DEMON

Detect and Evaluate Manipulation of ONline information

Deliverable number: **WP1**

Data collection, clustering of stories and categorization of sub-narratives (UNICA, POLIMI) M1-M9

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:	**WP1:** Data collection, clustering of stories and categorization of sub-narratives
Deliverable Type:	Report (R)
Dissemination Level:	Public (PU)
Authors:	Maurizio Atzori, Cecilia Di Ruberto, Andrea Loddo, Davide Mura, Manuela Sanguinetti, Marco Usai (UNICA); Stefano Ceri, Francesco Pierri (POLIMI)
Delivery Date:	September 15, 2024

Contents

1 Introduction to WP1

This work package focuses on the systematic identification, collection, analysis, and dissemination of data related to online news content, social media posts, and emerging narratives. The primary goal is to build a comprehensive dataset that will serve as a resource for analyzing the dynamics between news dissemination and public discourse, as well as to develop methodologies for grouping and categorizing content based on emerging stories and sub-narratives.

The first set of tasks involves identifying online news websites by leveraging existing third-party lists that assess news sources based on credibility and political bias (e.g., MediaBiasFactCheck, Iffy, Newsguard). This ensures that the news sources included in the study are rigorously vetted, representing a spectrum of political leanings and reliability. The subsequent step focuses on collecting news articles from these websites, creating a repository of media content that can later be used for studying how different stories evolve, spread, and gain traction across various parts of the political spectrum.

The second component of the work involves identifying and collecting social media posts from influential actors, including politicians, public officials, and prominent users. Twitter and Facebook, as primary platforms for public political discourse, will be the focus for gathering social media data. By cross-referencing the actors with their corresponding social media accounts, this task aims to capture the narratives and opinions that may influence or reflect public opinion, often complementing or contesting news media narratives [1, 3].

The next critical task is the development of clustering techniques to group related pieces of information (news articles, social media posts) into coherent stories. This clustering process allows for the construction of broader narratives, which are then categorized into sub-narratives, revealing deeper trends and patterns. These techniques will be pivotal for understanding how stories evolve and fragment into different interpretations across platforms.

Finally, this work package includes a dissemination phase where the collected data and methodologies will be shared with the broader academic community through open repositories (e.g., GitHub, DataVerse, Zenodo). By making these datasets publicly accessible, along with accompanying dataset papers in academic venues, the project aims to encourage further research and collaboration, contributing to a more informed and transparent understanding of news ecosystems and social media's role in shaping public narratives.

1.1 Contribution and Organization

The following contains the contributions of the DEMON project falling into the area of Work Package 1 described above. In particular:

- A new dataset on social media conversation on the 2022 Italian Elections has been collected and pulished (Section 2)
- An analyisis of user activities of the Bluesky social media platform has been performed (Section 3)
- Effects on bias of LLMs on Political news articles have been studied (Section 4)

- A thorough research on existing dataset relevant to DEMON, both textual, visual and multi-modal have been accomplished (Section 5
- A relevant testbed and challenge focusing on finding memes, a relevant technique to promote/spread propaganda, has been identified (Semeval Task 4) and a novel multi-modal model (Themis) have been developed with promising results (Section 6)
- An initial study on classifying fake news has been conduced (Section 7)

The deliverable is organized as follows: in Sect.2 a new dataset on social media conversation has

2 Social media conversations related to the 2022 Italian general election

We collected a unique, multi-platform dataset [10] designed to capture Italian-language political conversations surrounding the 2022 Italian general election, held on September 25th. The dataset, available in a public GitHub repository [1] represents the first public resource of its kind, aimed at helping researchers and academics understand the role of social media in shaping political discourse and opinion. By gathering data from a range of social media platforms, including Facebook, Instagram, Twitter, TikTok, and YouTube, this comprehensive dataset provides an in-depth look at how political narratives and conversations unfold across multiple online spaces.

The dataset was collected over a four-month period using public APIs and a keyword-based search strategy, ensuring that the captured content is relevant to the election and its key political actors. Millions of posts, along with associated metadata, were gathered from user accounts, public pages, and groups on Facebook, Instagram, and Twitter. In addition to this content, the collection also includes metadata of TikTok and YouTube videos that were shared across these platforms, offering insight into multimedia political communication.

To further enrich the dataset, we incorporated political advertisements sponsored on Meta platforms (Facebook and Instagram), as well as a list of social media handles corresponding to Italian political representatives. This provides researchers with additional context, allowing them to examine the role of paid content and official political figures in influencing online discussions. With this dataset, we aims to contribute to a broader understanding of how social media interacts with the democratic process, shedding light on the flow of information, the spread of political narratives, and the engagement of the public in political events.

3 An exploration of Bluesky decentralized social platform

We carried out a longitudinal analysis of user activity on Bluesky [4], a decentralized social media platform that gained significant traction after opening to the public on February 6th, 2024, following an initial invite-only phase. The analysis focuses on the two-month

[1] https://github.com/frapierri/ita-election-2022

period surrounding the platform's public release, capturing how the rapid expansion of the user base influenced the platform's dynamics. The study offers insights into the general characteristics of Bluesky, comparing its user behavior to more established social media platforms while also highlighting unique patterns emerging from its decentralized nature.

The findings reveal a broad distribution of activity on Bluesky, comparable to other mainstream platforms, but with key distinctions. Notably, users on Bluesky generated a higher proportion of original content compared to reshared content, and the overall level of toxicity was found to be very low. The platform's user base experienced a significant surge after the public opening, with a marked increase in posts, particularly in English and Japanese. This growth period also saw the emergence of suspicious activity, such as accounts following numerous users and sharing content from low-credibility news sources, identified using public (MediaBiasFactCheck) and proprietary (Newsguard) lists of news websites ratings. However, the platform's moderation system responded effectively, leading to the classification of some accounts as spam or their suspension.

By examining these trends, our findings provide an in-depth look at the growth of Bluesky, particularly how it handled the influx of users and potential challenges like misinformation and spam. The study contributes valuable insights into how decentralized social networks evolve, the nature of user-generated content, and the role of moderation in maintaining platform health as they scale.

4 Leveraging Large Language Models to analyze political news articles

We investigated the potential for media bias amplification in Large Language Models (LLMs) when used to summarize politically biased news articles [14]. As AI tools, especially LLMs, become more integrated into daily tasks, including within the news industry, their ability to shape information and potentially amplify existing biases raises concerns. Given the essential role that unbiased information plays in supporting democracy, it is crucial to understand how LLMs handle politically charged content and whether they contribute to the neutralization or reinforcement of media bias.

We studied the behavior of three recent LLMs by tasking them with summarizing and rewriting politically biased news articles related to the same event. To assess the polarity of the words used in these summaries, we employ arousal scores from the Valence-Arousal-Dominance (VAD) lexicon, offering insight into the emotional intensity of the language. In addition, the political bias of the generated summaries is classified into three categories—Left-biased, Right-biased, or Neutral—using a custom classifier developed for this study.

The results reveal that the LLMs generally lower the lexical arousal level of the original biased articles, suggesting a tendency to tone down emotionally charged language. However, this reduction in arousal does not consistently correlate with political bias, as the classifier often identified outputs as Right-biased, even in cases where lexical neutrality was achieved. While LLMs exhibit a significant capacity for neutralization from a linguistic standpoint, the study highlights the persistent challenge of political bias, with a notable portion of the generated outputs skewing toward the Right.

Overall, the work underscores the importance of critically evaluating LLMs' role in media and their potential to inadvertently propagate bias, despite their apparent linguistic neutrality. This work contributes to ongoing discussions on the ethical use of AI in news

production, particularly in contexts where balanced and unbiased reporting is vital for the democratic process.

5 Relevant Datasets in Literature

As required by the project specifications, one of the key points is the search for existing data regarding the manipulation of information from various sources such as newspapers and social media. During the search, several public datasets were identified as potentially relevant to the project. It is necessary to make an initial distinction between the data types identified, unimodal and multimodal. Unimodal dataset means data containing only one kind of data and, in this case, text-only data. In contrast, a multimodal dataset means data containing more than one data type: images and text.

5.1 Unimodal

- **Propaganda Techniques Corpus (PTC)**: PTC comes from the SEMEVAL challenge of detecting propaganda techniques in newspaper articles. The dataset consists of textual content only, in which sentences containing propaganda techniques were labeled. Using this data, it is possible to tackle different tasks such as span identification, whose goal is to identify within the text the propaganda phrase, or given the propaganda phrase, classify it based on a taxonomy of 18 persuasion techniques; finally, it is possible to unify the previous two tasks into a single task. More information on the dataset can be found at the official website [2].

- **Logical fallacy detection**: this dataset was proposed to identify locational fallacies within textual content. The dataset consists of two sub-datasets: the first is Logic, which contains general examples of logical fallacies. In contrast, LogicClimate contains logical fallacies in newspaper articles dealing with climate change. More information and data are available [3].

- **Mafalda** also focuses on detecting logical fallacies within a text. Mafalda is a benchmark for classifying fallacies according to a taxonomy, which unifies previous datasets on the same topic. The dataset includes a text part and labels indicating the beginning and end of each logical fallacy. More information and data are available [4].

5.2 Multimodal

- **MEME** comes from the SemEval-2024 Task 4 challenge to identify persuasion techniques within memes. The dataset is multimodal, consisting of image and text pairs labeled with 22 classes representing manipulation techniques, such as Name calling or Loaded Language. The dataset allows working on different subtasks, such as detecting manipulative content in binary form, then figuring out whether the meme contains manipulative or multiclass content by identifying the persuasion technique(s) used; finally, it is possible to do experiments in both multimodal and unimodal. More information on the dataset can be found at the official website [5].

[2] https://propaganda.math.unipd.it/ptc/

[3] https://github.com/causalNLP/logical-fallacy/tree/main/data

[4] https://github.com/ChadiHelwe/MAFALDA/

[5] https://propaganda.math.unipd.it/semeval2024task4/

- **Fakeddit** is a multimodal dataset containing image-text pairs of posts from the social network Reddit [9]. Each record has three types of labels: a binary, a ternary, and, finally, a senary. The first type classifies the record as fake or real; the second identifies a record as real or fake but with some real or fake text; finally, the 6-label labeling corresponds to true, satire, misleading content, impostor content, false connection, and manipulated content. More details on this dataset are provided in Section 6, while the data are available at the official website [6].

- **ReCOVery** is a multimodal dataset containing text-image pairs of newspaper articles covering the topic of COVID-19 [15]. Each instance has been labeled as reliable or unreliable. More details on this dataset are available in Section 6, whereas data are available at the following website [7].

- **Detecting and Grounding Multi-Modal Media Manipulation (DGM^4)** is a multimodal dataset for deepfake and text fake news detection. For each instance, several types of manipulation were created: a visual one that consists of a face swap, i.e., a person's face is artificially manipulated to change his or her identity and merge it with another person's, or a face attribute that consists of modifying a person's face to change the example emotion from neutral to happy. Another type of manipulation is textual manipulation, which consists of text swap, in which the text associated with the image is altered by changing the semantics but keeping the words used for the subject, or text attribute, in which words from the text are substituted to change the meaning. More info and data available can be found at the following reference [8].

6 Multimodal Model in the Context of Fake News Detection: Themis Model for Meme Classification in Multilingual Contexts

The ever-growing prevalence of social media platforms as communication tools necessitates sophisticated techniques to fight harmful content. In the context of our research, this study explores the proposal and application of the Themis model, a multimodal neural network architecture, to analyze and classify memes containing persuasive techniques. These methods aim to address the challenges of disinformation and polarization in the digital space, as memes often combine textual and visual elements to convey manipulative messages.

The Themis system was developed as part of the SemEval-2024 Task 4, specifically Subtask 2b, which focuses on binary classification to identify the presence of rhetorical or psychological persuasion techniques in memes [5]. Using a modular framework, Themis integrates contrastive learning-based image encoders, such as CLIP, with Large Language Models (LLMs), including TinyLlama and Phi, to process visual and textual content. A unique Token Merger module was employed to effectively align multimodal information, enabling the model to focus on salient features for classification.

[6] https://fakeddit.netlify.app/

[7] https://github.com/apurvamulay/ReCOVery/

[8] https://rshaojimmy.github.io/Projects/MultiModal-DeepFake

To optimize its performance, Themis leveraged Low-Rank Adaptation (LoRA), freezing the encoders' pre-trained weights while adapting specific parameters. Experiments with multilingual datasets revealed that the model achieved a Macro-F1 score of 79.8% in English but faced challenges in zero-shot settings for languages like Arabic and North Macedonian. Themis demonstrated significant potential for addressing harmful online content, offering a robust and adaptable solution for real-world applications.

7 Investigation of a Binary Multimodal Model for Classifying Fake News

Driven by the growing importance and impact that social networks have in our daily lives, we decided to undertake an in-depth analysis to explore the capabilities of a multimodal model to identify fake news. Indeed, social networks represent a major source of information for millions of people. Still, at the same time, they constitute a fertile ground for the spread of disinformation and fake news. In our study, we focused on data from social platforms such as Reddit and X (formerly Twitter), which are among the main vectors of user-generated content. The main objective was to assess how a multimodal approach, combining textual and visual information, can improve the identification of fake news compared to traditional methods.

In our work, we used Themis, a multimodal binary classification model originally developed for the SemEval23 task [5], which focused on the classification of memes. Themis effectively combines textual and visual information by exploiting an LLM (Large Language Model), such as TinyLlama, Phi, or other similar models, together with an image encoder, such as CLIP, to extract visual and textual features from the data. The features extracted from the two modules are then concatenated and processed by the LLM, which performs the final classification. To optimise the model's efficiency and reduce the number of parameters to be trained, Themis incorporates advanced techniques such as LoRA (Low-Rank Adaptation), which allows large models to be adapted more lightly. It also uses feature fusion methods, such as PatchMerger, to combine information from different modalities.

In this work, the Themis model was applied to two datasets concerning fake news. The first, Fakeddit [9], is a multimodal dataset of image-text pairs, extracted from the social network Reddit. It has been categorised in three ways: 2_way_label, which indicates whether the pair has been classified as **Fake** or **Real**; 3_way_label, where the instance may be **Real**, **False with some True text**, or **Completely False**; finally 6_way_label, which contains the following labels:

- **True**: content is completely accurate.
- **Satire**: content is true information with a satirical tone, which makes it false.
- **Misleading content**: content is purposely manipulated to fool the audience.
- **Imposter content**: content is bot-generated.
- **False connection**: the images don't match their correspondent text descriptions.
- **Manipulated content**: content has been purposely manipulated through manual photo editing or other forms of alteration.

As the Themis model is designed to perform binary classification, only the 2_way_label was used.

The second dataset is ReCOVery [15], a multimodal image-text dataset, containing articles concerning COVID 19. The articles were extracted from sites such as the New York Times, and after careful analysis were labelled as **Reliable** or **Unreliable**.

In order to improve the performance of the model, two types of multimodal data augmentation were performed for each dataset. The first type called MixGen [8], is based on merging two records to create a new one. In fact, given two image-text pairs (I_i, T_i) and (I_j, T_j), to create a new image I_k , a linear interpolation is performed between the respective pixels of the images I_i and I_j. While for the creation of the text T_k , T_i and T_j are concatenated. The second type of data augmentation, renamed **T**ext **S**ynonyms and **I**mage **T**ransformations (TSIT), consists of randomly replacing certain words in the text with synonyms and performing an image transformation.

Overall, Themis performed well in both datasets used. The experiments revealed that the model, supported by Lora, performs considerably better than its vanilla version (without the use of LoRA). In contrast, despite the introduction of data augmentation, the results did not improve on the Fakeddit dataset, whereas on the ReCOVery dataset, the model performed best with data augmentation, both TSIT and MixGen.

8 An Italian Corpus of Stereotypes

The presence, within news headlines or social media posts, of content that includes hate speech or stereotypes is a valuable indicator for identifying forms of propaganda and information manipulation techniques. These elements, in fact, play a crucial role in the implementation of strategies aimed at influencing the public, such as name calling, the use of loaded language or appeal to fear or prejudice. Such techniques rely on evoking intense emotions and creating a sense of belonging by manipulating the audience's critical judgment. For example, the use of racial, religious, or cultural stereotypes is often aimed at arousing feelings of fear, disgust, or distrust toward specific groups while fostering identification with the "virtuous" side promoted by propaganda. This polarizing process is one of the key tools for reinforcing the desired narrative. While not directly identifiable with the manipulation techniques at hand, these elements might serve as indicators of potential propaganda strategies, also offering support in analyzing and evaluating the degree of intensity with which these techniques are used. For these reasons, the availability of annotated corpora incorporating these aspects is central and represents a contribution to the activities of this Work Package.

Following a similar approach to the one proposed in [12] with the HaSpeeDe2 dataset, the QUEEREOTYPES corpus was developed and released. The corpus is based on data from social media, with a focus on stereotypes, specifically towards members of the LGBTQIA+ community [2]. The corpus features data from two different social media (Facebook and Twitter–now X) and comprises data in Italian, in view of its possible application in the Italian scenario. This results from a merger of two collections previously developed by independent groups that were eventually harmonized into a single resource. The Facebook section of the corpus includes data from public pages selected based on the main topics addressed and their content, both against and in support of LGBTQIA+ stances. It contains pages of right-wing politicians and pro-family movements on the one hand and groups or politicians that have been active in the defense of LGBTQIA+ rights on the other. The Twitter section contains tweets collected by activists from an

Italian LGBTQIA+ association using the Search Twitter API v1.118 from March 2018 to November 2018. Both corpus sections underwent first automatic pre-processing steps to reduce the collections to a tractable amount and remove noise and duplicate content, and then manual labeling by independent annotators. In line with the perspectivist data manifesto[9], two versions of the dataset were retained, a gold standard that conflates all labels provided by the annotators into a single final label (based on a majority voting), and in a non-aggregated form. This double facet of the dataset is a richness considering the growing research branch in which NLP models are not solely trained on a gold standard but rather on disagreeing annotations: i.e., *learning with disagreement* [13].

	Total number of instances	
Sub-corpus	**Non-aggregated**	**Gold standard**
Facebook	8,384	2,888
Twitter	5,310	3,427
Total	**13,694**	**6,215**

Table 1: Total number of texts with non-aggregated labels and of texts of the gold standard.

The dataset has also been validated in terms of its robustness as a training set for supervised predictive models for NLP tasks, such as precisely the automatic detection of stereotypes. We then used the above-mentioned HaSpeeDe2 dataset to obtain a baseline measure and considered two different experimental settings:

1. *HaSpeeDe2 Setting*, in which we trained and tested models on the training and test set provided within the HaSpeeDe2 shared task for the binary classification of stereotype.

2. *Expanded Setting*, in which we added QUEEREOTYPES to the training dataset of HaSpeeDe2 for training a model, and we tested it against the same test set provided by the evaluation campaign.

The models tested are the multilingual version of BERT [7] and AlBERTo [11], a BERT-based model pre-trained on a large collection of tweets in the Italian language [6]. Table 2 displays the averaged results of the two models on 5 runs and in both settings.

	HaSpeeDe2 Setting			*Expanded Setting*		
	P	**R**	**F1**	**P**	**R**	**F1**
mBERT	.740	.719	.698	.739	.740	.735
AlBERTo	.751	.729	.716	.746	.744	**.744**

Table 2: Results of textual classification experiments on the Stereotype dimension with mBERT and AlbERTo.

Focusing in particular on the *Expanded Setting*, the results show that models fine-tuned on a broader training set improve their performance, with Recall and F1-score being higher than in the *HaSpeeDe2 Setting*. This suggests that incorporating the additional QUEEREOTYPES data enhances the models' tendency to identify more instances as

[9] https://pdai.info/

positive examples of Stereotype. We conducted significance tests to verify whether the addition of the QUEEREOTYPES data has benefited the performance of both models and obtained a p-value of 0.0372 for mBERT and 0.0140 for the AlBERTo model. It is also worth pointing out that the HaSpeeDe2 dataset encodes the presence of stereotypes in Italian tweets towards immigrants, Muslims and Roma, while, on the other hand, QUEEREOTYPES encodes stereotypes towards LGBTQIA+ individuals. Overall, the obtained results highlight the importance of dataset diversity and extension in training models to enhance performance. The higher scores obtained in the *Expanded Setting*, where both models were fine-tuned on both racist and homophobic stereotypes, seem also to point out that stereotypes towards different targets share common traits. Therefore, the phenomenon of 'stereotyping' could be more generalizable, and the same models might be employed also for detection of stereotypes towards other vulnerable groups (women, elderly, people with disabilities, ethnic minorities, etc.).

References

Publications Acknowledging the DEMON Project

[1] Dominik Bär, Francesco Pierri, Gianmarco De Francisci Morales, and Stefan Feuerriegel. "Systematic discrepancies in the delivery of political ads on Facebook and Instagram". In: *PNAS nexus* 3.7 (2024), pgae247.

[2] Alessandra Teresa Cignarella, Manuela Sanguinetti, Simona Frenda, Andrea Marra, Cristina Bosco, and Valerio Basile. "QUEEREOTYPES: A Multi-Source Italian Corpus of Stereotypes towards LGBTQIA+ Community Members". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, May 2024, pp. 13429–13441. URL: https://aclanthology.org/2024.lrec-main.1176.

[3] Francesco Pierri. "Drivers of hate speech in political conversations on Twitter: the case of the 2022 Italian general election". In: *EPJ Data Science* 13.1 (2024), p. 63.

[4] 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.

[5] Luca Zedda, Alessandra Perniciano, Andrea Loddo, Cecilia Di Ruberto, Manuela Sanguinetti, and Maurizio Atzori. "Snarci at semeval-2024 task 4: Themis model for binary classification of memes". In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. 2024, pp. 853–858.

Other References

[6] Valerio Basile, Mirko Lai, and Manuela Sanguinetti. "Long-term Social Media Data Collection at the University of Turin". In: *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018.* 2018.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[8] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. "MixGen: A New Multi-Modal Data Augmentation". In: *arXiv* (2023). URL: https://arxiv.org/abs/2206.08358.

[9] Kai Nakamura, Sharon Levy, and William Yang Wang. "Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection". In: *arXiv preprint arXiv:1911.03854* (2019).

[10] Francesco Pierri, Geng Liu, and Stefano Ceri. "ITA-ELECTION-2022: A multi-platform dataset of social media conversations around the 2022 Italian general election". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 5386–5390.

[11] Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. "AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets". In: *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. Vol. 2481. CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[12] Manuela Sanguinetti, Gloria Comandini, Elisa di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, Irene Russo, and Pisa. "HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task". In: *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. CEUR-WS.org. 2020.

[13] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. "Learning from disagreement: A survey". In: *Journal of Artificial Intelligence Research* 72 (2021), pp. 1385–1470.

[14] Giulia Venturini. "Media Bias and AI: Evaluating LLMs' Neutralization of Politically Biased News Articles". In: *Master Thesis in Computer Engineering, Politecnico di Milano* (2024).

[15] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 3205–3212.