XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics

Paolo Rosso[1,2], Berta Chulvi[7], Damir Korenčić[1], Mariona Taulé[3], Xavier Bonet Casals[3], David Camacho[4], Angel Panizo[4], David Arroyo[5], Juan Gómez[6] and Francisco Rangel[7]

[1]PRHLT - Pattern Recognition and Human Language Technology, Universitat Politècnica de València (UPV)
[2]ValgrAI - Valencian Graduate School and Research Network of Artificial Intelligence
[3]Universitat de Barcelona (UB)
[4]Universidad Politécnica de Madrid (UPM)
[5]Consejo Superior de Investigaciones Científicas (CSIC)
[6]Universidad de Granada (UGR)
[7]Symanto Research, Spain

Abstract

The aim of the XAI-DisInfodemics project is to investigate strategies for fighting disinformation based on insights from social science. To address the challenges of disinformation, we need interdisciplinary collaboration and the development of tools that private and public entities can use. The developed tools need to address the problem of disinformation detection from an eXplainable Artificial Intelligence (XAI) perspective. We aim to counter disinformation and conspiracy theories on the basis of fact checking of scientific information. Moreover, our aim is to be able to explain not only the AI models in their decision-making but also the narratives that are employed to trigger emotions in the readers and make disinformation and conspiracy theories believable and propagate among the social networks users. We also focus on the important problem of distinguishing between conspiracies and texts which are simply critical and oppositional from a mainstream perspective. The final AI tool should help users to spot in documents those parts whose aim is to grab readers' attention by emotional appeals and that alert about a poor quality of the information. The tool is thought for the general public to improve users' digital literacy and its use will allow media and information platforms to be rated based on the quality of their health information.

Keywords

Infodemics, Disinformation Detection, Oppositional Thinking Analysis, Conspiracy Theories, Critical Thinking, COVID-19, Telegram

1. Motivation and Related Work

The problem of the automatic detection of disinformation and conspiracy theories has recently gained popularity [1, 2, 3, 4, 5, 6]. It is framed as a binary classification problem with fine-grained versions corresponding to multi-label or multi-class classification. However, the prevalent true vs. false paradigm runs into difficulties when dealing with conspiracy theories in everyday communication exchanges. Conspiracy Theories (CTs) are complex narratives that attempt to explain the ultimate causes of significant events as cover plots orchestrated by secret, powerful and malicious groups (for a review, see 7). Once the explanation regarding the agency of these groups has entered the public imaginary, these narratives are invoked in social media messages alongside a very small number of factual elements, making them difficult to be debunked by fact-checkers.

In addition to this lack of factual information, another challenging aspect of combating CTs with NLP models stems from the difficulty of distinguishing critical thinking from conspiratorial thinking in automatic content moderation. This distinction is vital because labeling a message as conspiratorial when it is only oppositional could drive those who were simply asking questions into the arms of the conspiracy communities. As several authors from social science suggest, a fully-fledged conspiratorial worldview is the final step in a progressive "spiritual journey" that started questioning social and political orthodoxies [8, 9].

This approach begs the question of what makes people pass from criticizing mainstream views to joining conspiracy communities. Phadke et al. [10] have recently

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.
prosso@dsic.upv.es (P. Rosso); berta.chulvi@symanto.com (B. Chulvi); damir.korencic@gmail.com (D. Korenčić); mtaule@ub.edu (M. Taulé); david.camacho@upm.es (D. Camacho); angel.panizo@upm.es (A. Panizo); david.arroyo@csic.es (D. Arroyo); jgomez@ugr.es (J. Gómez); francisco.rangel@symanto.com (F. Rangel)
0000-0002-8922-1242 (P. Rosso); 0000-0003-1169-0978 (B. Chulvi); 0000-0003-4645-2937 (D. Korenčić); 0000-0003-0089-940X (M. Taulé); xbonet@mac.com (X. B. Casals); 0000-0002-5051-3475 (D. Camacho); 0000-0002-2195-3527 (A. Panizo); 0000-0001-8894-9779 (D. Arroyo); 0000-0003-0439-3692 (J. Gómez); 0000-0002-6583-3682 (F. Rangel)

established that the ratio of dyadic interaction with current conspiracist users is the most important feature in predicting whether or not users join conspiracy communities, even after controlling for individual factors. This work has an essential implication: if models do not differentiate critical from conspiracist thinking, mindless censorship may push people toward conspiracy communities. Another important but still though neglected issue in the computational analysis of the conspiratorial texts is the role that these narratives play in intergroup conflict (for a recent review of intergroup conflict concept see 11). The increased involvement of conspiracist communities in political processes, including violence, suggests that the purpose of CTs is to enforce group dynamics and coordinate action [12]. Therefore, from a computational linguistic approach, we need to pay attention not only to the topics [6] but also to the elements of narrative relating to the intergroup conflict. This requires fine-grained span-level detection that has been used as an approach to other problems [13, 14] but, to the best of our knowledge, not yet in the domain of computational analysis of conspiracy theories.

2. Disinformation Detection in XAI-DisInfodemics

In the framework of the XAI-DisInfodemics project[1], several works have been published on disinformation detection. Disinformation was studied considering also the role that bots and trolls may have [15] and their polarisation dynamics [1]. False information in health was investigated in [16], also with respect to vaccines [17].

Disinformation detection was also addressed in [18] and semi-automated fact-checking through semantic similarity in [19]. **(author?)** [20] investigated the impact that psycholinguistic patterns may have in discriminating between disinformation spreaders and fact checkers. The correlation between false information spreaders and political bias were also investigated and a new dataset was provided [21].

Moreover, a widget was designed to analyse cloaked science [22] disinformation and content spread by bots [23]. Rumor and clickbait detection were addressed by combining information divergence measures and deep learning techniques [24], and multiplatform dynamics were investigated in order to study negationists on Twitter and Telegram [25].

[1]Grant PLEC2021-007681 funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR.

3. Conspiracy Theories Detection in XAI-DisInfodemics

The MediaEval 2022 FakeNews challenge [6] aimed to tackle the spread of COVID-19 conspiracy theories through tweets, encompassing three subtasks: identifying the stance of tweets towards conspiracy theories, detecting misinformation posters based on social network graphs, and an enhanced version of the first subtask incorporating graph data. The challenge utilized a dataset comprising 1,913 development and 830 test tweets/users, supplemented by a large user graph. Performance was measured using the Matthews correlation coefficient (MCC) [26].

Two teams composed of paper authors contributed their approaches to address the challenge's subtasks. The UPV team [27] focused on enhancing a transformer-based system with additional features, model ensembles, and GPT-3-augmented training data for Subtask 1, while exploring Graph Neural Networks (GNNs) for Subtasks 2 and 3. On the other hand, the UPM team [28] applied representational learning techniques to automatically discover relevant features from raw data for user classification, utilizing Node2vec, FastRP, Random Forest, and XGBoost algorithms.

Both teams demonstrated the effectiveness of their respective approaches, with UPV and UPM obtaining the best results in Subtask 1 and Subtask 2, respectively. In Subtask 1, the UPV team achieved an MCC of 0.738 [27], surpassing the second-best team's MCC of 0.710 [29]. For Subtask 2, the UPM team achieved an MCC of 0.459, outperforming the second-best team's MCC of 0.355 [29].

Building upon the experience from the MediaEval challenge, we proceeded to explore the capabilities of large language models (LLMs) for handling the task of conspiracy theory classification [30]. Our investigation utilized the same dataset to examine the zero-shot performance of GPT-3 in accurately classifying fine-grained, multi-label conspiracy theories. We also utilized the dataset to analyze the GPT's ability to interpret and utilize definitions effectively. We experimented with several types of definitions, including descriptive noun phrases and human-crafted definitions, and proposed methods for both generating definitions from examples and assessing GPT-3's comprehension of the definitions. The results demonstrate a positive correlation between the quality of class definitions and the zero-shot performance [30].

4. XAI-DisInfodemics Dataset

For the creation of XAI-DisInfodemics dataset, we first manually compiled a list of 2,273 public Telegram channels in **English** and **Spanish** that contain oppositional non-mainstream views on the COVID-19 pandemic. We

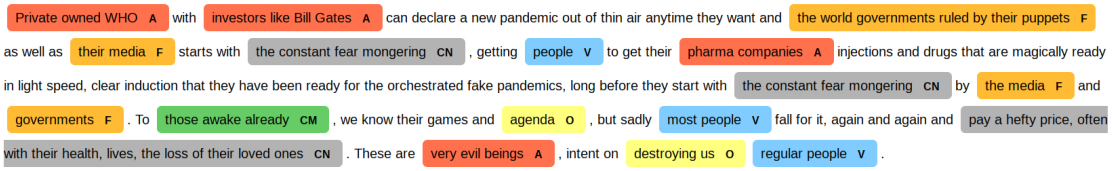

Figure 1: A Telegram text annotated with elements of oppositional narrative.

retrieved and filtered messages from the channels based on a set of oppositional and conspiracy keywords related to COVID-19. Then the messages were cleaned by removing duplicates, short texts, and texts with a large proportion of non-regular words (such as URLs and mentions). Finally, the messages were ranked using an index of quality based on the properties of a message and its channel. The index is composed of several criteria capturing the prevalence of COVID-19 topics and the channel's activity.

We developed an annotation schema to differentiate between the messages criticizing the mainstream views on COVID-19 and the messages evoking the existence of a conspiracy. A message was labeled "conspiracy" if any of these four criteria were met: (1) it framed COVID-19 or a related public health strategy as the result of the agency of a small and malevolent secret group; (2) it claimed that the pandemic is not real (e.g. a plandemic); (3) it accused critics of the conspiracy theory of being a part of the plot; (4) it divided society into two: those who know the truth (the conspiracy theorists) and those who remain ignorant. A message was labeled "critical" if it opposed publicly accepted understandings of events but had none of these four characteristics of the conspiratorial mindset.

Using this annotation scheme, 5,000 messages per language were annotated as "conspiracy" or "critical" thinking. For these messages we performed anonymization by removing sensitive and identifiable information such as nicknames, user IDs and e-mail addresses. The average text length is 128 tokens for Spanish texts and 265 tokens for English texts that tend to elaborate more on conspiracy theories.

Each message was annotated by three linguists and the inter-annotator agreement (IAA) was calculated. Disagreements were discussed with the social psychologist who created the annotation scheme. For English messages, the IAA in terms of Krippendorf's α is 0.79 for "conspiracy" messages and 0.60 for "critical" messages, while the average observed percentage of agreement between the three annotators is 91.4% , and 80.3%, respectively. For Spanish messages, Krippendorf's α is 0.80 for "conspiracy" messages and 0.70 for "critical" messages, corresponding the percentage agreements of 90.9% and 84.9%.

Moreover, a new fine-grained annotation scheme was developed with the goal of identifying, at the text span level, how oppositional and conspiracy narratives use intergroup conflict. The annotation was performed for the described 5,000 binary-labeled messages per language. Inspired by Lasswell's paradigm [31], we identify the following six categories of narrative elements at the span level (an example, with the abbreviations defined below, is displayed in Figure 1):

Agents (A): The hidden power that pulls the strings of the conspiracy. In critical messages, agents are actors that design the mainstream public health policies (Government, WHO, among others).

Objectives (O): Parts of the narrative that answer the question "what is intended by the agents of the CT or by the promoters of the action being criticized from a critical thinking perspective?"

Consequences (CN): Parts of the narrative that describe the effects of the agent's actions.

Facilitators (F): The facilitators are those who collaborate with the conspirators. In critical messages, facilitators are those who implement the measures dictated by the authorities.

Campaigners (CM): In conspiracy messages, the campaigners are the ones who uncover the conspiracy theory. In critical messages, campaigners are those who resist the enforcement of laws and health instructions.

Victims (V): Victims are the people who are deceived into following the conspiratorial plan or the ones who suffer due to the decisions of the authorities.

In the process of span-level annotation, each of the 5,000 Spanish and English messages was annotated by two linguists. Currently, the annotation instructions are being discussed and improved and, to this end, we are using the Gamma (γ) measure of the IAA test [32]. The preliminary annotation round (first 150 messages) yielded an average γ of 0.43. The following batch had an average gamma of 0.53, and the last one a γ of 0.61. We deemed this a good agreement because it is close to or above the average agreement of other highly conceptual span-level schemes [33, 34]. So far, 2,000 messages in both

languages have been fully annotated with an average density of six spans per message.

5. XAI-DisInfodemics Task at PAN

At the PAN Lab[2] we are organising a shared task on oppositional thinking analysis with the aim of addressing the following two new challenges for the NLP research community: (1) to distinguish the conspiracy discourse from other oppositional narratives that do not express a conspiracy mentality, and (2) to identify in online messages the key elements of a narrative that fuels the intergroup conflict in oppositional thinking. Accordingly, we propose two substasks:

Subtask 1 A binary classification task consisting of differentiating between (1) critical messages that question major decisions in the public health domain, but do not promote a conspiracist mentality; and (2) messages that view the pandemic or public health decisions as a result of a malevolent conspiracy by secret, influential groups.

Subtask 2 A token-level classification task aimed at recognizing text spans corresponding to the key elements of oppositional narratives. Since conspiracy narratives are a special kind of causal explanation, we developed a span-level annotation scheme that identifies the goals, effects, agents, and the groups in conflict.

In Subtask 1, model performance is equally important for both the "conspiracy" and the "critical" classes. Additionally, high-performance classifiers are desirable since errors in automatic content moderation can directly or indirectly promote the conspiracist mentality. To this end, we use MCC since the dataset is balanced, more reliable and less optimistic then the macro-averaged F1 [26], and compares favorably to other alternatives [35]. For Subtask 2, we will use an adaptation of the F1 measure suited for a sequence-labeling scenario with long and overlapping spans [33], which was applied in previous SemEval evaluation of systems for span-level propaganda annotation [13].

We already performed experiments with transformer-based baseline models for both subtasks. For Subtask 1 we used pre-trained BERT transformers and fine-tuned them for the binary tasks. This baseline yielded a MCC of 0.68 for Spanish and 0.79 for English texts. For Subtask 2, we experimented on currently annotated data using a pre-trained BERT model, with 6 token classification heads (one per category), and fine-tuning the model using multi-task learning. This approach yielded the results of 0.54 (English) and 0.45 (Spanish) in terms of the adapted F1 measure of **(author?)** [33]. The baseline results show that both tasks are feasible, although there is still room for improvement, especially for the challenging Subtask 2. We intend to motivate participants to use advanced classification techniques and architectures, with the goal of discovering most accurate solutions for the real-world deployment.

[2] https://pan.webis.de/clef24/pan24-web/oppositional-thinking-analysis.html

6. XAI-DisInfodemics App

In collaboration with Symanto[3], we are developing an application based on the research on detection and analysis of oppositional narratives. The application is envisioned as a tool that will enable users to determine whether a social media text in English or in Spanish contains elements of conspiratorial or critical narratives, and to detect fine-grained narrative elements. Target audience for the application are journalists, students and researchers in social sciences, as well as anyone interested in analysing and learning about oppositional narratives.

In addition to predict narrative categories, the application will highlight key parts of text for making the predictions, using XAI techniques[36]. This feature will facilitate the users' analysis of text, and increase the users' confidence in the AI model. The application will be easily accessible via a web browser, and feature an easy-to-use graphical user interface.

[3] https://www.symanto.com/es/

Acknowledgements

This work was carried out in the framework of Project XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (PLEC2021-007681) funded by MICIU/AEI/ 10.13039/501100011033 and by "European Union NextGenerationEU/PRTR".

References

[1] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language, Computer Science Review 47 100531. doi:10.1016/j.cosrev.2022.100531. URL https://www.sciencedirect.com/science/article/pii/S157401372200065X

[2] M. Gambini, S. Tardelli, M. Tesconi, The anatomy of conspiracy theorists: Unveiling traits using a comprehensive twitter dataset, Computer Communications 217 25–40. doi:10.1016/j.comcom.2024.01.027. URL https://www.sciencedirect.com/science/article/pii/S0140366424000264

[3] A. Giachanou, B. Ghanem, P. Rosso, Detection of conspiracy propagators using psycho-linguistic

characteristics, Journal of Information Science 49 (1) (2021) 3–17. doi:10.1177/0165551520985486. URL https://doi.org/10.1177/0165551520985486

[4] J. D. Moffitt, C. King, K. M. Carley, Hunting conspiracy theories during the covid-19 pandemic, Social Media + Society 7 (3) (2021). doi:10.1177/20563051211043212.

[5] K. Pogorelov, D. T. Schroeder, S. Brenner, J. Langguth, FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021, in: Working Notes Proceedings of the MediaEval 2021 Workshop Bergen, Norway and Online, 2021.

[6] K. Pogorelov, D. T. Schroeder, S. Brenner, A. Maulana, J. Langguth, Combining tweets and connections graph for fakenews detection at mediaeval 2022, in: Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023., 2023.

[7] K. M. Douglas, R. M. Sutton, What are conspiracy theories? a definitional approach to their correlates, consequences, and communication, Annual Review of Psychology 74 (1) (2023) 271–298. URL https://doi.org/10.1146/annurev-psych-032420-031329

[8] R. M. Sutton, K. M. Douglas, Rabbit hole syndrome: Inadvertent, accelerating, and entrenched commitment to conspiracy beliefs, Current Opinion in Psychology 48 (2022) 101462. doi:https://doi.org/10.1016/j.copsyc.2022.101462.

[9] E. Funkhouser, A tribal mind: Beliefs that signal group identity or commitment, Mind & Language 37 (3) (2022) 444–464. doi:https://doi.org/10.1111/mila.12326.

[10] S. Phadke, M. Samory, T. Mitra, What makes people join conspiracy communities? role of social factors in conspiracy engagement, Proc. ACM Hum.-Comput. Interact. 4 (CSCW3) (jan 2021). doi:10.1145/3432922.

[11] R. Böhm, H. Rusch, J. Baron, The psychology of intergroup conflict: A review of theories and measures, Journal of Economic Behavior & Organization 178 (2020) 947–962. doi:https://doi.org/10.1016/j.jebo.2018.01.020.

[12] P. Wagner-Egger, A. Bangerter, S. Delouvée, S. Dieguez, Awake together: Sociopsychological processes of engagement in conspiracist communities, Current Opinion in Psychology 47 (2022) 101417. doi:https://doi.org/10.1016/j.copsyc.2022.101417.

[13] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1377–1414. doi:10.18653/v1/2020.semeval-1.186.

[14] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, SemEval-2021 Task 5: Toxic Spans Detection, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 59–69. doi:10.18653/v1/2021.semeval-1.6.

[15] A. Giachanou, X. Zhang, A. Barrón-Cedeño, O. Koltsova, P. Rosso, Online information disorder: fake news, bots and trolls, International Journal of Data Science and Analytics 13 (4) 265–269. doi:10.1007/s41060-022-00325-0. URL https://doi.org/10.1007/s41060-022-00325-0

[16] I. B. Schlicht, E. Fernandez, B. Chulvi, P. Rosso, Automatic detection of health misinformation: a systematic review, Journal of Ambient Intelligence and Humanized Computingdoi:10.1007/s12652-023-04619-4. URL https://doi.org/10.1007/s12652-023-04619-4

[17] J. M. Noguera Vivo, M. d. M. Grandío-Pérez, G. Villar-Rodríguez, A. Martín, D. Camacho, Desinformación y vacunas en redes: Comportamiento de los bulos en twitter, Revista Latina de Comunicación Social (81) 44–62. doi:10.4185/RLCS-2023-1820. URL https://nuevaepoca.revistalatinacs.org/index.php/revista/article/view/1820

[18] J. Del Ser, M. N. Bilbao, I. Laña, K. Muhammad, D. Camacho, Efficient fake news detection using bagging ensembles of bidirectional echo state networks, in: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–7, ISSN: 2161-4407. doi:10.1109/IJCNN55064.2022.9892331. URL https://ieeexplore.ieee.org/document/9892331

[19] A. Martín, J. Huertas-Tato, A. Huertas-García, G. Villar-Rodríguez, D. Camacho, FacTeR-check: Semi-automated fact-checking through semantic similarity and natural language inference, Knowledge-Based Systems 251 109265. doi:https://doi.org/10.1016/j.knosys.2022.109265. URL https://www.sciencedirect.com/science/article/pii/S0950705122006323

[20] A. Giachanou, B. Ghanem, E. A. Ríssola, P. Rosso, F. Crestani, D. Oberski, The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers, Data & Knowledge Engineering 138 101960. doi:10.1016/j.datak.2021.101960. URL https://www.sciencedirect.com/science/article/pii/S0169023X21000835

[21] F. Sakketou, J. Plepi, R. Cervero, H. J. Geiss, P. Rosso, L. Flek, FACTOID: A new dataset for identify-

ing misinformation spreaders and political bias, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proc. of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, pp. 3231–3241.
URL https://aclanthology.org/2022.lrec-1.345

[22] D. Arroyo, S. Degli-Esposti, A. Gómez-Espés, S. Palmero-Muñoz, L. Pérez-Miguel, On the design of a misinformation widget (MsW) against cloaked science, in: S. Li, M. Manulis, A. Miyaji (Eds.), Network and System Security, Lecture Notes in Computer Science, Springer Nature Switzerland, pp. 385–396. doi:10.1007/978-3-031-39828-5_21.

[23] J. M. Camacho, L. Perez-Miguel, D. Arroio, WIDISBOT: Widget to analyse disinformation and content spread by bots, in: Proceedings of the 4th Conference on Language, Data and Knowledge, NOVA CLUNL, Portugal, pp. 514–519.
URL https://aclanthology.org/2023.ldk-1.55

[24] C. Oliva, I. Palacio-Marín, L. F. Lago-Fernández, D. Arroyo, Rumor and clickbait detection by combining information divergence measures and deep learning techniques, in: Proceedings of the 17th International Conference on Availability, Reliability and Security, ARES '22, Association for Computing Machinery, pp. 1–6. doi:10.1145/3538969.3543791.
URL https://doi.org/10.1145/3538969.3543791

[25] A. de Paz, M. Suárez, S. Palmero, S. Degli-Esposti, D. Arroyo, Following negationists on twitter and telegram: Application of NCD to the analysis of multiplatform misinformation dynamics, in: J. Bravo, S. Ochoa, J. Favela (Eds.), Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2022), Lecture Notes in Networks and Systems, Springer International Publishing, pp. 1110–1116. doi:10.1007/978-3-031-21333-5_110.

[26] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (1) (2020) 6. doi:10.1186/s12864-019-6413-7.
URL https://doi.org/10.1186/s12864-019-6413-7

[27] D. Korenčić, I. Grubišić, A. H. Toselli, B. Chulvi, P. Rosso, Tackling Covid-19 Conspiracies on Twitter using BERT Ensembles, GPT-3 Augmentation, and Graph NNs, in: Working Notes Proceedings of the MediaEval 2022 Workshop, 2023.
URL https://ceur-ws.org/Vol-3583/paper48.pdf

[28] A. G. Jiménez, A. Panizo-LLedot, J. Torregrosa, D. Camacho, Representational learning for the detection of COVID related conspiracy spreaders in online platforms, in: Working Notes Proceedings of the MediaEval 2022 Workshop, 2023.

[29] Y. Peskine, P. Papotti, R. Troncy, Detection of COVID-19-Related Conpiracy Theories in Tweets using Transformer-Based Models and Node Embedding Techniques, in: Working Notes Proceedings of the MediaEval 2022 Workshop, 2023.
URL https://ceur-ws.org/Vol-3583/paper46.pdf

[30] Y. Peskine, D. Korenčić, I. Grubisic, P. Papotti, R. Troncy, P. Rosso, Definitions matter: Guiding GPT for multi-label classification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 4054–4063. doi:10.18653/v1/2023.findings-emnlp.267.
URL https://aclanthology.org/2023.findings-emnlp.267

[31] H. D. Lasswell, Politics: Who Gets What, When, How, Whittlesey House, 1936.

[32] Y. Mathet, A. Widlöcher, J.-P. Métivier, The Unified and Holistic Method Gamma for Inter-Annotator Agreement Measure and Alignment, Computational Linguistics 41 (3) (2015) 437–479. doi:10.1162/COLI_a_00227.
URL https://doi.org/10.1162/COLI_a_00227

[33] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-Grained Analysis of Propaganda in News Articles, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5636–5646. doi:10.18653/v1/D19-1565.
URL https://aclanthology.org/D19-1565

[34] A. M. Weimer, F. Barth, T. Dönicke, L. Gödeke, H. Varachkina, A. Holler, C. Sporleder, B. Gittel, The (In-)Consistency of Literary Concepts. Operationalising, Annotating and Detecting Literary Comment, Journal of Computational Literary Studies 1 (1), number: 1 Publisher: Universitäts- und Landesbibliothek Darmstadt (Dec. 2022). doi:10.48694/jcls.90.
URL https://jcls.io/article/id/90/

[35] D. Chicco, N. Tötsch, G. Jurman, The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Mining 14 (1) (2021) 13. doi:10.1186/s13040-021-00244-z.

[36] A. Madsen, S. Reddy, S. Chandar, Post-hoc interpretability for neural nlp: A survey, ACM Comput. Surv. 55 (8) (dec 2022). doi:10.1145/3546577.