HAMiSoN Project

Anselmo Peñas[1], Jan Deriu[2], Rajesh Sharma[3], Guilhem Valentin[4] and Julio Reyes-Montesinos[1]

[1]NLP & IR group at UNED, Madrid, Spain
[2]CAI, ZHAW, Zürich, Switzerland
[3]University of Tartu, Estonia
[4]Synapse Développement, France

Abstract

In order to face the current context of organised intentional misinformation campaigns, collectively known as disinformation, society must be aware of not just fake news, but also of the agents introducing misleading information, their supporting media, the nodes they use in social networks, their propaganda techniques and their narratives and intentions. This is a challenge that must be addressed in a holistic way, considering all these dimensions in order to identify, characterise and describe orchestrated disinformation campaigns. The HAMiSoN project aims at treating misinformation from this holistic view. The main challenge is integrating the message and the network level. To tackle this challenge, we propose to reveal misinformation's hidden intents: which agents introduce disinformation in the social media, which narratives do they use and with which concrete aims (such as polarising, destabilising, generating distrust, destroying reputation, etc.) We must also identify malicious and harmed agents and provide this information to the final analysts and users in explainable ways. Identifying misleading messages, knowing their narratives and hidden intentions, modelling the diffusion in social networks, and monitoring the sources of disinformation will also give us the chance to react faster to the spreading of disinformation.

Keywords

LaTeX misinformation, fake news, media analysis, multi-modal analysis, natural language processing, social networks, CEUR-WS

1. Introduction

Among the different kinds of misinformation, perhaps the most dangerous is the one created with the intention to harm, polarise, destabilise, generate distrust, destroy reputation, etc. by means of spreading untrue information [1, 2]. In this scenario of organised intentional misinformation campaigns (also called disinformation[1]) current fact-checking strategies are not enough [3].

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
anselmo@lsi.uned.es (A. Peñas); jan.deriu@zhaw.ch (J. Deriu); rajesh.sharma@ut.ee (R. Sharma); guilhem.valentin@synapse-fr.com (G. Valentin); jreyes@lsi.uned.es (J. Reyes-Montesinos)
https://nlp.uned.es/~anselmo (A. Peñas); https://www.zhaw.ch/en/about-us/person/deri/ (J. Deriu); https://rajeshsharma.cs.ut.ee (R. Sharma); https://linkedin.com/guilhem-valentin (G. Valentin)

[1]From here we use misinformation and disinformation interchangeably.

Fact Checkers need Artificial Intelligence tools to help them to identify the most important claims to check (check-worthiness), detect claims that they already have checked (verified claim retrieval), and check claims as soon as possible [4, 5]. However, fake news spreads 6 times faster than true ones [6], and 50% of the fake news propagation occurs in the first 10 minutes [7]. They are carefully prepared to behave in this way, they have an intention (not always explicit) and a coordinated spreading.

Given this scenario of organised intentional misinformation campaigns, we need strategies to anticipate and mitigate the spreading of disinformation [8]. We, as a society, must be aware not only about fake news, but also about the agents that introduce false or misleading information, their supporting media, the nodes they use in the social networks, the propaganda techniques they use, their narratives and their intentions. Therefore, we must address this challenge in a holistic way, considering the different dimensions involved in the spreading of disinformation and bring them together to really identify and describe the orchestrated disinformation campaigns. These dimensions are:

1. Detect misinformation
2. Acknowledging their organised spreading in social networks
3. Identifying its malicious intent
4. Bring everything together

To address the challenge of working on these dimensions, we have proposed the HAMiSoN (Holistic Analysis of Organised Misinformation Activity in Social Networks) project[2] under the CHIST-ERA[3] 2021 call. Our main audience are the analysts that make use of services such as fact-checkers for a further analysis and better understanding of the agents and narratives involved in disinformation campaigns. Making explicit the hidden intention behind disinformation campaigns will raise citizens' awareness. But for this purpose we need to move from just checking single messages or just analysing alterations in the social network, to seeing the complete picture. For example, one of our use cases is related to the international observers in political elections. These observers analyse the whole bunch of fake news as a whole, and identify the communication campaigns and their narratives employed to destroy the political opponent's reputation.

2. Project goals

The overall goals of this project are to *raise social awareness* and *mitigate disinformation propagation* by *making explicit the context* behind the intentional spreading of misleading information: sources and means of diffusion, stance and bias, intentionality and narratives.

This project is articulated around the following specific goals:

1. Develop models and systems for disinformation identification at message level: Claim check-worthiness detection, Stance detection, Multilingual verified claim retrieval.

[2] http://nlp.uned.es/hamison-project/
[3] https://www.chistera.eu/

2. Analyse the organised diffusion of disinformation and their narratives at social network level.
3. Integrate both sources of evidence (message level and network level) for better identification of organised misinformation campaigns.
4. Create evaluation datasets in multiple languages (English, Spanish, French, German, Estonian), and two modalities (text in Twitter messages, and video streams).
5. Organize shared tasks for competitive evaluation on stance detection in Twitter and claim-checking worthiness on videos.
6. Develop demonstration applications and a simulation tool for analysing potential coordinated disinformation campaigns in social networks.

3. Description

Integrating multimodal models [9] for misinformation detection with network models of misinformation diffusion to identify large misinformation campaigns [10] and their narratives constitutes a novel, holistic view of misinformation. It poses considerable and exciting challenges both on the conceptual and technical level.

There are two main current technologies to deal with the detection of disinformation. One, related to the needs of fact-checkers, focuses on the processing and analysis of single messages. The other, related to the detection of disinformation campaigns organised to influence a social network, relies on social network analysis: highly similar behaviour of different user accounts along time series are indications of a disinformation campaign.

However, both research lines remain separate research fields, although one gives context to the other. In fact, current AI models for misinformation detection are limited in the ability to represent and consider contextual information. It is still a research frontier we want to address.

HAMiSoN's most breakthrough goal is the integration of different technologies at both message and social network levels into a single system. Although we plan to take advantage of the hidden variable they share (their intentionality), there are many research questions that have to be addressed.

A straightforward approach would be to run all involved systems separately and then compare and combine their output. However, they don't leverage each other's signals and, in fact, the current state of the art achieves rather low performance.

The alternative we want to explore is what we call an "holistic" approach, where all tasks are considered simultaneously by one integrated system. This resembles in a way the end-to-end approach with neural networks which replaced component-based architectures for several NLP tasks.

Apart from solving the "whole" task - i.e. detection and description of organised disinformation campaigns - we see a great potential to also improve each single subtask, since they have access to much more data and insights. This hope is motivated by the success of multi-task learning, where additional unrelated subtasks help each other (Zhang and Yang, 2021). Messages that would be missed by local analysis level could be uncovered at this deeper latent level if they are strongly connected to an identified potential harmful network and, provided with

contextual information to better interpret their intention, eventually bring them to the attention of analysts.

The project aims to establish solid proofs of concept to validate the core of the technology developed to articulate the various AI models to detect, analyse and mitigate coordinated disinformation on social networks, bringing it to a Technology Readiness Level of 4. Qualitative evaluation will be performed to validate the holistic modelling of disinformation and in vitro experiments will be performed in simulated environments, using the simulation tools that will be developed.

The project includes several areas of computer science (CS), together with network science and social sciences. With respect to CS, the project requires expertise on: (i) Natural Language Processing for textual disinformation detection; (ii) Automatic Speech Recognition and Image Analysis for multimodal disinformation detection. Network science and theoretical modelling of disinformation diffusion is required to simulate various real scenarios. The holistic approach proposed in the project brings a natural way to integrate techniques from these different fields enabling cross-fertilization and synergy.

4. Expected impacts

4.1. Scientific and technological impact

The integration of evidence coming from the message, the social network, and the intended hidden goals, will be an important step towards a new modelling that can improve the state of the art, and produce more effective tools for the detection of organised misinformation campaigns.

The project will apply and combine Natural Language Processing, Artificial Intelligence techniques and Social Network analysis (Machine Learning, Multi-Agent based simulation, speech-to-text for audio transcription, image recognition, lifelong learning, etc.) in the identification of disinformation in text, images and video in several platforms. This joint approach will enable better modelling of coordinated disinformation campaigns, helping to train more efficient and adaptive methods for stance detection, claim worthiness checking, verified claim retrieval, similar message clustering and disinformation propagation modelling.

Whereas some techniques to perform misinformation identification tasks at both message level and network level may be validated with existing metrics, the combination of textual and non-textual features as well as the aggregation of both message content and network perspectives for orchestrated disinformation campaigns identification are still open research problems and involves finding new metrics correlating with the disinformation network modelling, as well as with the impact of mitigation actions over the disinformation propagation through this network.

In particular, multilingual claim similarity models will have an impact in verified claim retrieval and clustering of similar messages across different languages. The consideration of textual and non-textual features will have an impact in stance detection and news fact-checking. The simulation of disinformation spreading in social networks will have an impact in the evaluation of mitigation actions. The design of new architectures to gather and leverage evidence from different levels (message, network, intention) will have an impact in the methodologies for disinformation detection. The release of datasets in languages other than English for stance

detection, for claim detection in transcripts of video posts, and for harming narratives during political elections will have an impact in the research community and in technology transfer.

4.2. Social impact

By analysing the sources of disinformation, how they interconnect in various posts and messages, as well as in various modalities and languages, shedding light on their narratives, the project will provide means for a deeper understanding of disinformation in social networks and media and, indirectly, better anticipate and limit its propagation.

The project will develop tools to assist Fact Checking organisations, civil society, NGO and political observers to better and sooner detect disinformation campaigns, which will indirectly benefit the whole community of social media users, providing means to contextualise disinformation messages.

In particular, being able to make explicit and attach to the messages the intentions behind, the propaganda techniques they may use, the sources, their subjectivity, stance and bias degree will help provide evidence of disinformation activity and increase awareness in society for social media users. By modelling the disinformation propagation in simulated social network environments, the project will also provide means to adjust mitigation actions against the spreading of disinformation, by evaluating and measuring their potential effect on the agents and on the propagation.

Acknowledgments

This work was supported by the CHIST-ERA HAMiSoN project grant CHIST-ERA-21-OSNEM-002, by AEI PCI2022-135026-2, SNF 20CH21 209672, ANR ANR-22-CHR4-0004 and ETAg.

References

[1] M. Schütz, A. Schindler, M. Siegel, K. Nazemi, Automatic fake news detection with pre-trained transformer models, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII, Springer, 2021, pp. 627–641.

[2] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, defend: Explainable fake news detection, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 395–405.

[3] G. Da San Martino, A. Barron-Cedeno, P. Nakov, Findings of the nlp4if-2019 shared task on fine-grained propaganda detection, in: Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda, 2019, pp. 162–170.

[4] A. Patwari, D. Goldwasser, S. Bagchi, Tathya: A multi-classifier system for detecting check-worthy statements in political debates, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 2259–2262.

[5] S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, P. Nakov, It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction, arXiv preprint arXiv:1908.07912 (2019).
[6] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.
[7] T. Zaman, E. B. Fox, E. T. Bradlow, A bayesian approach for predicting the popularity of tweets (2014).
[8] A. Zubiaga, M. Liakata, R. Procter, Learning reporting dynamics during breaking news for rumour detection in social media, arXiv preprint arXiv:1610.07363 (2016).
[9] M. Dhawan, S. Sharma, A. Kadam, R. Sharma, P. Kumaraguru, Game-on: Graph attention network based multimodal fusion for fake news detection, arXiv preprint arXiv:2202.12478 (2022).
[10] S. Sharma, R. Sharma, Identifying possible rumor spreaders on twitter: A weak supervised learning approach, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.