Overview of NLP-MisInfo 2023: Workshop on NLP applied to Misinformation

Roberto Centeno[1,*], Rodrigo Agerri[2]

[1]Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
[2]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Spain

Abstract

The 2023 Workshop on NLP applied to MisInformation (NLP-MisInfo 2023) is at its First Edition, held as part of SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing. NLP-MisInfo aims at fostering research both at the theoretical and at the level of practical real-world applications of NLP technologies applied to misinformation mitigation. The workshop aims at bringing together researchers, developers and industries interested in the problem of mitigating misinformation through NLP technologies. We will discuss recent trends and research projects, as well as developments and advances being made in the area of NLP to address the problem of misinformation from different perspectives.

Keywords

Misinformation, NLP, disinformation, fake news, Harmful Information Detection, fact-checking

1. Introduction

The impact of fake news on the global economy, public health and even the creation of panic in society has been extensively documented in the past few years with countless examples [1, 2]. Thus, the high cost associated with the spread of fake news: is the absence of control and verification of the information, which makes social media a fertile ground for the spread of unverified or false information. With this in mind, we can affirm that the magnitude, diversity and substantial dangers of fake news and, in more general terms, the disinformation circulating on social media is becoming a reason for concern due to the potential social cost it may have in the near future [3, 4]. As a consequence, the research community in the field of Natural Language Processing has been focusing on the detection and intervention of fake news using techniques such as Machine Learning and Deep Learning [5, 6], and taking into account: *i)* Content-based features contain information that can be extracted from the text, e.g. linguistic features [7]. *ii)* Context-based features contain surrounding information such as user characteristics, social network propagation features, or users' reactions to the information [8, 9].

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

[*]Corresponding author.

✉ rcenteno@lsi.uned.es (R. Centeno); rodrigo.agerri@ehu.eus (R. Agerri)
🌐 http://nlp.uned.es/~rcenteno/index.php (R. Centeno); https://ragerri.github.io/ (R. Agerri)
iD 0000-0001-9095-4665 (R. Centeno); 0000-0002-7303-7598 (R. Agerri)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3525/overview.pdf

These approaches handled the phenomenon from a validity aspect, where they labelled a claim as “False“ or “True“ [10]. However, others tried to tackle it from a stance perspective trying to determine whether a tweet (or claim in general) is in favour, against, or neither to a given target entity (person, organization, etc.) [11], or even from a social perspective where how the spreading of misinformation is performed through social networks [12].

In spite of the current advances in the field, we, as a society, must be aware not only of fake news but also of the agents that introduce false or misleading information, their supporting media, the nodes they use in social networks, the propaganda techniques they use, their narratives and their intentions [13].

Therefore, we must address these challenges, providing new techniques and methods to really identify and describe the orchestrated disinformation campaigns, such as: detecting misinformation; claim worthiness checking, stance detection and verified claim retrieval; models of disinformation propagation, source detection using social network analysis; Identifying its malicious intent: narratives that want to be spread, benefited and injured agents and final goals; etc.

2. Topics of Interest

All those approaches that can serve, from different perspectives, to tackle the misinformation problem, in general, and by using NLP tools in particular, find their place in NLP-MisInfo 2023. Specifically, the topics of interest include, but are not limited to:

- Stance detection and polarization
- Automated claim verification (inference, counter-argument generation, multilinguality/crosslinguality)
- Reliability detection
- Misleading headlines
- Collective intelligence
- Digital entities
- Harmful Information Detection: fake news and hate speech
- Semi-automatic disinformation annotation
- Disinformation resources
- Misinformation and economic with the truth
- Automatic fact-checking
- Evidence Extraction
- QA-based Verification
- Representation of Misinformation Propagation with Knowledge Graphs
- Conspiracy detection
- Spreading disinformation simulation

Besides the aforementioned topics of interest, papers can be of the following three types:

- **Dataset submissions**. Present and describe a dataset related to the topic of the workshop that has been or is being developed.

- **Projects submissions**. Describe ongoing projects within the workshop's topic, both academic and industrial.
- **Original**, unpublished contributions are also welcome.

3. Submissions

The NLP-MisInfo 2023 Workshop received 7 submissions, of which 6 were accepted. Articles have been submitted from seven different countries, i.e., Spain, Poland, United Kingdom, Saudi Arabia, Switzerland, Estonia, and France. The accepted articles, collected in these Proceedings, have primarily addressed two topics. The first issue concerns the usage of NLP techniques for detecting misinformation; the second issue concerns more general approaches based on research projects for addressing the misinformation problems from a multi-perspective approach.

With respect to the first issue, in the article by Álvaro Huertas-García et al., entitled: *"Countering Malicious Content Moderation Evasion in Online Social Networks: Simulation and Detection of Word Camouflage"*, the authors present a set of resources for addressing the misinformation problem. In particular, it introduces novel methodologies and tools to combat content evasion in multilingual Natural Language Processing on social networks. A unique Python package, "pyleetspeak", is developed, offering a customizable system for simulating multilingual content evasion through word camouflage techniques. The study also presents a synthetic multilingual dataset of camouflaged words, facilitating the training of models for camouflage detection. They show the utility of the tool in improving content moderation, enhancing online security, and serving as a potential data augmentation tool for AI systems.

In this issue, we can find the article entitled: *"ELAINE: rELiAbility and evIdence-aware News vErifier"*, by Carlos Badenes-Olmedo et al. It presents ELAINE, a hybrid proposal to detect the veracity of news items that combines content reliability information with external evidence. The external evidence is extracted from a scientific knowledge base that contains medical information associated with coronavirus, organized in a knowledge graph created from a CORD-19 corpus. The information is accessed using Natural Language Question Answering and a set of evidences are extracted and their relevance measured. By combining both reliability and evidence information, the veracity of the news items can be predicted, which is very promising for the veracity detection task.

With the same objective of improving misinformation detection through NLP techniques, we find the article entitled: *"Where Does It End? Long Named Entity Recognition for Propaganda Detection and Beyond"*, by Piotr Przybyła and Konrad Kaczyński. They investigate how the extensive span lengths affect the recognition of propaganda, showing that the task difficulty indeed increases with the span length. They also propose a new solution, including an adaptive convolution layer that facilitates the sharing of information between distant words. This allows for improved length preservation without sacrificing overall performance.

Finally, with regard to the first issue, the article entitled: *"Google Snippets and Twitter Posts; Examining Similarities to Identify Misinformation"* by Saud Althabiti et al., investigates the applicability of Google search and its results as a practical tool for detecting fake news on platforms like Twitter. The research focuses explicitly on comparing Google search result snippets with tweets to assess their similarity and determine if such similarity can serve as an indicator of misinformation. However, the study reveals that the observed similarity between tweets and snippets does not necessarily correlate with news credibility.

With respect to the second aspect, research projects about misinformation, two articles were submitted. The first, entitled: *"ERINIA: Evaluating the Robustness of Non-Credible Text Identification by Anticipating Adversarial Actions"*, by Piotr Przybyła and Horacio Saggion, presents the ERINIA project. This project is aimed to address the challenges posed by the increasing importance of automatic assessment of text credibility. Text classifiers are commonly used by platforms hosting user-generated content, including social media, to aid or replace human moderation in filtering out text that is undesirable for some reason – bullying, hate speech, fake news, etc. Unfortunately, deep neural networks are known for their vulnerability to adversarial examples, i.e. data instances with small modifications that preserve the original meaning, yet change the prediction of the target classifier. Here we describe the research actions of the ERINIA project, planned to tackle this challenge by assessing the robustness of currently used classifiers in the misinformation context, creating better methods for discovering adversarial examples and detecting machine-generated content.

Finally, the article entitled *"HAMiSoN Project"* by Anselmo Peñas et al. presents the HAMiSoN project which aims at treating misinformation from this holistic view. The main challenge is integrating the message and the network level. To tackle this challenge, they propose to reveal misinformation's hidden intents: which agents introduce disinformation in social media, which narratives they use and which concrete aims (such as polarising, destabilising, generating distrust, destroying reputation, etc.). They propose also to 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.

4. Keynote Speeches

As part of the Workshop, two Keynote Speeches were given. The first was centred on the industry point of view, i.e. how the industry works to mitigate the misinformation. It was entitled *"Understanding the discourse: NLP in the fight against disinformation"*, and was given by Carlos Ponce, IT engineer in the well-known fact-checker Maldita.es[1]. The second, with an NLP point of view, entitled *"Fake news and conspiracy theories: distinguishing conspiracy narrative from critical thinking"*, was given by Professor Paolo Rosso. Further details are in the following.

4.1. Understanding the discourse: NLP in the fight against disinformation

Abstract: Disinformation is a global problem: it wins and loses elections, generates fear and distrust in the population, and affects the security and integrity of people. At Maldita.es they know this very well, they have been fighting against it and its effects for years. In this workshop, we will take a practical tour of the workflow and the tools that the Maldita team relies on to stand up to this battle. We will talk about their use of NLP to engage with their

[1] https://maldita.es/

audience and monitor public discourse and the -sometimes unfathomable- use cases of Machine Learning in the fight against misinformation and the creation of evidence-based content.

Carlos Ponce is a Computer Engineer from the UPM, theatre director and development manager at Maldita.es. The Maldita.es Foundation exists to help citizens make decisions with all the verified information in hand and so that they do not miss it in the battle against misinformation. It does this through journalism, technology, education and new narratives. Misinformation affects all strata of society and is present in our daily lives; The Foundation develops tools to combat it and generates information based on evidence so that the different actors involved, from legislators to content distribution platforms, including journalists, citizens and governments, have verified data to rely on.

4.2. Fake news and conspiracy theories: distinguishing conspiracy narrative from critical thinking

Abstract: The ease of generating content online has increased the amount of harmful information that is published. Disinformation is published mostly on social media and propagated on a daily basis. In this seminar I will try to stress the importance of going beyond the analysis of *(i)* words, *(ii)* textual information, and *(iii)* fake news. In order to do that we should: *(i)* integrate in the architecture of AI deep learning models emotional signals and psycholinguistics characteristics; *(ii)* address disinformation detection from a multilingual perspective; and *(iii)* consider that often fake news could be part of a conspiracy theory and a disinformation campaign. Related to the latter, it is important to be able to distinguish between conspiracy theories and critical thinking. A shared task on this topic will be organised in 2024 at PAN[2], both in Spanish and in English, with data from Telegram.

Paolo Rosso is a Full Professor at the Universitat Politécnica de Valéncia, where he is also a member of the Pattern Recognition and Human Language Technology (PRHLT) research centre. His research interests are focused on social media data analysis, mainly on fake news and hate speech detection, author profiling, and sarcasm detection.

He has published **50+ articles in journals (34 Q1)** and 400+ articles in conferences and workshops; he has an **H-index of 69** (source: Google Scholar) and he is in the **ranking of the top H-index scientists in Spain** (http://www.guide2research.com/scientists/ES). He has been **PI of several national and international research projects funded by EC, U.S. Army Research Office, Qatar National Research Fund, and Vodafone Spain**.

[2] https://pan.webis.de/shared-tasks.html

Currently, he is the **PI of the research project XAI-DisInfodemics** on eXplainable AI for disinformation and conspiracy detection during infodemics (Spanish Ministry of Science and Innovation), and of the Public Procurement with OBERAXE, the **Spanish *Observatory* on racism and xenophobia** of the Secretary of State for Migration. Moreover, he is a member of the **EC IBERIFIER** project on Monitoring the threats of disinformation (**European Digital Media *Observatory***), the project on Resources and Applications for Detecting and Classifying Polarized Hate Speech in Arabic Social Media (Qatar National Research Fund).

He has been **advisor of 26 PhD theses** and **actually he is the advisor of 8 PhD students**. He gave several keynotes (TSD-2020, CICLing-2019 etc.) and has helped **organising 30+ shared tasks at the PAN Lab at CLEF and FIRE evaluation forums, SemEval, IberLEF and Evalita** on topics such as author profiling (e.g. profiling bots, haters, and fake news spreaders), hate speech detection, irony detection, misogyny, sexism and toxic language identification, as well as of the MAMI shared task at SemEval 2022 on misogyny identification in memes. He has been the chair of *SEM-2015, and organised conferences in Valencia such as CERI-2012, CLEF2013, EACL-2017, and NLDB-2022. He helped as senior chair or track chair in conferences such as SIGIR, ACL etc.

Since 2014 he is **Deputy Steering Committee Chair of the CLEF Association**. He is also **Associate Editor at the Information Processing & Management journal**. He gave several tutorials on plagiarism detection at ICON-2010, author profiling at RuSSIR-2014, RANLP-2015, FIRE-2016 and CLiC-it-2018, and harmful information (fake news and hate speech) at CIKM-2020. During the last 10 years, the obtained results in plagiarism detection, irony detection, author profiling, and credibility detection (fake news) were covered by Spanish (El País, ABC, La Vanguardia, El Mundo, El Levante, El Confidencial, Radio Nacional de España, La Cope) and international media (Reforma, Informador, CNN-Español). In 2022 he received the **UPV Research Award in the category of Excellent Publication in Engineering and Technology** for his work on the automatic identification and classification of misogynistic language on Twitter.

5. Organizing Team

The NLP-MisInfo 2023 Organizing Team was composed of the following people with respect to their distinct roles:

- Two Co-chair Workshop Organizers;
- Fourteen Members of the Program Committee.

5.1. Co-chairs

Roberto Centeno is an Associate Professor at the Universidad Nacional de Educación a Distancia (UNED), Department of Languages and Informatics Systems (LSI), Madrid, Spain, where he has developed his teaching and research career since 2010. In 2012 he obtained his PhD in Computer Science from Rey Juan Carlos University, where he developed his doctoral thesis as an

FPI-MEC fellow from 2007 to 2010. In 2007 he obtained the Official Master's Degree in Information Technology and Computer Systems and since 2006 he is a Computer Engineer, both from Rey Juan Carlos University. He is currently a member of the Language Processing and Information Retrieval Research Group of the UNED, as well as the Center for Intelligent Information Technologies and their Applications (CETINIA) of the URJC.

In recent years, his research lines have focused on the areas of misinformation mitigation, fake news detection and stance on social networks, on reputation and trust mechanisms based on opinion systems. He is the author of around 20 JCR-indexed publications and conferences classified as highly relevant and relevant. According to Google Scholar, he has an h-index of 13 with over 470 citations. He has participated in various international and national research projects collaborating with several different institutions, focused on the application of artificial intelligence techniques to solve real-world problems. Web site: http://nlp.uned.es/~rcenteno/index.php

Rodrigo Agerri is a Ramon y Cajal Research Fellow (tenure-track) at IXA Group, part of the HiTZ Centre of the University of the Basque Country UPV/EHU, where he is head of the Text Analysis unit. He got a PhD in Computer Science at City, University of London (2007), and he has since been working on Natural Language Processing at several British and Spanish institutions, including a two-year stint in the industry as a research project director. He has been involved as PI or collaborator in more than 40 research projects funded by the European Commission, UK research councils, Spanish Ministry of Science and Basque Goverment and published in major journals (Artificial Intelligence, etc.) and conferences (ACL, EMNLP, EACL, IJCAI, etc.) related to Artificial Intelligence and Natural Language Processing.

Currently, his research is focused on Computational Semantics and Information Extraction, with a strong focus on multilingual and cross-lingual approaches. He was the creator and main developer of IXA pipes, a set of ready-to-use multilingual tools for linguistic processing. He is also PMC and committer in the OpenNLP project of the Apache Software Foundation. Web site: https://ragerri.github.io/

5.2. Program Committee

- **Óscar Araque**, GSI, Universidad Politécnica de Madrid (UPM)
- **Carlos Badenes-Olmedo**, Ontology Engineering Group (OEG), Universidad Politécnica de Madrid (UPM)
- **David Camacho**, Applied Intelligence & Data Analysis group, Universidad Politécnica de Madrid (UPM)
- **Jorge Carrillo-de-Albornoz**, NLP & IR, Universidad Nacional de Educación a Distancia (UNED)
- **Pablo Hernandez**, Maldita.es

- **Manuel Montes**, Laboratory of Language Technologies of the Computational Sciences Department (INAOE), México
- **Borja Lozano**, Newtral
- **Laura Plaza**, NLP & IR, Universidad Nacional de Educación a Distancia (UNED)
- **Anselmo Peñas**, NLP & IR UNED, Universidad Nacional de Educación a Distancia (UNED)
- **Álvaro Rodrigo**, NLP & IR, Universidad Nacional de Educación a Distancia (UNED)
- **Paolo Rosso**, PRHLT Research Center, Universitat Politècnica de València (UPV).
- **Fernando Sánchez**, GSI, Universidad Politécnica de Madrid (UPM)
- **Estela Saquete**, Natural Language Processing and Information Systems Group, University of Alicante
- **Mariona Taule Delor**, CLiC- The Language and Computation Center-CLiC, University of Barcelona

Acknowledgments

We would like to thank the authors of the submitted articles for their interest in the considered problem, the Keynote Speakers for the interest aroused in new research directions, and the members of the Program Committee for their valuable contribution to the success of the NLP-MisInfo 2023 Workshop.

Roberto Centeno would like to acknowledge DeepInfo project (PID2021-127777OB-C22) (MCIU/AEI/FEDER,UE) and the CHIST-ERA HAMiSoN project grant CHIST-ERA-21-OSNEM-002 by AEI PCI2022-135026-2 funded by the Spanish Research Agency (Agencia Estatal de Investigación).

Rodrigo Agerri would like to acknowledge the Basque Government (Research group funding IT-1805-22) and the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of making Europe; (ii) Disargue (TED2021-130810B-C21) and European Union NextGenerationEU/PRTR. Furthermore, Rodrigo Agerri is funded by the RYC-2017-23647 fellowship (MCIN/AEI/10.13039/501100011033 and ESF Investing in Your Future).

References

[1] R. N. Zaeem, C. Li, K. S. Barber, On sentiment of online fake news, in: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp. 760–767. doi:10.1109/ASONAM49781.2020.9381323.

[2] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Inf. Process. Manage. 57 (2020). URL: https://doi.org/10.1016/j.ipm.2019.03.004. doi:10.1016/j.ipm.2019.03.004.

[3] P. S. Kulkarni, R. T. Aghayan, L. Huang, S. Gupta, Misinformation detection in online content, 2020. US Patent App. 16/019,898.

[4] R. K. Kaliyar, N. Singh, Misinformation detection on online social media-a survey, 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (2019) 1–6. URL: https://api.semanticscholar.org/CorpusID:209695880.

[5] Q. Su, M. Wan, X. Liu, C.-R. Huang, Motivations, methods and metrics of misinformation detection: An nlp perspective, Natural Language Processing Research 1 (2020) 1–13. URL: https://doi.org/10.2991/nlpr.d.200522.001. doi:10.2991/nlpr.d.200522.001.

[6] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, 2020. arXiv:1811.00770.

[7] Q. Su, The Routledge Handbook of Chinese Applied Linguistics, 32, Routledge, London, UK, 2019, p. 16.

[8] A. Ihsan, M. Ayub, P. Shivakumara, N. Noor, Fake news detection techniques on social media: A survey, Wireless Communications and Mobile Computing 2022 (2022) 1–17. doi:10.1155/2022/6072084.

[9] N. R. de Oliveira, P. S. Pisa, M. A. Lopez, D. S. V. de Medeiros, D. M. F. Mattos, Identifying fake news on social networks based on natural language processing: Trends and challenges, Information 12 (2021). URL: https://www.mdpi.com/2078-2489/12/1/38. doi:10.3390/info12010038.

[10] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online., Science 359 (2018) 1146–1151. doi:10.1126/science.aap9559.

[11] S. Dungs, A. Aker, N. Fuhr, K. Bontcheva, Can rumour stance alone predict veracity?, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3360–3370. URL: https://aclanthology.org/C18-1284.

[12] S. Chen, L. Xiao, A. Kumar, Spread of misinformation on social media: What contributes to it and how to combat it, Computers in Human Behavior (2022) 107643.

[13] Y. Shen, C. T. Lee, L. Pan, C. Lee, Why people spread rumors on social media: developing and validating a multi-attribute model of online rumor dissemination, Online Inf. Rev. 45 (2021) 1227–1246. URL: https://doi.org/10.1108/OIR-08-2020-0374. doi:10.1108/OIR-08-2020-0374.

A. Online Resources

More information and materials about the 2023 Edition of the NLP-MisInfo Workshop can be found at the following URL: https://sites.google.com/view/nlp-misinfo-2023/