Multilinguality in the VIGILANT project

Brendan Spillane[1], **Carolina Scarton**[2], **Robert Moro**[3], **Petar Ivanov**[4], **Andrey Tagarev**[2,4], **Jakub Smiko**[3], **Ibrahim Abu Farha**[2], **Gary Munnelly**[5], **Filip Uhlárik**[6], and **Freddy Heppell**[2]

[1] School of Information and Communication Studies, University College Dublin, Ireland
[2] Department of Computer Science, University of Sheffield, UK
[3] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
[4] Sirma AI EAD, Sofia, Bulgaria
[5] School of Computer Science and Statistics, Trinity College Dublin, Ireland
[6] Gerulata Technologies, Bratislava, Slovakia
brendan.spillane@adaptcentre.ie

Abstract

VIGILANT (Vital IntelliGence to Investigate ILlegAl DisiNformaTion)[1] is a three-year Horizon Europe project that will equip European Law Enforcement Agencies (LEAs) with advanced disinformation detection and analysis tools to investigate and prevent criminal activities linked to disinformation. These include disinformation instigating violence towards minorities, promoting false medical cures, and increasing tensions between groups causing civil unrest and violent acts. VIGILANT's four LEAs require support for English, Spanish, Catalan, Greek, Estonian, Romanian and Russian. Therefore, multilinguality is a major challenge and we present the current status of our tools and our plans to improve their performance.

1 Introduction

Disinformation and other related forms of harmful content has an increasingly detrimental effect on society. It is used to reduce trust in healthcare (Naeem et al., 2021), politics and rule of law (Bayer et al., 2019), and influence voting behaviour (Cantarella et al., 2023) or to increase funding and support for criminal networks. It has been classified as a strategic threat to the EU and its member states. Due to its nature, disinformation is extremely difficult for LEAs to identify, investigate and link to criminal activities. The Internet and social media platforms, where anonymity amplifies conspiracy, have provided ideal conditions for it to grow.

Research undertaken in previous projects, e.g. Horizon 2020 PROVENANCE (Yousuf et al., 2021) and WeVerify (Marinova et al., 2020), focused on developing supporting tools for journalists, fact-checkers or to inform the general public about disinformation. Meanwhile, most European LEAs have only recently set-up units to investigate crime related to disinformation. Thus, there is a lack of technical capabilities and institutional knowledge necessary to identify and investigate it. Disinformation that interests LEAs needs to be related to crimes or have the potential to affect the security of citizens. Therefore, general purpose tools do not capture the nuances of these specific cases.

One of the key challenges for VIGILANT is how to provide tools for multiple LEAs in different countries and targetting different languages. The consortium includes LEAs from Greece, Spain, Estonia and Moldova. Therefore, as a minimum, the VIGILANT platform (and its tools) should support: English (EN), Spanish (ES), Catalan (CA), Romanian (RO), Russian (RU), Greek (EL) and Estonian (ET). A Community of Early Adopters (CoEA), which currently has members from Ireland, Spain and Portugal, has been set up for LEAs who are not in the consortium but who wish to adopt VIGILANT. The project aims to provide support for all current and future CoEA required languages to create a common European platform to investigate disinformation linked to criminal activities.

2 Multilingual approaches in VIGILANT

To date, most work in disinformation analysis has been done for English. Developing monolingual approaches for each language from scratch is not feasible, given our project's time-frame and the need for large amounts of data for training state-of-the-art models. Approaches that leverage the

[1] https://www.vigilantproject.eu/

Name	EN	ES	ET	CA	EL	RO	RU
Event detection	●	–	–	–	–	–	–
Fact-checked claim detection	●	●	○	●	●	●	○
Central claim detection	●	●	○	○	○	○	○
Synthetic text detection	●	●	○	●	○	○	●
Stance classifier	●	○	○	○	○	○	○
Hate speech detection	●	●	○	○	○	○	○
Multilingual entity linking*	●	●	●	●	●	●	●
Paraphrase-resistant similarity*	●	●	○	○	○	○	●
Narrative analysis*	●	○	○	○	○	○	○

Table 1: VIGILANT tools (● = fine-tuning; ○ = zero-shot; – = no support). * means that the tool is under development.

knowledge learnt in models developed for the English language are thus needed.

Language adaptation in VIGILANT is a challenge for (i) natural language processing (NLP) and information retrieval (IR) tools; (ii) user interface and documentation; and (iii) training materials. We focus on (i), since (ii) and (iii) will be done by LEA professionals.

Multiple tools are being adapted for VIGILANT and here we discuss the challenges for multilingual support and our planned approaches. Table 1 presents a list of NLP and IR tools selected to appear in the VIGILANT platform and their current language support. Most tools support languages other than English in a zero-shot way, i.e., the model is pre-trained on multilingual data, but fine-tuned for the task only on English data. Although creating datasets for fine-tuning models in each language is not feasible, we will aim to create development sets (for domain adaptation) and test sets (to accurately assess whether our tools deliver multilinguality). These should support (i) the languages required by the four LEAs who are partners in the VIGILANT project, (ii) the languages of the members of the CoEA, and ultimately, (iii) as many European languages as possible. They will be created following strict ethical protocols and will be made publicly available when possible.

We will explore multiple approaches for multilingual adaptation:

- Further evaluating the few-shot and zero-shot models capabilities: both for encoder-based (e.g., multilingual BERT) and decoder-based models (e.g., GPT-4).
- Machine translation (MT) of the input to EN to use EN-trained models.
- Using MT for data augmentation and further fine-tuning of our EN-trained models with data in different languages.
- Investigating unsupervised domain adaption techniques, leveraging knowledge from unlabelled monolingual data for adapting our tools to other languages.

3 Project timeline and future work

VIGILANT is 16 months into its 36 month lifespan. In the 1st year of the project, we focused on analysing the tools to be deployed, further developing and adapting their APIs and on developing new tools for image/video and network analysis. We are releasing a minimum viable prototype in 2024 that will integrate NLP and IR tools, supporting EN only. We expect to develop new approaches adapted to multilingual settings in the next 12 months.

Acknowledgements

This work has been co-funded by the EU under the Horizon Europe VIGILANT project, GA No. 101073921 and the Innovate UK (grant 10039039).

References

Bayer, Judit, Natalija Bitiukova, Petra Bard, Judit Szakács, Alberto Alemanno, and Erik Uszkiewicz. 2019. Disinformation and propaganda–impact on the functioning of the rule of law in the eu and its member states. *European Parliament, LIBE Committee, Policy Department for Citizens' Rights and Constitutional Affairs*.

Cantarella, Michele, Nicolò Fraccaroli, and Roberto Volpe. 2023. Does fake news affect voting behaviour? *Research Policy*, 52(1):104628.

Marinova, Zlatina, Jochen Spangenberg, Denis Teyssou, Symeon Papadopoulos, Nikos Sarris, Alexandre Alaphilippe, and Kalina Bontcheva. 2020. Weverify: Wider and enhanced verification for you project overview and tools. In *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*.

Naeem, Salman Bin, Rubina Bhatti, and Aqsa Khan. 2021. An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, 38(2):143–149.

Yousuf, Bilal, M. Atif Qureshi, Brendan Spillane, Gary Munnelly, Oisin Carroll, Matthew Runswick, Kirsty Park, Eileen Culloty, Owen Conlan, and Jane Suiter. 2021. Provenance: An intermediary-free solution for digital content verification. In *Proc. of the CIKM 2021 Workshops co-located with 30th ACM Int. Conf. on Inf. and Knowledge Management (CIKM 2021)*, volume 3052. CEUR Workshop Proceedings.