
Proceedings Paper:
Vasilakes, J., Zhao, Z. orcid.org/0000-0002-3060-269X, Vykopal, I. et al. (3 more authors) (2024) ExU: AI models for examining multilingual disinformation narratives and understanding their spread. In: Scarton, C., Prescott, C., Bayliss, C., Oakley, C., Wright, J., Wrigley, S., Song, X., Gow-Smith, E., Forcada, M. and Moniz, H.L., (eds.) Proceedings of the 25th Annual Conference of the European Association for Machine Translation, EAMT 2024. 25th Annual Conference of the European Association for Machine Translation, 24-27 Jun 2024, Sheffield, United Kingdom. European Association for Machine Translation (EAMT) , pp. 39-40. ISBN 9781068690716

ExU: AI Models for Examining Multilingual Disinformation Narratives and Understanding their Spread

Jake Vasilakes[1], Zhixue Zhao[1], Ivan Vykopal[2], Michal Gregor[2], Martin Hyben[2], and Carolina Scarton[1]

[1] Department of Computer Science, University of Sheffield, UK
[2] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
{j.vasilakes, zhixue.zhao, c.scarton}@sheffield.ac.uk
{michal.gregor, ivan.vykopal, martin.hyben}@kinit.sk

1 Project Overview

Online disinformation is a major challenge, with potential to cause economic, social, and medical harm (Zubiaga et al., 2018). Disinformation can be disseminated in multiple languages, which can be an overwhelming challenge for fact-checkers and journalists. It is therefore necessary to develop multilingual methods for analysing disinformation. The ExU project[1] aims to do just that, targeting stance classification and claim retrieval, two central tasks for assisting fact-checkers.

Stance classification predicts whether a piece of content (e.g., a social media post or news article) agrees or disagrees with a claim. Claim retrieval aims to find relevant fact-checks for a given claim. Previous research in these areas, predominantly in English, is largely focused on single languages (Küçük and Can, 2020). Still, there is no research that focuses on developing and evaluating at large-scale a single stance detection or claim retrieval model for multiple languages.

Given these challenges, the objectives of the ExU project are to (1) develop novel methods for multilingual disinformation analysis via the tasks of stance detection and claim retrieval and (2) follow a multilingual user-centric evaluation which focuses on providing explainability of model predictions to end users. Besides English, ExU will work with a set of 20+ languages, providing evaluation frameworks for Portuguese, Spanish, Polish, Slovak, Czech, Hindi and French (languages spoken in the UK and Slovakia). ExU started in November 2023 and is an 18-month project.

[1] https://exuproject.sites.sheffield.ac.uk

2 Progress

We conducted a survey of user requirements for our proposed tools at the Voices Festival of Journalism and Media Literacy,[2] which brings together journalists, fact-checkers, researchers, and educators. Participants were recruited among the visitors to the EMIF (European Media and Information Fund) booth. The survey consisted of 24 questions covering basic demographic information, exposure to multiple languages, and features of stance classification and claim retrieval.

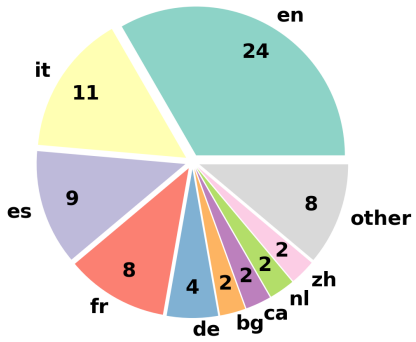

Figure 1: Counts of languages from responses to the survey question "Which languages do you encounter most often in your work?". The "Other" category is comprised of Czech, Hindi, Polish, Portuguese, Russian, Sinhala, Slovak, and Turkish, all of which had a count of one.

We obtained 29 survey responses. Almost all of participants (97%) encountered content in multiple languages when performing fact-checks. Figure 1 indicates the counts of the languages from the participant's answers.[3] For a fact-checking tool in general, participants would like content translated into a language of their choice, so it will be

[2] https://voicesfestival.eu/

[3] The event was held in Florence, Italy, so the results are biased towards EU languages and Italian specifically. We plan to obtain survey data from other demographics in the future.

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2), pages 39–40
June 24-27, 2024 ©2024 European Association for Machine Translation

necessary to ensure accuracy of translations both between our target languages and into other user-specified languages. For stance detection, respondents deemed it most important to automatically predict the stance of the post regarding the target claim and to automatically highlight the main argument of posts. For claim retrieval, respondents would like a high-level summary of the claim's fact-checks in addition to the fact-checks themselves.

3 Future work

Based on the initial survey results, we aim for our stance classification models to output accurate and explainable predictions for content across the target languages. Towards this we will utilise multilingual transformers such as Aya (Üstün et al., 2024), which covers all the languages in Figure 1. To address the lack of data for low-resource languages we may obtain small amounts of target language fine-tuning data, as previous work found this improved results (Scarton and Li, 2021). Additionally, we may translate low-resource languages into English before performing classification to leverage the knowledge from English models. We are exploring explainability via feature attribution and rationale extraction, and our preliminary research shows promise for using extractive rationales in multiple languages. Still, explanations ought to be consistent across languages and invariant to translation, yet previous work showed a performance gap in explainability methods between mono- and multi-lingual models (Zhao and Aletras, 2023), so we plan to explore this in depth.

For multilingual claim retrieval, we aim to employ a retrieval augmented generation model (Lewis et al., 2020) to help end users efficiently discern factual claims from debunked ones. This model may facilitate the existing tools for extraction of textual claims from any textual content found online and match them with existing fact-checks contained in our MultiClaim dataset (Pikuliak et al., 2023). The dataset contains 293,169 fact-checked articles and their corresponding claims in 39 languages. The output of the model will include the list of fact-checked claims relevant for each textual claim from the analysed textual content, their language and source references, along with the central claim summarisation of the retrieved claims in natural language.

4 Acknowledgements

The ExU project is funded by the European Media and Information Fund (grant number 291191). The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute.

References

Küçük, Dilek and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1):12:1–12:37, February.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, page 9459–9474. Curran Associates, Inc.

Pikuliak, Matús, Ivan Srba, Róbert Móro, Timo Hromadka, Timotej Smolen, Martin Melisek, Ivan Vykopal, Jakub Simko, Juraj Podrouzek, and Mária Bieliková. 2023. Multilingual previously fact-checked claim retrieval. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16477–16500. Association for Computational Linguistics.

Scarton, Carolina and Yue Li. 2021. Cross-lingual rumour stance classification: a first study with BERT and machine translation. In *Truth and Trust Online*, pages 50–59.

Zhao, Zhixue and Nikolaos Aletras. 2023. Incorporating attribution importance for improving faithfulness metrics. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4732–4745, Toronto, Canada, July. Association for Computational Linguistics.

Zubiaga, Arkaitz, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2):32:1–32:36, February.

Üstün, Ahmet, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.