-

[7] https://huggingface.co/datasets/HiTZ/CONAN-EUS

lations and professional post-editions of the original English CONAN. The corpus consists of 6654 machine translated HS-CN pairs and 6654 gold-standard human curated HS-CN pairs (per language) which makes it a unique resource to investigate CN generation from a multilingual and crosslingual perspective. Experimental results show that CN generation is better when mT5 is fine-tuned on post-edited training data, rather than on the output of MT. The paper will appear at LREC-COLING 2024 [28].

4.2. Automatic Generation of Critical Questions

Critical questions can be particularly helpful in the debunking process of misinformation. DISARGUE will study the automatic generation of these questions by exploring argumentation schemes, which represent different types of arguments illustrated through different premises. In argumentation theory, each argumentation scheme may be associated to a set of critical questions [29].

Based on this theory, we are currently working on building a model that, given an argument, outputs the critical questions needed to question the argument. Additionally, the automatic generation of critical questions would potentially enhance DISARGUE's quality of argumentation-based explainability. The limitations we are currently facing include: few and small datasets annotated with argumentation schemes, mainly in English; the great amount of different argumentation schemes (over 60, and it is not a closed set); and the automated transformation of the datasets does not result in particularly natural critical questions.

4.3. Multilingual TruthfulQA

A popular benchmark to evaluate the truthfulness of current LLMs is TruthfulQA, which evaluates truthfulness in English [30]. The dataset consists of question-answer pairs, each question with both true and false reference answers. No similar task on truthfulness has been done before for Basque, Catalan or Spanish, which means that currently is not possible to evaluate truthfulness of LLMs for those languages. DISARGUE will explore the truthfulness of monolingual and multilingual LLMs for those languages and English. The manual translated dataset and complementary experiments will be released soon.

5. Concluding Remarks

This paper outlines the DISARGUE project, which focuses on developing novel automatic argumentation techniques to enhance explainability and improve existing methods for detecting and mitigating misinformation. Currently, ongoing work has focused on analyzing the automatic generation of counterarguments in Basque and Spanish, as well as novel experimentation of critical question generation and text veracity authentication via the development of new benchmarks such as TruthfulQA for Basque, Catalan and Spanish.

Future work includes further experimentation on argument generation using LLMs and on the evaluation of the generated text, a crucial topic to understand the performance of our models.

Acknowledgments

Disargue (TED2021-130810B-C21) is a project funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR. Iker García-Ferrero is supported by a doctoral grant from the Basque Government (PRE_2021_2_0219). Rodrigo Agerri was also funded by the RYC-2017-23647 fellowship (MCIN/AEI/10.13039/501100011033 and by ESF Investing in your future).

References

[1] U. K. Ecker, Z. O'Reilly, J. S. Reid, E. P. Chang, The effectiveness of short-format refutational fact-checks, British Journal of Psychology 111 (2020) 36–54.

[2] R. Kouzy, J. A. Jaoude, A. Kraitem, M. B. E. Alam, B. S. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, K. Baddour, Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter, Cureus 12 (2020).

[3] I. Augenstein, Towards explainable fact checking, arXiv preprint arXiv:2108.10274 (2021).

[4] U. K. Ecker, J. L. Hogan, S. Lewandowsky, Reminders and repetition of misinformation: Helping or hindering its retraction?, Journal of applied research in memory and cognition 6 (2017) 185–192.

[5] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 809–819.

[6] I. Augenstein, T. Rocktäschel, A. Vlachos, K. Bontcheva, Stance detection with bidirectional conditional encoding, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 876–885.

[7] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, SemEval-2016 task 6: Detecting stance

in tweets, in: S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, T. Zesch (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 31–41.

[8] M. Hardalov, A. Arora, P. Nakov, I. Augenstein, Few-shot cross-lingual stance detection with sentiment-based pre-training, in: AAAI Conference on Artificial Intelligence, 2021.

[9] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2021) 1 – 35.

[10] D. Küçük, F. Can, Stance detection: A survey, ACM Comput. Surv. 53 (2020).

[11] R. Agerri, R. Centeno, M. Espinosa, J. F. de Landa, Álvaro Rodrigo Yuste, Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection, in: Procesamiento del Lenguaje Natural., 2021.

[12] M. S. Espinosa, R. Agerri, Á. Rodrigo, R. Centeno, Deepreading @ sardistance 2020: Combining textual, social and emotional features, EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020).

[13] M. Lai, A. T. Cignarella, L. Finos, A. Sciandra, Wordup! at vaxxstance 2021: Combining contextual information with textual and dependency-based syntactic features for stance detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings, 2021.

[14] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, A diagnostic study of explainability techniques for text classification, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 3256–3274.

[15] 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, 2019, pp. 2819–2829.

[16] T. Brown, e. a. Mann, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[17] H. Touvron, L. M. et al., Llama 2: Open foundation and fine-tuned chat models, ArXiv abs/2307.09288 (2023).

[18] A. Q. Jiang, A. S. et al., Mistral 7b, ArXiv abs/2310.06825 (2023).

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[20] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, ArXiv abs/2312.10997 (2023).

[21] T. Schick, H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in: Conference of the European Chapter of the Association for Computational Linguistics, 2020.

[22] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, ArXiv abs/2109.01652 (2021).

[23] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heinz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Computing Surveys 56 (2021) 1 – 40.

[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, ArXiv abs/1904.09675 (2019).

[25] K. Wang, X. Wan, Sentigan: Generating sentimental texts via mixture adversarial networks, in: International Joint Conference on Artificial Intelligence, 2018.

[26] Y.-L. Chung, S. S. Tekiroğlu, M. Guerini, Italian counter narrative generation to fight online hate speech, Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 (2020).

[27] L. Zhu, X. Wang, X. Wang, Judgelm: Fine-tuned large language models are scalable judges, ArXiv abs/2310.17631 (2023).

[28] J. Bengoetxea, Y. Chung, M. Guerini, R. Agerri, Basque and spanish counter narrative generation: Data creation and evaluation, in: LREC-COLING 2024, 2020.

[29] D. M. Godden, D. Walton, Advances in the theory of argumentation schemes and critical questions, Informal Logic 27 (2008) 267–292.

[30] S. C. Lin, J. Hilton, O. Evans, Truthfulqa: Measuring how models mimic human falsehoods, in: Annual Meeting of the Association for Computational Linguistics, 2021.