DEMON

Detect and Evaluate Manipulation of ONline information

Deliverable number: **WP4**

Integration of content- and network-based approaches to measure manipulation of online information (M18–M24)

Università degli Studi di Cagliari

POLITECNICO MILANO 1863

Project Acronym:	DEMON
Project Title:	Detect and Evaluate Manipulation of ONline information
Project Type:	PRIN 2022
Project Code:	2022BAXSPY
Project Website:	`https://demon.unica.it/`
Workpackage:	**WP4:** Integration of content- and network-based approaches to measure manipulation of online information
Deliverable Type:	Report (R)
Dissemination Level:	Public (PU)
Authors:	Maurizio Atzori, Cecilia Di Ruberto, Andrea Loddo, Davide Mura, Manuela Sanguinetti, Marco Usai (UNICA), Emanuele Della Valle, Francesco Pierri (POLIMI)
Delivery Date:	March, 2026

Contents

1 Introduction to WP4

The following contains the contributions of the DEMON project falling into the area of Work Package 4. During this period, WP4 activities concentrated on advancing narrative modeling and content analysis through several key initiatives:

- UNICA Unit: During the project's final phase, the team consolidated previous research lines into an integrated framework for analyzing persuasive communication (Section 2). A Transformer-based model was engineered to identify persuasive techniques and perform span detection to highlight specific influential words. Moving beyond binary classification, this model supports a "persuasion score" that quantifies intensity along a continuum. To explore mitigation, the team experimentally verified the ability of Large Language Models (LLMs) to "sanitize" content—rewriting persuasive text into neutral versions while preserving the original semantic meaning. Finally, the project expanded into Italian-language social media analysis, developing a pipeline to detect fallacies (Section 3). This system addressed data scarcity and class imbalance through data augmentation and feature extraction. To optimize performance, the team benchmarked Fine-tuning against Retrieval-Augmented Generation (RAG) across several state-of-the-art models, including Mixtral-8x7B, Gemma-3-12B, and GPT-5.
- POLIMI Unit: WP4 represented the consolidation and integration phase of the POLIMI activities. Building on the network-based analyses developed in WP3, this work package focused on combining content-based signals from WP2 with network-level indicators such as interaction structures, coordination patterns, and temporal dynamics. In this perspective, the empirical evidence previously gathered on emerging platforms such as Bluesky served as a foundation for identifying robust indicators of manipulation, amplification, and anomalous diffusion (Section 4). WP4 extended this line of work by systematizing such signals within integrated analytical pipelines and by assessing how they can support the interpretation of manipulation phenomena across different application scenarios, including misinformation and harmful content. Through performance evaluations and ablation studies, POLIMI identified the most effective combinations of content and network features. These findings culminated in detailed use-case reports that allow users to explore manipulation through dimensions such as actor involvement, coordination, and amplification intensity (Section 5).

The next sections will provide further details on these activities and the results obtained.

2 Calculating the Persuasion Score

When we read a newspaper article or an online post about polarising topics such as politics or religion, an unwary reader may fall victim to manipulative textual content. This type of content aims to persuade the reader to change their beliefs. To achieve this, specific persuasion techniques known as logical fallacies are used. Some appeal to emotion, while others shift the focus to the speaker rather than the argument (*ad hominem)*. In order to develop greater critical thinking skills and preserve the integrity of public discourse, it is necessary to use tools that can recognise these techniques and alert the reader.

Motivated by this context, our work is based on two main questions: is it possible to measure the gradient of persuasion within a text? Furthermore, once detected, is it possible to remove it or at least reduce it?

With the aim of answering these questions, a deep learning model based on XLM-RoBERTa was developed to analyse texts from journalistic sources for the recognition of persuasion techniques. Based on this recognition, a calculation was devised for a score called a persuasion score, which could assign a value to a generic text representing the amount of persuasion present in the text.

To validate the persuasion score, a pipeline was created using the Persuasive Pairs dataset [5], which contains a pair of texts for each instance, one of which is more persuasive than the other.

Finally, in order to answer the second question, a Large Language Model was used to rephrase the most persuasive texts and recalculate their persuasive score after rewriting, to verify whether there had been a reduction or elimination of persuasiveness.

2.1 Approach Overview

The overall framework consists of three main steps, briefly outlined below and illustrated in Figure 1:

- Persuasion Scoring: the input text is processed by a Transformer-based language model fine-tuned for identifying persuasion techniques (e.g., appeals to emotion, loaded language, etc.). The model's output is then passed to an additional module that computes an indicator representing the degree of persuasion within the text, referred to as the Pscore.

- Validation: the reliability of the Pscore in measuring a text's level of persuasion is evaluated by applying it to a dataset of sentence pairs. This dataset includes human annotations, which serve as the ground truth for validation.

- Text Sanitization: a large language model (LLM) rewrites the original text into a "sanitized" version, meaning that persuasive elements are minimized or removed. The resulting text is thus more neutral and objective compared to the original.

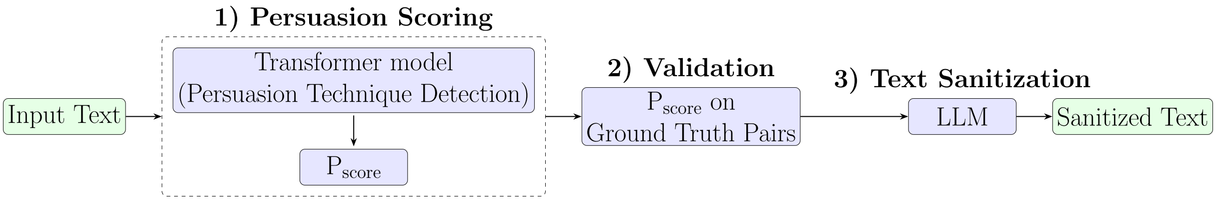

Figure 1: Summary of the overall framework.

2.2 Persuasion Scoring

To address this task, the first stage was divided into two main sub-steps:

- developing a deep learning model capable of identifying potential persuasion techniques within a text;

- using the trained model to compute the degree of persuasion of a given text, hereafter referred to as the Pscore.

Persuasion Technique Detection The approach employed three state-of-the-art pre-trained language models (BERT, T5, and RoBERTa) to detect linguistic patterns linked to persuasive reasoning. The task was divided into two subtasks: (1) binary sentence classification, determining whether a sentence contains persuasion, and (2) binary token classification, labeling individual words involved in persuasive expressions.

A multi-task learning framework was adopted to integrate both sentence-level and token-level insights, enhancing contextual understanding of persuasion. All models were fine-tuned using the SemEval 2023 Task 3 dataset [6] on detecting persuasion techniques in multilingual online news.

The dataset includes annotations for 23 persuasion techniques grouped into six categories (e.g., Attack on Reputation, Simplification, Manipulative Wording) and provides span-level annotations, identifying the exact words forming persuasive elements. For this study, only the English portion of the dataset was used.

Figure 2 shows the unified multi-task architecture that enables simultaneous sentence classification and token classification within a single forward pass. Token embeddings are extracted from the base Transformer model, with a special token added to represent the entire sentence. This design enables the model to produce both a global sentence-level prediction and a fine-grained token-level analysis to detect tokens involved in fallacy techniques.

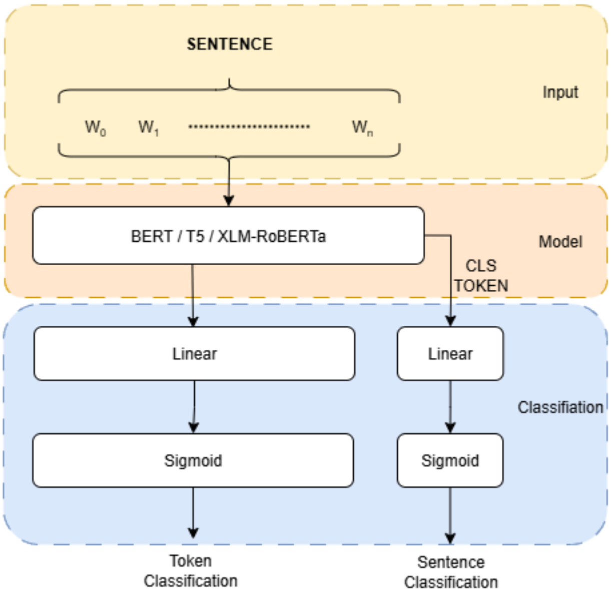

Figure 2: Overview of the multi-task model architecture (sentence classification and token classification). The input part is represented by the sentence, where W_0, ...,W_n correspond to the words in the sentence. The model part represents one of the three models used (BERT, T5, RoBERTa). Finally, the classification part returns token classification and sentence classification.

Pscore To address the first research question, whether it is possible to measure the quantity of a persuasive text, we introduce the persuasive score (or simply P_{score}), a metric derived from the model's output that assigns a numerical value between 0 and 1 to represent the gradient of manipulative or persuasive content.

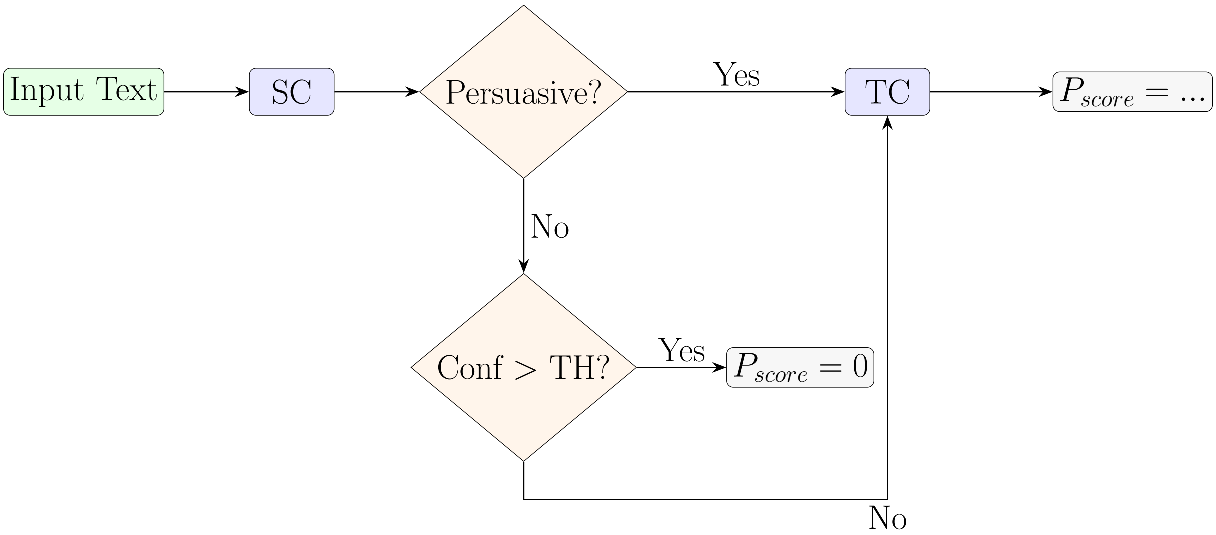

Figure 3: Flow chart of the persuasion scoring pipeline. SC and TC denote Sentence and Token Classification, respectively, while TH represents the threshold value of the confidence score from the sentence classification.

To validate this metric, the Persuasive Pairs dataset [5] was employed. It contains 2697 pairs of semantically similar short texts, with one text in each pair judged as more persuasive by three human annotators. The texts were drawn from multiple sources PT-Corpus [6], Webis-Clickbait-17 [7], Winning-Arguments [11], PersuasionForGood [9], and ElecDeb60to20 [3] and rewritten by LLMs to amplify or reduce their persuasive strength. Each pair received a human-assigned persuasiveness score ranging from -3 to 3, indicating the degree and direction of change in persuasiveness.

Using this dataset, a dedicated pipeline, shown in Figure 3, was designed to identify the more persuasive text in each pair based on the computed P_{score}. Once persuasive words are detected by the multi-task model, information from both token-level and sentence-level classifications is merged. The number of persuasive words is normalized by the total number of words in the sentence, and an ϵ value, derived from the sentence-level output, is included to handle cases where no specific tokens are flagged, but the sentence is still globally persuasive. This ensures a consistent and reliable scoring mechanism.

$$P_{score} = \max\left(\epsilon,\ \frac{\#\text{pers\_words}}{\#\text{tot\_words}}\right) \tag{1}$$

In Equation (1), $\#pers\_words$ denotes the number of words identified as persuasive, $\#tot\_words$ the total number of words, and ϵ represents the baseline from the sentence classification. The RoBERTa-based model, which achieved the best results in persuasion technique detection (Section 2.5), was ultimately used to compute the final P_{score}.

2.3 Validation

To validate the reliability of the P_{score}, a dedicated pipeline was designed to assess persuasion within pairs of sentences from the Persuasive Pairs dataset. For each pair $(T_{i,0}, T_{i,1})$, the P_{score} is computed using the Persuasion Scoring module. The sentence with the higher score, $T_{i,MAX}$, is then identified and compared against the human-annotated ground truth to evaluate the model's accuracy (see Figure 4).

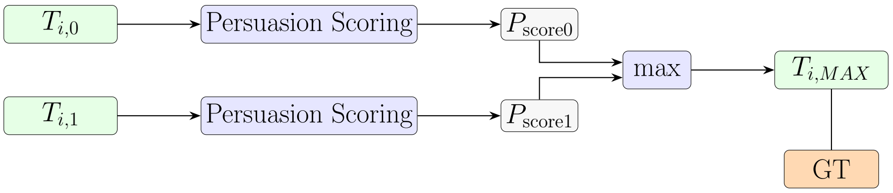

Figure 4: Flow chart of the Validation step for each i^{th} pair of the dataset.

2.4 Text Sanification

After establishing a method to quantify a text's persuasiveness through a numerical score and validate its effectiveness on unseen data, the next step explored whether a persuasive text could be "sanitized". Following the approach of [8], an LLM was employed to rewrite persuasive sentences into more neutral versions. Using the Persuasive Pairs dataset, the most persuasive sentence from each pair, identified through the procedures in Section 2.2 and 2.3, was reformulated by the LLM to reduce or remove its persuasive elements. The Qwen3 LLM was employed for this task, tested in four parameter sizes (0.6B, 1.7B, 4B, and 8B) to evaluate performance across lighter and larger models. To ensure deterministic generation, the temperature was set to 0, the random seed to 42, and the maximum output length to 2048 tokens. Additional chat options, such as add_generation_prompt (to mark response start) and enable_thinking (to activate reasoning), were enabled. The model was then prompted to rephrase each persuasive sentence.

2.5 Results

This section presents the results obtained from the different tasks conducted. First, the outcomes of the Persuasion Technique Detection task are described, followed by the results from the persuasive score validation. Finally, the findings of the sanitization task are presented.

Table 1 presents the evaluation results for all models across both tasks, using accuracy, macro-F1, micro-F1, and Persuasion-F1 (focused on the persuasion class due to dataset imbalance). Overall, sentence classification outperformed token classification, reflecting the lower complexity of sentence-level predictions. In contrast, token-level performance declined for all models, especially in the Persuasion-F1 metric, indicating greater difficulty in detecting persuasive tokens. Among the tested models, BERT-based architectures surpassed T5, which is less suited for classification tasks. RoBERTa achieved the best overall performance across both tasks and was therefore chosen for subsequent experiments.

The pipeline described in Section 2.3 was evaluated using the Persuasive Pair dataset. Table 2 displays the confusion matrix corresponding to the classification of Text 1 and Text 2 within each dataset pair. According to these results, the model misclassified 341 cases, 180 involving Text 1 and 161 involving Text 2. Overall, the pipeline's predictions aligned with human annotations in 79% of cases.

Finally, the sanitization task was evaluated using standard natural language generation metrics. The main metric, ROUGE, measured lexical overlap between the original (persuasive) and rewritten (sanitized) texts to assess syntactic similarity and content preservation. Specifically, ROUGE-1 captured unigram overlap, while ROUGE-L used the Longest Common Subsequence to evaluate structural consistency.

Additionally, two complementary metrics were adapted:

Task	Model	Accuracy	Macro-F1	Micro-F1	Persuasion-F1
Sentence	BERT	0.80	0.77	0.80	0.84
	RoBERTa	**0.83**	**0.80**	**0.82**	**0.86**
	T5	0.77	0.75	0.77	0.81
Token	BERT	0.77	0.65	0.76	0.44
	RoBERTa	**0.80**	**0.68**	**0.80**	**0.50**
	T5	0.67	0.56	0.66	0.35

Table 1: Classification metric scores are reported for each task. The first row presents the results for the sentence classification task, while the second row refers to the token classification task. We highlight the best result for each task and metric in bold.

		Prediction	
		Text 1	Text 2
Actual	Text 1	891	245
	Text 2	235	877

Table 2: Confusion matrix of the most persuasive text classification. Text 1 is the first text, and Text 2 is the second one.

- Style Transfer Accuracy (STA): assessed whether the sanitized text showed a reduced persuasion level compared to the original, based on sentence-level classification scores.
- Semantic Similarity: measured how closely the sanitized text preserved the original meaning, computed via LaBSE embeddings using cosine similarity.

A combined score, obtained by multiplying STA and semantic similarity, captured both stylistic change and content preservation. The Pscore was also computed before and after sanitization to verify reduction in persuasiveness.

Results Table 3 show that the LLM successfully preserved meaning (high cosine similarity) while reducing persuasion, even with the smallest 0.6B-parameter model. Larger models further decreased the persuasive score by up to 5%, though overall performance remained comparable. Finally, between 70% and 80% of sentences showed a reduction in Pscore after sanitization.

3 Experiments on Fallacy Detection with an Italian Benchmark

In order to explore possible content-based approaches on the Italian scenario – thus departing from a purely anglo-centric view typically encountered in tasks addressing online information manipulation – we conducted a study on fallacy detection as part of the FadeIT shared task at EVALITA 2026[1], which precisely focuses on identifying fallacies

[1] https://sites.google.com/fbk.eu/fadeit2026

	ROUGE		**Sanitization Metrics**			**P$_{score}$**
Model	R-1 F1	R-L F1	Cosine Sim	STA	Joint	
Qwen0.6B	**0.557**	**0.513**	0.792	**0.718**	0.570	0.151
Qwen1.7B	0.446	0.400	0.743	0.714	0.531	0.049
Qwen4B	0.524	0.488	**0.806**	0.715	**0.576**	0.051
Qwen8B	0.498	0.461	0.797	0.713	0.569	**0.048**
Original Texts						0.357

Table 3: Evaluation metrics for the Qwen models on the sentence sanitization task. The last row shows the overall P$_{score}$ computed over the original data.

within Italian social media posts. We developed two systems to address this multi-label classification task using state-of-the-art Large Language Models (LLMs). Specifically, we implemented a comprehensive pipeline consisting of sophisticated pre-processing and distinct technical approaches (see what presented in [2]).

Pre-processing Pipeline To address issues like class imbalance, where some fallacies appeared in less than 2% of the dataset, we employed several strategies:

- Data Augmentation with Prompt Engineering: GPT-5.1 was used to generate 50 new examples for each minority fallacy class.
- Data Augmentation with Back-translation: Posts were translated from Italian into Spanish or French and then back to Italian using Helsinki-NLP models to create semantically equivalent paraphrases.
- Feature Extraction: GPT-5 was instructed to act as a rhetorical feature detector, extracting 11 specific descriptors (e.g., *unsupported_ claim*, *generalization*, *profanity*) from each post to serve as additional context for classification.

Approach 1: RAG and Dynamic Few-Shot Prompting This method utilized Retrieval-Augmented Generation (RAG) to enhance model performance: The training dataset was stored in a Qdrant vector database using OpenAI's text-embedding-3-small model. For each new post, the system retrieved the top five most similar examples from the database based on cosine similarity. The final prompt sent to GPT-5.1 included the problem definition, fallacy descriptions, retrieved examples, and the pre-extracted rhetorical features to guide the model's reasoning.

Approach 2: Supervised Fine-Tuning The second approach involved fine-tuning open-source models to specialize them for fallacy detection.We used QLoRA (Quantized Low-Rank Adaptation) to enable parameter-efficient training on a single A100 GPU. A custom loss function incorporated label weights to balance the influence of rare fallacy classes during training. We experimented with Mixtral-8x7B-Instruct, Gemma-3-12B-it, and Meta-Llama-3.1-70B-Instruct.

Results and discussion The final results of the FadeIT subtask A (fallacy detection in Italian social media) show that the RAG-based approach using GPT-5 was the most effective, outperforming the models optimized through supervised fine-tuning.

The approach using GPT-5 with dynamic few-shot prompting via RAG achieved the highest F1-score. While the RAG approach was the most balanced, Gemma-3 achieved the highest precision (58.70) among the three runs but recorded the lowest F1-score due to its significantly lower recall (38.44). Interestingly, the fine-tuned models (Mixtral and Gemma) initially showed superior performance during the development phase (with Mixtral reaching a 0.64 F1-score). However, their performance declined on the official test set, eventually falling behind the GPT-5 RAG system. All final scores remained below 50%, highlighting the intrinsic difficulty of the task, which involved classifying posts into 20 different fallacy categories.

As regards in particular the reduced effectiveness of the fine-tuning approach, the performance drop suggests a potential issue with the distribution of fallacies between the training and test datasets. If the test set contained a different balance of fallacies or used them in different contexts than the training data, the fine-tuned models—which were optimized specifically on the training data—would likely struggle to generalize. Furthermore, certain fallacies, such as *circular reasoning*, *strawman*, or *cherry picking*, are difficult for transformer-based models to grasp immediately. These require complex reasoning capabilities to identify. While fine-tuning helps a model specialize, the RAG-based approach leveraged GPT-5, a model with significantly higher baseline reasoning power and "enhanced reasoning capabilities" compared to the smaller open-source models. In addition, despite data augmentation efforts, the dataset remained characterized by a suboptimal distribution where some fallacies appeared in less than 2% of the posts. Fine-tuning on such imbalanced data can often lead to models that are less robust when encountering the rare classes in a live test environment. The RAG approach, instead, used dynamic few-shot prompting, meaning it selected the five most semantically similar examples from a vector database for every specific post it had to classify. This provided the model with highly relevant, "on-the-fly" context that the fine-tuned models (which rely on static weights learned during training) did not have.

4 From WP3 evidence to integrated indicators: the Bluesky case

The Bluesky case study [4, 10] played a bridging role between WP3 and WP4. In WP3, we investigated the early evolution of Bluesky around its public opening, with particular attention to user growth, posting behavior, suspicious activity, and the circulation of low-credibility information. WP4 builds on that empirical foundation by using the Bluesky case as a testbed for consolidating network-based evidence into a broader framework for measuring manipulation through the joint interpretation of content-based and structural indicators.

The results obtained in the previous work package showed that rapid platform expansion may generate abrupt changes in activity and interaction volumes that do not necessarily correspond to manipulation, but that may nonetheless create favorable conditions for opportunistic or coordinated amplification. This makes it essential to interpret activity bursts in light of additional dimensions, rather than relying on volume alone.

A second important result concerned the distribution of low-credibility information.

While such content represented only a limited portion of the overall information ecosystem, its spread was highly concentrated among a small number of accounts. This suggests that manipulation-related phenomena may be better captured through concentration, amplification, and actor-level asymmetry indicators than through aggregate prevalence alone.

Finally, the Bluesky analyses highlighted the importance of combining multiple perspectives, including political leaning, community structure, conversational toxicity, and moderation outcomes. Taken together, these dimensions show that potentially manipulative behavior should be interpreted as a multi-layered phenomenon, emerging from the interaction between content, actors, network structure, and platform governance.

In this sense, the Bluesky study should be seen not as a separate contribution from WP3, but as an empirical predecessor whose findings were consolidated in WP4 into a more general analytical perspective. These results informed the design of integrated pipelines and dashboards aimed at supporting the exploration of manipulation through dimensions such as actor involvement, coordination, amplification intensity, and exposure to harmful or low-credibility content.

5 Modeling the Impact of Online Misinformation on Real World Epidemics

This work [1] develops a data-informed, multi-level modeling framework to quantify how exposure to online misinformation can amplify epidemic spread through behavioral changes such as reduced compliance with public health guidance and increased vaccine hesitancy. The approach explicitly couples two distinct but interdependent layers: an *information diffusion network* capturing how low-credibility content spreads on social media, and a large-scale *physical contact network* capturing opportunities for disease transmission. By integrating these layers, the study moves beyond purely theoretical or stylized simulations and provides quantitative bounds on the potential harms associated with misinformation exposure.

The information layer is grounded in a large corpus of English-language discussions about COVID-19 vaccines on Twitter collected over approximately nine months in 2021. Users are geolocated to U.S. counties, political alignment is inferred from shared news domains (and propagated through the retweet network when direct domain evidence is missing), and misinformation exposure is operationalized at the source level by labeling tweets that link to low-credibility outlets. These data are used to construct a directed, weighted retweet network, where edge weights represent repeated retweeting and serve as a proxy for diffusion intensity and potential exposure.

To model how exposure can convert additional users beyond those who directly share low-credibility links, the study adopts a threshold-based complex contagion process. A single parameter ϕ represents "resilience" to misinformation: as ϕ increases, individuals require more exposures through their neighbors to become misinformed, whereas low values correspond to highly vulnerable settings (including the simple contagion case $\phi = 1$). Varying ϕ generates misinformed subpopulations of different sizes that are empirically anchored in observed diffusion patterns, enabling systematic comparisons across best- and worst-case misinformation scenarios.

The physical contact layer is constructed as a large, mobility-informed network with approximately $N \approx 20$ million nodes (a 10% sample), designed to reproduce key empirical constraints at the county level. County samples are calibrated to match population size and ideological composition using voting records, while the fraction of misinformed individuals in each county is inherited from the information diffusion layer. Edges in the contact network are generated using aggregated cell-phone mobility data, yielding expected contact probabilities within and between counties and producing a network with target average degree (set to a pre-pandemic contact rate). This construction captures geographic structure in physical mixing while preserving ideological and misinformation-related heterogeneity across locations.

Disease transmission is simulated on the contact network using an extension of the classical SIR model that distinguishes between ordinary susceptible individuals and a misinformed susceptible subpopulation (SMIR). Misinformed individuals are assumed to adopt riskier behaviors (e.g., lower adherence to masking, distancing, vaccination), modeled through extreme transmission parameters to bound outcomes: a worst-case scenario in which misinformed individuals have very high transmission probability, versus a best-case scenario where ordinary individuals have very low transmission probability. Agent-based simulations initialized with a small set of infected seeds show that larger misinformed subpopulations (lower ϕ) lead to earlier and higher epidemic peaks and substantially greater cumulative infections. Quantitatively, the worst-case setting produces a peak infection level roughly six times larger and occurring about two weeks earlier than the most resilient scenario, with an additional ∼14% of the population infected over the course of the epidemic in the worst case.

Overall, the study provides a principled mechanism to translate misinformation exposure—inferred from social media diffusion data—into measurable amplification of epidemic outcomes on a realistic contact network. By bounding best- and worst-case scenarios and testing robustness to alternative parameterizations (including mean-field approximations and sensitivity analyses), the framework offers a reusable template for assessing how misinformation-driven behavioral heterogeneity can propagate risk to the broader population, and for informing policy discussions on the societal costs of health misinformation.

References

Publications Acknowledging the DEMON Project

[1] Matthew R DeVerna, Francesco Pierri, Yong-Yeol Ahn, Santo Fortunato, Alessandro Flammini, and Filippo Menczer. "Modeling the amplification of epidemic spread by individuals exposed to misinformation on social media". In: *npj Complexity* 2.1 (2025), p. 11.

[2] Matteo Fenu and Maurizio Atzori. "Unica at FadeIT: Adapting Large Lanuage Models to Fallacy Identification in Social Networks". In: *Proceedings of the 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026)*. Ed. by Franco Cutugno, Alessio Miaschi, Alessio Palmero Aprosio, Giulia Rambelli, Lucia Siciliani, and Marco Antonio Stranisci. Bari, Italy: CEUR-WS.org, Feb. 2026.

[4] Gianluca Nogara, Erfan Samieyan Sahneh, Matthew R DeVerna, Nick Liu, Luca Luceri, Filippo Menczer, Francesco Pierri, and Silvia Giordano. "A longitudinal analysis of misinformation, polarization and toxicity on Bluesky after its public launch". In: *Online Social Networks and Media* 51 (2026), p. 100342.

[10] Erfan Samieyan Sahneh, Gianluca Nogara, R DeVerna Matthew, Nick Liu, Luca Luceri, Filippo Menczer, Francesco Pierri, Silvia Giordano, et al. “The Dawn of Decentralized Social Media: An Exploration of Bluesky’s Public Opening”. In: *LECTURE NOTES IN COMPUTER SCIENCE* (2024), pp. 406–421.

Other References

[3] Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. “Argument-based Detection and Classification of Fallacies in Political Debates”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 11101–11112. DOI: 10.18653/v1/2023.emnlp-main.684. URL: https://aclanthology.org/2023.emnlp-main.684/.

[5] Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. “Measuring and Benchmarking Large Language Models’ Capabilities to Generate Persuasive Language”. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 10056–10075. ISBN: 979-8-89176-189-6. DOI: 10.18653/v1/2025.naacl-long.506. URL: https://aclanthology.org/2025.naacl-long.506/.

[6] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. “SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup”. In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2343–2361. DOI: 10.18653/v1/2023.semeval-1.317. URL: https://aclanthology.org/2023.semeval-1.317/.

[7] Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. “Crowdsourcing a Large Corpus of Clickbait on Twitter”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1498–1507. URL: https://aclanthology.org/C18-1127/.

[8] Daniel Russo, Oscar Araque, and Marco Guerini. “To Click It or Not to Click It: An Italian Dataset for Neutralising Clickbait Headlines”. In: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-It 2024)*. Ed. by Felice Dell’Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli. Pisa, Italy: CEUR Workshop Proceedings, 2024, pp. 829–841. URL: https://aclanthology.org/2024.clicit-1.90/.

[9] Saumajit Saha, Kanika Kalra, Manasi Patwardhan, and Shirish Karande. “Performance of BERT on Persuasion for Good”. In: *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*. Ed. by Sivaji Bandyopadhyay, Sobha Lalitha Devi, and Pushpak Bhattacharyya. National Institute of Technology

Silchar, Silchar, India: NLP Association of India (NLPAI), Dec. 2021, pp. 313–323. URL: https://aclanthology.org/2021.icon-main.38/.

[11] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. "Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 613–624. ISBN: 9781450341431. DOI: 10.1145/2872427.2883081. URL: https://doi.org/10.1145/2872427.2883081.