Beyond Fake News Detection: a Community-based Study of the Multicultural Nature of Information Disorder

Sara Gemelli[1,8], Giulia Di Cristina[2], Yiran Zhang[5], Md Azizul Hoque[5], Alberto De La Torre Solís[4], Mohamad Mojtaba Behboudi Eshkiki[2], Nikolai Efimov[2], Mariia Everstova[2], Caterina Maria Cappello[2], Maziar Kianimoghadam Joneghani[2], Payam Latifi[2], Yashar Mahboudi[2], Farzaneh Mohseni[2], Dario Placenti[6], Tommaso Caselli[3], Manuela Sanguinetti[7], Aurora Scarpellini[2], Chiara Zanchi[8], Usman Naseem[5], Marco Antonio Stranisci[2] and Simona Frenda[9]

[1]University of Bergamo, [2]University of Turin, [3]University of Groningen, [4]Universidad de Huelva, [5]Macquarie University, [6] Politecnico di Torino, [7]University of Cagliari, [8]University of Pavia, [9]Heriot-Watt University
Corresponding author: **s.frenda@hw.ac.uk**

Abstract

Recognizing disinformation is a challenging task for humans and AI systems. News can be false, misleading, or harmful, and its interpretation often depends on the cultural context of the audience. However, existing datasets rarely account for these contextual and cultural differences, as they are typically not designed from the perspective of news consumers. To address this gap, in this paper, we present the Information Disorder (InDor) corpus, a multilingual dataset of news articles in English, Farsi, Italian, and Russian, annotated for information disorder detection and explanation. The corpus was developed through a participatory process involving contributors from diverse cultural and professional backgrounds, who engaged in data collection, annotation, and evaluation of Large Language Model (LLM) performance on the task. Our findings highlight that false and manipulated news manifest differently across cultural settings, and that current LLMs fail to adequately capture this complexity. This underscores the need for culturally aware computational approaches in the study of information disorder. Additional material and the InDor dataset can be found in the GitHub repo: https://github.com/citizen-dataset/InDor.

WARNING: The InDor corpus may contain content that is offensive, including racist, sexist, or violent language.

Keywords: Information Disorder, Multi-Cultural Dataset, Multi-Lingual Dataset, LLM Alignment, Human Evaluation, Cross-Cultural Media Analysis

1. Introduction

Defining what fake news is remains a challenging issue due to its multifaceted nature (Nakov and Da San Martino, 2021), despite its frequent mention in the media and widespread presence on social media. Very different phenomena such as propaganda (Da San Martino et al., 2019) and conspiracy theories (Korenčić et al., 2024) are associated with this concept, and they often co-occur in determining its manifestations. For this reason, Wardle and Derakhshan (2017) proposed the concept of **Information Disorder** to highlight the multiple ways in which news can be problematic: *i.* Misinformation refers to the unintentional spreading of false content; *ii.* Disinformation to the spreading of false content with intentionality; *iii.* Malinformation to the framing of true news in a way that triggers reactions by readers and causes harm.

Previous work on fake news in Natural Language Processing (NLP) does not account for this complexity and is fragmented in different lines of research: works on disinformation focus on claim verification and fact-checking (Guo et al., 2022; Schlichtkrull et al., 2024); works on disinformation and malinformation study the presence of propaganda narratives[1] (Da San Martino et al., 2019), subjectivity (Alam et al., 2025), and media bias (Hamborg, 2020). Another limitation of existing research on fake news detection is the lack of works that adopt the perspective of news consumers. Despite the perspectivist turn in NLP (Cabitza et al., 2023), few works on disinformation detection consider the subjectivity of readers by design (Gabriel et al., 2022) and annotators are not active participants in the data collection process, increasing the risk of data selection bias (Søgaard et al., 2014) in these datasets.

In this paper, we present the **In**formation **D**is**or**der Corpus (InDor): a multilingual dataset of false and manipulated news. InDor is composed of $4,155$ news in English, Farsi, Italian, and Russian, collected and annotated through the adoption of a community-based approach. Participants with different socio-demographic backgrounds, professions, and origins have been involved throughout the entire

[1]https://propaganda.math.unipd.it/semeval2025task10/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6496–6508
11–16 May 2026.

corpus creation process. They contributed to a) the collection stage by reporting news sites that consistently spread suspicious news in their language, b) the annotation of the news with a free-text annotation scheme, and c) the evaluation of the outputs of one of the most recent LLMs (LLaMA 4 Maverick). By developing the InDor corpus, we investigated two main research questions:

RQ1: How do cultural and contextual factors influence the characteristics and perceived severity of problematic news across different countries? The adoption of a community-based approach in data selection and annotation allowed us to observe how the perception of what is manipulated and false is very sensitive to the different contexts where the news is spread. Participants agree the most when they must evaluate news in their native language, while English news amplifies the disagreement among them. The lexical analysis of news classified as problematic shows that the Information Disorder issue affects very different topics depending on the country they originated from.

RQ2: To what extent can LLMs assess if a piece of news is problematic, and with which groups of human annotators do their evaluations align most closely? We tested the ability of two LLMs from the LLaMA and Mixtral families to *i.* generate explanations and evaluate the severity of the problematic multilingual news and *ii.* explain why a given news in English, as annotated by non-English native speakers, is problematic and at which level. While the former set of experiments was measured with automatic metrics only, the latter was evaluated using both automatic scores and human judges. Both experimental studies show the language- and situational-based nature of information disorder, which can hardly be addressed by LLMs if not correctly guided.

2. A community-centered approach to corpus creation

Since our work aims at investigating the multicultural nature of information disorder, a crucial step of our corpus creation process has been the adoption of a methodology that emphasizes the collection of news relevant in specific local contexts and the definition of an annotation scheme that leverages the subjectivity of annotators. Inspired by the principles of participatory design in AI and NLP (Delgado et al., 2023; Caselli et al., 2021) and driven by the interest to highlight the nuances of local contexts in Information Disorder (Moorosi et al., 2023), we created the InDor corpus framing the whole task as a process where participants co-shape the artefact with the team that leads the research (Palacin et al., 2020). The methodology adopted to create InDor is based on the purpose of reducing the risks of marginalizing the communities that AI-based models seek to serve and empowering every community, encouraging people through calls to action to take part in the project. Below, we describe the most important steps of our approach.

Community engagement Participants' engagement happened in two phases: firstly, we engaged people interested in combating Information Disorder online; secondly, the small group of activists spread the designed calls to action in their networks, recruiting other interested individuals or making the action available to other online connections. All participants were constantly asked to provide feedback about each step and to share suggestions throughout the whole process.

Open data collection The first call to action aimed at collecting suspicious news or publicly declared fake news[2]. Participants in this call reported URLs of news conveying equivocal messages in a form. Together with the link to a news article, participants were asked to add a brief explanation about their choice to report it and to specify the language of the news (see 8). Reported news have been used as a starting point for the collection of texts, reducing the impact of data selection bias (Hovy and Prabhumoye, 2021) in our pipeline.

Descriptive annotation scheme Another call to action asked to recruit more annotators, namely native speakers of the available languages. Since we want to preserve the subjective perception of the annotators, we designed a descriptive annotation framework that combines question-guided reasoning with macro-level classification (Röttger et al., 2022). This structure encourages deeper reflection on both the severity of the news content and the types of misrepresentation affecting the individuals and events featured in the news. Participants were asked to select problematic spans and to write free-text explanations about their selection. This process facilitated the emergence of annotators' perceptions of issues in fake/suspicious news and their subjectivity.

Downstream evaluation Participants' role extended beyond corpus creation to include the evaluation of automatically generated explanations of problematic news. Testing if one of the most recent models (such as LLaMA 4 Maverick) is able to provide a good explanation of news problems opens

[2]Although we used the news collected until March to create the dataset used in this work, the collection of data through the form is still open and could be used to further increase the current dataset, encouraging diachronic investigation.

the discussion on the ability to help speed up news assessment and what the weaknesses and open research questions require being addressed in the future. To this end, we have also provided different pre- and post-annotation surveys and run interviews with a focus group. While this served as a feedback mechanism on the overall method adopted, highlighting both its strengths and weaknesses, it also opened up a broader reflection on the very notion of Information Disorder and the differing perceptions of it among participants.

Based on this theoretical background, our annotation focused specifically on the (mis)representation of events and participants in news discourse. This focus complemented the descriptive annotation approach we adopted, offering annotators a reference point to support their reasoning throughout the task.

3. The Information Disorder Corpus

In the following sections, we describe the process of creating the InDor dataset, which sees the active involvement of our participants.

3.1. From community to dataset

Stakeholders Different stakeholders have been involved in various moments in the creation of the dataset, coming from various environments and 5 countries: Italy, Spain, the Netherlands, Iran, and Russia. *i.* As described in Section 2, the first small group of individuals interested in combating Information Disorder online has helped substantively adjust the schema of annotation, translate and assess guidelines, and engage a broader public audience. *ii.* Subsequently, the extended participatory group was involved in the annotation process and the evaluation of automatically generated explanations. Engaging individuals outside the initial team allowed the construction of a corpus that encompasses a wide range of topics, sources, and viewpoints. *iii.* Especially during the annotation design, we organized online meetings to ask academic experts from various universities, newsrooms, and non-profit organizations for feedback on the annotation scheme and engagement of the broader community.

Dataset creation The first call to action disseminated by participants through a card in their social media[3] aimed at collecting suspicious or publicly declared fake news from different online newspapers active in different countries. We asked participants to fill out a form reporting the URL, the news language, and the motivation. The collection phase was intentionally designed with minimal constraints to encourage the inclusion of a wide range of perspectives and maximize dataset diversity. This collection strategy, employed between February and March 2025, helped us gather 799 suspicious news items whose website domains have been used as seeds for collecting a total of 11,199 news across four languages. To balance the load of data to be annotated, we distributed a minimum of 20 and a maximum of 250 news to the annotators depending on the time they were involved in the annotation and the load they could support. Table 1 shows the distribution of collected and annotated news per language.

[3]To broaden participation and facilitate international engagement, the card was originally written in English and translated into nine languages: Italian, Arabic, Spanish, Japanese, French, Chinese, Catalan, Russian, and Dutch. Alongside this post, we also shared a second card on social media, which provided simple guidelines to help participants identify problematic content.

4. Corpus-based analysis

4.1. Annotation

Each news has been processed to facilitate the annotation process. We automatically extracted headlines and the three most significant sentences in the news articles using the Sumy library, based on extractive summary methods. This resulted in shorter text pieces, which made the annotation load lighter.

The annotation task is characterized by three components: text classification, span annotation, and text writing. For each news item, annotators were asked:

1. To identify the span of text that they consider problematic, looking in particular at the misrepresentation of events and individuals/communities in the news. To help guide the annotation, annotators were led to assign an *Eventive* or *Attributive* label. The *Eventive* label was to be used when the problematic span referred to how the events were presented or framed. The *Attributive* label was applied when the span concerned the participants or entities involved, focusing on how they were described or referenced, or more generally, on information that was provided about them in the text[4].

2. To write an explanation of their choice with a semi-structured template. Rationales could be written in free text, but annotators were encouraged to write their rationales using an *if... / then...* structure (see Example 4.1). This format served two main purposes: first, it was conceived as a means to stimulate reflective reasoning during the annotation process; second,

[4] *Eventive* and *Attributive* labels have not been used in the experiments. Their use is aimed at guiding the annotation.

language	website domains	annotated news	avg	#annotators	speakers
ru	63	500	2.00	2	native
it	242	1,656	1.88	64	native
en	358	999	2.01	9	non-native
fa	47	1,000	1.95	6	native

Table 1: Distribution of collected news and annotators who annotated data across the four languages in InDor. We report the amount of website used to extract news articles, the number of annotated news (composed of a headline and three sentences), the average number of annotators per news, the total number of annotators per language, and their language knowledge.

First Language	Gender		Ethnicity				Education				Employment					Age	
	M	F	White	Black	Asian	other	HighSchool	Bachelor	Master	PhD	Fu	Pa	Un	NP	other	<30	>30
Italian	5	26	29	0	0	0	0	29	1	1	3	6	13	1	8	28	3
Persian	4	1	4	0	0	1	1	4	0	0	1	1	1	1	1	2	3
French	0	2	1	0	0	0	0	2	0	0	0	0	1	0	1	2	0
Russian	1	1	0	0	2	0	0	2	0	0	0	0	2	0	0	2	0
Arabic	0	1	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0
English	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0

Table 2: Demographic distribution of annotators by their first language.

it could facilitate the automatic processing of annotators' comments.

3. To evaluate whether the news conveyed a problematic message and to what extent, by responding to the question: *How problematic is this news?* The available options were: *Highly, Moderately, Slightly, None*, where *None* indicated no problematic content, and *Highly* denoted the presence of strongly problematic elements. The *N/A* option was provided for cases in which annotators felt unable to evaluate the item (e.g., the text was not a news content but advertisement, caption, and so on)[5].

In addition, we provided a set of six questions designed to stimulate reflection on what aspects of a news report might be considered problematic, in line with the goals of the scheme[6].

The output of the annotation scheme is a set of triplets that include the context - which, in turn, includes the news title and the sentences conveying a problematic message - the labeled problematic span, and the explanation. An instance of a fully annotated data point is reported below:

CONTEXT: Melania Trump Makes Apparent Retaliation For 'Stormy Daniels' Affair; Donald In Shambles
EVENTIVE: Apparent Retaliation
EXPLANATION: ***if*** you use 'retaliation' to describe Melania Trump's response, ***then*** you are using sensationalist language that aims to create clickbait while also misleading the reader by diverting attention to a different topic than what the article is actually about

[5]These cases have been excluded from the analysis and the experiments.

[6]The complete guidelines for the annotation can be found in the GitHub repository.

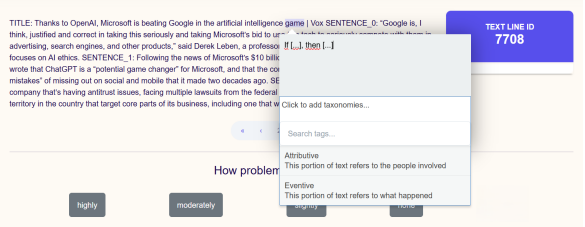

Figure 1: Screenshot of the annotation platform.

The annotation has been performed through a dedicated platform (see Fig. 1), inviting voluntary annotators via email. Annotators have been asked to fill out a brief questionnaire about their socio-demographic information, precisely gender, age, first language, nationality, educational level, and employment status. Table 2 shows the distribution of the various traits.

Throughout the whole annotation phase, we held meetings to introduce a first draft of the annotation schema and guidelines and progressively refine the process through collaborative feedback. The initial sessions were dedicated to outlining the objectives of the project and explaining the annotation procedure. In the following meetings, we invited feedback on both the schema and the guidelines, with the aim of improving their clarity and usability. After an initial annotation step, participants were again encouraged to share reflections on the annotation experience, including challenges encountered and suggestions for refining both the annotation framework and the dataset itself.

The annotation process in particular was articulated in two phases: firstly, non-native English speakers (see Table 1) have been asked to annotate an English set of news; secondly, they have been asked

to annotate news in their own languages. The English set of news allowed us to investigate deeply the different perceptions of how news are manipulated across several annotators coming from different cultures and countries (Italy, Iran, Russia, Cameroon, France, and the US).

InDor[7] is a corpus of $4,155$ news that received on average 2 annotations (from a minimum of 1 and a maximum of 4, depending on the available number of annotators) regarding the severity of news, the presence of misleading portions of texts and their explication. The dataset covers the four languages used in collected data (Farsi, English, Italian, and Russian)[8] and a total of 81 people were involved in the evaluation, none of whom were native English speakers. This enabled the comparison between a setting in which participants annotated news in their native language and a setting where they did not. The corpus is imbalanced by design: the adoption of a citizen-based approach led to a high variation of data splits and annotation workload depending on annotators' availability. In Table 1 statistics about the corpus and the annotators are reported.

In this section we leverage InDor to investigate the impact of cultural and contextual factors in perceiving problematic news across different countries (RQ1).

Label Distribution. Our first analysis focuses on differences emerging among annotated news in different languages. In Table 3 we report the classification of news severity on a scale from *None* to *Highly* across the four languages. As can be observed, the percentage of news perceived as non-problematic by annotators is always the highest, even if there is a significant variation between Farsi, which includes 43.08% non-problematic news, and other languages that are all above 56%. On the opposite side of the scale, there emerges a difference between highly problematic news in English and Farsi, which are both below 6%, and the percentage in Italian and Russian above the 12%. During the interactions, annotators of Russian news reported that the use of high severity scores is due to the repetitive presence in news of emotional and polarized language related to the war between Russia and Ukraine. This includes ideological expressions and repeated rhetorical devices[9]. Deepening the analysis to span labels, it emerges that the annotations of Italian, Farsi, and Russian spans are more focused on the misrepresentation of events (> 60%), e.g., *'La popolazione deve essere aumentata in tutti i modi possibili'*[10] in a news on declining birth rates. English annotated spans highlight the misrepresentation of people and communities (41.37%), see e.g. the use of 'Hamas goons' in news on the ongoing conflict in Palestine. These differences appear to be the result of the many ways in which Information Disorder might be shaped in different contexts.

Score	en	it	ru	fa
Highly	5.95	12.07	16.57	4.34
Moderately	11.91	13.13	12.94	10.96
Slightly	25.41	18.13	12.37	41.62
None	56.72	56.67	58.12	43.08

Table 3: Distribution of severity score in percentage.

	en	it	ru	fa
severity	0.03	0.12	0.06	0.10
spans	0.16	0.25	0.33	0.29

Table 4: IAA in terms of Krippendorff's alpha and token-match accuracy, respectively, for the severity score and the portion of text selected by annotators.

Annotators' agreement. A second focus of our analysis is on disagreement. We observed if and to which extent annotators disagree on news in their primary language and in English. Given the unbalanced number of annotators per language (Table 1) and the sparsity of annotation (not all the news have been annotated by all the annotators), we used Krippendorff's alpha for calculating the inter-annotator agreement (IAA) on news severity and the token-match accuracy (Wang and Yu, 2023) for the agreement on problematic spans. Results in Table 4 show that the task is highly subjective. The highest agreement on severity classification is $\alpha = 0.12$, the highest score for token-match accuracy is 0.33. The high subjectivity emerging from annotators' disagreement is accentuated in English news, which are annotated with the lowest agreement both on severity ($\alpha = 0.03$) and span selection (0.16). The subjectivity of annotators in the perception of problematic news is amplified by language, which is both a linguistic and a cultural barrier (Kramsch, 2014). To investigate deeper the agreement among annotators per language, we extracted the most frequent words selected in the problematic portion of text and used by annotators in their explanation (Tables 5 and 6).

Lexical Analysis As a final part of our analysis, we observed the lexical differences emerging from span annotations. For each language, we identified

[7] https://github.com/citizen-dataset/InDor

[8] No enough data in other languages have been collected during the data collection phase.

[9] For instance, the fact that Zelensky is often compared to Hitler delegitimizes the Ukrainian government and evokes strong emotional responses from the audience.

[10] "Population must be increased in every possible way".

en	it	ru	fa
make	solo	россия	کشور
year	volere	формат	ایران
say	dire	остаться	اسلامی
people	leader	антироссии	آمریکا
new	parte	год	استان
one	altro	который	سازمان
trump	europeo	страна	رئیس
take	anno	техника	اقتصادی
muslim	nazionale	уничтожить	تولید
get	sempre	западный	شهر
world	proprio	находиться	کشورهای
see	europa	враг	تهران
state	magistrato	металлолом	دولت
work	poco	военный	انقلاب
right	russo	котором	توجه
time	alcuno	новый	ملی
well	molto	противостояние	حوزه
go	uniti	готовить	حضور
medium	vita	уничтожить	افزایش
need	ora	антиРоссия	دلار

Table 5: Most frequent words used by annotators in the identified spans.

en	it	ru	fa
risk	rischiare	рисковать	محدود
reader	utilizzare	внешний	سیاسی
term	rischio	описываете	بررسی
context	creare	антиукраинский	بیان
sentence	riportare	россия	تحلیل
perception	espressione	нормализовать	ارزیابی
language	titolo	государство	درک
frame	negativo	утверждать	اطلاعات
state	uso	легитимность	متن
shape	politico	создавать	نقش
neutrally	informazione	суверенитет	ایدئولوژیک
provide	parola	которая	تأکید
potentially	contesto	военный	مطرح
toward	linguaggio	российский	منابع
author	volere	игнорировать	آمار

Table 6: Most frequent words used by annotators in their explanations.

the top 20 frequent tokens from spans labeled as problematic by annotators. Most frequent tokens in each language appear to show a specific way in which problematic news is conveyed. English spans mostly refer to people (e.g., 'Trump') and religious groups (e.g., 'Muslim'), confirming our previous findings on the major focus of these news on people's misrepresentation. Italian news are characterised by a focus on geopolitical relations in Europe (e.g., 'Europa', 'Europeo') and on immigration. The word 'magistrato'–*magistrate* is often retrieved in news that represent the migration topic as the product of a conflict between politics and judiciary, where the latter is framed as the defender of migrants against the government. Russian annotated spans are more focused on the war between Russia and Ukraine, with words like военный–*military* and уничтожить–*destroy* among the top 20 tokens. Adverbs or adjectives like антиРоссия–*anti-Russia* carry strong ideological and geopolitical connotations. Specifically, the term *anti-Russia* used to frame Ukraine or other post-Soviet states in opposition to Russia, suggesting deliberate antagonism or Western alignment. Situational words are more common in Persian spans, where annotators tend to select portions of texts containing terms such as 'اقتصادی'–*economic*, 'دلار'–*dollar* and 'آمریکا'–*US* connected to economic issues that cause daily stress for the population and the ongoing political tensions in the international relations of Iran. Similarly, in Russian news, words like металлолом– *scrap metal* appear, which, although less abstract, are metaphorically employed to mock or delegitimise Western military aid and Ukrainian capabilities, contributing to the discursive construction of superiority and derision.

The lexical analysis of words occurring in users' explanations shows the tendency to inform the journalist about the 'risk' of using some expressions or the need to take care about 'information' (*informazione*), different perceptions, context and language used (*linguaggio*). The same tendency is evident in Persian explanations, where annotators underline the need of certain 'statistics' (آمار) and 'sources' (منابع) not mentioned in the news. This selective use of information demonstrates 'ideological' (ایدئولوژیک) biases.

5. LLM-based Experiments

In this section we present two sets of experiments aimed at testing the performance of LLMs in the recognition of problematic news in a multilingual setting (RQ2). In Section 5.1 we evaluate two English-centric LLMs, developed in Europe and in the USA, respectively, on the tasks of span detection and explanation generation of the identified span in news written in four languages and published in as many national contexts. In Section 5.2 we study the alignment of a SOTA model with non-English speakers with different backgrounds who annotated the English subset of InDor. Both experiments are carried out on news where at least one span or severity label has been identified by annotators (see Table 7)[11].

[11]The cleaning process of the dataset per language included: filtering of news identified as n/a, filtering of news where the explanation is missing, and filtering of news where the language used in the explanation does not correspond to the news language.

language	#news 1	#severity	#news 2	#spans-explanations
ru	481	760	193	632
it	1,495	2,275	334	888
en	877	1,371	418	1,201
fa	932	1,507	660	1,862

Table 7: Amount of news used per task after the filtering process is applied: 1. severity classification and 2. span-explanation generation. We report the total number of provided annotations per severity and span-explanation pairs.

5.1. LLM Performance on Multilingual News

Our first analysis aims to test the ability of LLMs in detecting problematic spans of text in news and generating effective explanations about the problematic nature of news. We provide a comparative analysis of two English-centric LLMs developed in two different national contexts to assess their effectiveness in handling multilinguality: LLaMA 4 Maverick and Mixtral-8x22B-Instruct. We tested models on all languages included in our corpus, adopting two different settings: a zero-shot and a few-shot setting. All experiments have been performed with OpenRouter.ai selecting the standard configuration for each model (see 8).

We chose for the evaluation of both tasks the BERTScore measure (Zhang et al., 2019). Since the number of predicted spans or explanations is variable and may always be different from gold-standard ones, we employed the following strategy: i) we compute all pairwise similarities between predicted spans/rationales and human-identified spans/rationales, ii) select the highest-scoring one-to-one matches, iii) average within each sample, and then iv) average across the dataset. Results are shown in Table 8.

As the table shows, we can notice the relatively low performance of both models in the detection of problematic spans in English news in the few-shot settings. We observe the opposite dynamics for Farsi and Russian, which show a clear increase of performances from zero-shot to few-shot. Since no annotators are native English speakers, results may be interpreted as an effect of the standard English bias in models (Lucy et al., 2024), whose outputs are not representative of non-native English speakers.

The evaluation of generated explanations seems to confirm this trend. Both Mixtral and LLaMA have a drop in performance on English explanations, shifting from a zero-shot to a few-shot setting. The models' results diverge on other languages: LLaMA has a drop on Italian explanations, and Mixtral on Farsi news. These differences might be the results of annotators' subjectivity and of the origin-specific dimensions related to the spreading of false and manipulated news.

5.2. LLM Alignment with non-Native Speakers

This experiment analyzes the performance of LLaMA 4 Maverick against a pool of 11 annotators who are not native English speakers. The model's evaluation followed a twofold approach: *i.* We assessed its performance in predicting news severity by calculating the F-score across clusters of annotations grouped by annotators' gender and primary language. *ii.* We asked annotators to rate the quality of the model-generated explanations ("How much do you agree with the explanation provided by the model") on a 1–4 scale, then analyzed the average scores within the same demographic clusters. Detailed results, broken down by annotators, gender, and primary language, are provided in Appendix 8.

Table 9 indicates that in both tasks, gender and native language seem to have a significant impact. The average F-score of LLaMA in news severity classification is higher by 0.07 points on annotations provided by men; the difference between Farsi-speaking and Italian-speaking annotators is 0.11 F-score points.

The downstream evaluation shows the same pattern. Men evaluate news with an average score of 3.28 out of 4, women with a score of 2.68; Italian speakers with an average score of 2.44 against Farsi speakers with an average score of 3.53. As for the analysis in Section 5.1, culture-specific factors and socio-demographic traits seem to have a relevant impact on Information Disorder which is characterized by a common global reach but a high variety of local instantiations. The adoption of a community-based approach appears to be promising for facing such a complex phenomenon and identifying cultural biases in LLMs.

6. Discussion

A key feature of this study was the sustained engagement of participants beyond the creation of the InDor dataset. Periodic meetings served not only as a feedback mechanism but also as a place for annotators to reflect on their evolving perspectives on the concept of "problematic" in the context of news, share interpretations, and collectively artic-

Language	Setting	LLaMA		Mixtral	
		Span	Explanation	Span	Explanation
en	Zero-Shot	40.82	45.83	48.83	45.72
	Few-Shot	33.44	41.68	41.06	45.56
it	Zero-Shot	43.78	45.62	38.39	41.36
	Few-Shot	42.77	42.83	45.03	41.03
ru	Zero-Shot	52.79	44.90	48.59	42.06
	Few-Shot	56.28	46.83	54.96	46.09
fa	Zero-Shot	38.35	50.72	49.69	45.60
	Few-Shot	60.93	53.37	63.28	44.30

Table 8: BERTscore evaluation for spans and rationales across models, settings, and languages.

Task	men-women	Italian-Farsi
Severity Classification	0.44-0.37	0.36-0.47
Downstream Evaluation	3.28-2.68	2.44-3.53

Table 9: LLaMA performance on severity news classification and downstream evaluation of generated explanation. Results are broken down by gender and native language of annotators.

ulate the challenges encountered during the task. Most of them concerned the reasoning side of the task, while others related to the limited diversity of the news samples, such as the Russian-language articles, which participants noted as being strongly influenced by the contemporary sociopolitical context.

Along with meetings during the annotation phase, we employed two structured surveys, specifically a pre-annotation, a post-annotation test, and a focus group[12]. The pre-annotation survey, provided to annotators before the annotation process, contains questions to capture the knowledge, perception, and confidence of annotators about the detection of information disorder. The post-annotation survey, provided after the annotation, tests if annotators acquired a new perception or improved their sensibility toward the recognition of false and misleading news. Differently from pre- and post-surveys, the focus group was a free interaction guided mainly by the following questions to which citizens could report further insights:

1. What did you find interesting during the annotation?
2. Which differences did you notice between annotating in English and in your own language?
3. On which other topics would you organize another annotation task?
4. Which are the functionalities and usages that a technology trained on these data must have?

The outcomes showed that this experience of engagement in the creation of InDor has been perceived as enriching and acknowledged its potential value for future applications. Further details are presented below.

Evolving Perception and Subjectivity Responses to the pre-annotation survey indicated that participants sought to improve their critical reading and bias detection skills. Over time, both the post-annotation survey and the focus group revealed a noticeable evolution in participants' perception of media content. They reported a heightened awareness of what they described as "more subtle manipulative strategies"- in contrast to the more overtly offensive, violent, or stereotypical ones they were used to noticing. In particular, participants reported becoming increasingly attentive to indicators such as biased reporting, lack of verified sources, and the use of inflammatory language. The post-annotation survey further suggested that the task also served as a **learning-by-doing** experience, where repeated exposure to texts deepened their critical awareness and challenged surface-level assumptions.

Cultural and Linguistic Sensitivity Focus group discussions underscored the role of cultural and linguistic context in shaping perceptions of the severity of news and bias. Participants noted significant differences in how news content was framed across languages and media systems. For example, Persian news was often seen as embedding religious ideology, while Russian texts displayed leader-centric framing and censorship in contrast to the relative English-language pluralism. Participants emphasized that identical discursive strategies could have varying connotations depending on the surrounding cultural and political context, highlighting the importance of linguistic and cultural sensitivity in media interpretation. And this relevance is also evident in the language-based behaviors of the two LLMs used in the first experimental study (Section 5.1).

Civic Orientation and Impact Participants also recognized the broader civic value of their contribution. Several noted that the InDor dataset could help future annotators better understand problematic

[12]This material is released in the GitHub repo.

news content. During the focus group, many viewed the annotated dataset as a potential foundation for applications such as explainable AI tools, educational resources, and media literacy interventions. This pedagogical dimension highlights its potential dual function: producing valuable data while cultivating informed individuals. This **civic orientation** reflects a growing consensus in the NLP community that participatory methods should not only serve model performance but also empower communities to interrogate and shape the technologies that affect them (Bender and Grissom, 2024).

7. Related work

Online disinformation is a pressing challenge for our societies. Its role in influencing elections (Allcott and Gentzkow, 2017) and behaviors (van Der Linden et al., 2020) gathered the attention of different societal actors (e.g., fact-checkers, scholars, and media) aimed at mitigating its negative impact.

Information Disorder Especially in NLP, the growing interest in Information Disorder led to a growing number of contributions in this field. Several datasets have been released on misinformation Alhindi et al. (2018); Wang et al. (2025), disinformation Gabriel et al. (2022); Bondielli et al. (2024), and polarizing narratives Minnema et al. (2021); Rosso et al. (2024). Various technologies[13] have been launched as well, like the tools proposed by Lupi et al. (2023) and Wührl et al. (2023) for the automatic recognition of fake news. However, this field of research suffers from a lack of a common theoretical framework on this topic that causes a fragmentation of approaches: initiatives like FEVER (Schlichtkrull et al., 2024) and CheckThat! Lab (Alam et al., 2025) focus on fact-checking and news verification; others, like Rangel et al. (2020) proposed a task on the detection of fake spreaders, and Piskorski et al. (2023) on news framing. Differently from previous works, InDor is the first dataset built through a campaign of collection and annotation aimed to recruit volunteers interested in addressing "Information Disorder" following the framework introduced by Wardle and Derakhshan (2017). Moreover, it is also the first dataset in such a domain containing the explanations in natural language of the portions of text considered false or misleading.

Framing in news reports The same news story can be presented from different perspectives, depending on the linguistic choices made by the writer. Studies from various disciplines and approaches, including critical discourse analysis (Van Dijk, 2013; Hart, 2018), psychology (Bohner, 2001), pragmatics (Meluzzi et al., 2021), computational linguistics (Minnema et al., 2021; Remijnse et al., 2024), and sociology (Te Brömmelstroet, 2020), have demonstrated that specific semantic and syntactic choices can influence readers' perceptions of news reports.

Prior work has shown that both event descriptions and participant portrayals contribute to shaping interpretation and perception (Fowler et al., 2018; Samaie and Malmir, 2017; Li and Zhang, 2022; Stabile et al., 2019). Representations of agency, blame, and stereotypes often emerge not only from how events are reported but also from how involved actors are characterized. Inspired by these works, we designed a schema of annotation able to guide the annotators to identify misleading news in InDor, reporting also their explications that revealed a deeper level of subjectivity, affected by historical events, personal beliefs, and cultural elements.

8. Conclusions

In this paper, we introduced the Information Disorder (InDor) Corpus, a multilingual dataset of problematic news articles constructed through a community-centered annotation process. By engaging participants from diverse cultural and professional backgrounds, we sought to capture the plurality of perspectives that shape how problematic content is perceived and interpreted. This participatory approach grounded the annotation process in real-world judgements, underscoring the value of incorporating cross-cultural viewpoints in the study of information disorder. Our experimental results show that current LLMs struggle to reliably identify and explain problematic content, often aligning more closely with annotators who share their native language or socio-demographic characteristics. This suggests a lack of cultural robustness in existing models, with implications for fairness and generalizability in information disorder detection. Future work will focus on expanding the diversity of annotators across linguistic and cultural dimensions, with the goal of experimenting with multilingual speakers and improving both the dataset's coverage and our understanding of how computational models can better accommodate global perspectives.

Limitations

A persistent challenge in our methodology was dealing with **power imbalances** typical of traditional research hierarchies. Despite our explicit commitment to collaborative principles, we noticed that perceived hierarchical relationships inevitably affected participation dynamics, especially in the early stages of the project.

[13] https://www.veraai.eu/home, https://www.askvera.org/en, https://www.bellingcat.com/

To overcome these barriers, we have tried to constantly ask for feedback and suggestions, explicitly acknowledging and validating all valuable contributions during meetings. By demonstrating that participant input was being seriously considered through visible implementation in the design and overall project development, we attempted to create a more equitable collaborative environment. For future projects, we want to introduce more ice-breaking activities, which have proved to be useful in focus groups and workshops.

An ulterior limitation was the **limited number of languages and the unbalanced number of annotators** and of their socio-demographic traits due to the community-based nature of the creation of InDor. This could influence the creation of models trained on InDor. To this purpose, we have maintained the collection and annotation processes of the dataset ongoing in order to make an updated dataset on information disorder extended also to other languages.

Moreover, unfortunately, not all annotators participated in the downstream evaluation, leading to less data for the **human-based evaluation**. To involve a broader segment of the community beyond the activists' and researchers' networks, future efforts will explore the use of incentive-based strategies or reward mechanisms to enhance engagement and sustained participation. We also plan to perform in future work a comparative analysis between human- and model-produced explanations to test the possible presence of prejudices towards automatically generated texts.

Finally, although we recognize that the use of the citizen-participatory approach and the lack of specific guidelines in data collection could have introduced news articles that are **intentionally fabricated but non-malicious** (i.e., satire and parody), the data annotation stage did not reveal the presence of similar cases. However, to guarantee the quality of collected data, in further approaches, we will design adequate guidelines and reliability checks in the data collection stage.

Ethical Statement

This research relies on the voluntary work of those who participated in the initiative. All the involved annotators freely accepted taking part in the laboratory, for which no compensation was provided. We adopted all the measures to protect data privacy and safeguard personal information. The work has been approved by the Ethical Committee of the institution of one of the authors. The dataset has been collected according to the existing laws on copyright, and the corpus is licensed for only non-commercial uses.

Acknowledgment

All authors thank the support of aequa-tech, which funded the data collection and annotation through the community-centered participatory approach and provided the Citizen-Dataset annotation platform (https://citizen-dataset.aequa-tech.com/) and the Google Workspace, in which the annotation, annotation adaptation scheme, and focus groups were performed.

References

Firoj Alam, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and Venktesh V. 2025. The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In *Advances in Information Retrieval*, pages 467–478, Cham. Springer Nature Switzerland.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Emily M. Bender and Alvin Grissom. 2024. 199power shift: Toward inclusive natural language processing. In *Inclusion in Linguistics*. Oxford University Press.

Gerd Bohner. 2001. Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim. *British journal of social psychology*, 40(4):515–529.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, et al. 2019. Fine-grained analysis of propaganda in news article. 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)*, pages 5636–5646. Association for Computational Linguistics.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23.

Roger Fowler, Bob Hodge, Gunther Kress, and Tony Trew. 2018. *Language and control*. Routledge.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Felix Hamborg. 2020. Media bias, the social sciences, and NLP: Automating frame analyses to identify bias by word choice and labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 79–87, Online. Association for Computational Linguistics.

Christopher Hart. 2018. Event-frames affect blame assignment and perception of aggression in discourse on political protests: An experimental case study in critical discourse analysis. *Applied Linguistics*, 39(3):400–421.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Damir Korenčić, Berta Chulvi, Xavier Bonet Casals, Alejandro Toselli, Mariona Taulé, and Paolo Rosso. 2024. What distinguishes conspiracy from critical narratives? a computational analysis of oppositional discourse. *Expert Systems*, 41(11):e13671.

Claire Kramsch. 2014. Language and culture. *AILA review*, 27(1):30–55.

Ke Li and Qiang Zhang. 2022. A corpus-based study of representation of islam and muslims in american media: Critical discourse analysis approach. *International Communication Gazette*, 84(2):157–180.

Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. 2024. Aboutme: Using self-descriptions in webpages to document the effects of english pretraining data filters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7393–7420.

Arthur Thomas Edward Capozzi Lupi, Alessandra Teresa Cignarella, Simona Frenda, Mirko Lai, Marco Antonio Stranisci, Alessandra Urbinati, et al. 2023. Debunker assistant: a support for detecting online misinformation. In *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, volume 3596, pages 1–5. Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini, Nicole Novielli.

Chiara Meluzzi, Erica Pinelli, Elena Valvason, and Chiara Zanchi. 2021. Responsibility attribution in gender-based domestic violence: A study bridging corpus-assisted discourse analysis and readers' perception. *Journal of pragmatics*, 185:73–92.

Nyalleng Moorosi, Raesetje Sefala, and Sasha Luccioni. 2023. Ai for whom? shedding critical light on ai for social good. In *NeurIPS 2023 Computational Sustainability: Promises and Pitfalls from Theory to Deployment*.

Preslav Nakov and Giovanni Da San Martino. 2021. Fake news, disinformation, propaganda, and media bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4862–4865.

Victoria Palacin, Matti Nelimarkka, Pedro Reynolds-Cuéllar, and Christoph Becker. 2020. The design of pseudo-participation. In *Proceedings of the 16th Participatory Design Conference 2020-Participation (s) Otherwise-Volume 2*, pages 40–44.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. 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)*, pages 2343–2361.

Francisco Rangel, Anastasia Giachanou, Bilal Hisham Hasan Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CEUR workshop proceedings*, volume 2696, pages 1–18. Sun SITE Central Europe.

Levi Remijnse, Pia Sommerauer, Antske Fokkens, and Piek TJM Vossen. 2024. Tracking perspectives on event participants: a structural analysis of the framing of real-world events in co-referential corpora. In *Proceedings of the First Workshop on Reference, Framing, and Perspective@ LREC-COLING 2024*, pages 1–12.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 175–190.

Mahmoud Samaie and Bahareh Malmir. 2017. Us news media portrayal of islam and muslims: a corpus-assisted critical discourse analysis. *Educational Philosophy and Theory*, 49(14):1351–1366.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, et al. 2024. Proceedings of the seventh fact extraction and verification workshop (fever). In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*.

Anders Søgaard, Barbara Plank, and Dirk Hovy. 2014. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13.

Bonnie Stabile, Aubrey Grant, Hemant Purohit, and Kelsey Harris. 2019. Sex, lies, and stereotypes: Gendered implications of fake news for women in politics. *Public Integrity*, 21(5):491–502.

Marco Te Brömmelstroet. 2020. Framing systemic traffic violence: media coverage of dutch traffic crashes. *Transportation research interdisciplinary perspectives*, 5:100109.

Sander van Der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about covid-19. *Frontiers in psychology*, 11:566790.

Teun A Van Dijk. 2013. *News analysis: Case studies of international and national news in the press*. Routledge.

Hongwei Wang and Dong Yu. 2023. Going beyond sentence embeddings: A token-level matching algorithm for calculating semantic textual similarity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 563–570.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Amelie Wührl, Lara Grimminger, and Roman Klinger. 2023. An entity-based claim extraction pipeline for real-world biomedical fact-checking. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 29–37.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Language Resource References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90.

Alessandro Bondielli, Pietro Dell'Oglio, Alessandro Lenci, Francesco Marcelloni, and Lucia Passaro. 2024. Dataset for multimodal fake news detection and verification tasks. *Data in Brief*, 54:110440.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Viviana Patti, Tommaso Caselli, Malvina Nissim, et al. 2021. Frame semantics for social nlp in italian: Analyzing responsibility framing in femicide news reports. In *CEUR Workshop Proceedings*, volume 3033, pages 1–8. CEUR-WS.

Paolo Rosso, Berta Chulvi, Damir Korenčić, Mariona Taulé, Xavier Bonet Casals, David Camacho, Angel Panizo, David Arroyo, Juan Gómez, and Francisco Rangel. 2024. Xai-disinfodemics: explainable ai for disinformation and conspiracy detection during infodemics.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2025. When automated fact-checking meets argumentation: Unveiling fake news through argumentative evidence. *Argument & Computation*, 16(1):19462174251330980.

Appendix A: Data collection

Table 10 reports one extract per language of the type of news signalled by participants in the data collection campaign.

Appendix B: Parameters in Experiments

All experiments were conducted using two large language models (LLMs): *(i)* Llama 4 Maverick 400B

language	news domain	motivation
Italian	www.milanotoday.it	"unnecessarily mentions the origin of the people involved as if to emphasize the fact that it was not Italians who committed the crime but foreigners, and in this case it ferments prejudice"
Farsi	www.shomalnews.com	"it is fake because it says someone who died is actually alive"
English	www.foxnews.com	"While not outright lying, the wording of the title portrays the situation as sonething akin a violent gendered aggression, rather than a boxing match"
Russian	www.news.ru	"Highly misleading and inflammatory headline. The wording suggests that California is already leaving the U.S because it uses the present tense instead of a conditional or speculative form. The article inflates the importance of Calexit and is tailored to political agenda"

Table 10: Extract from answers received in the form used for the data collection.

and *(ii)* Mixtral Instruct 140B. Both models were accessed via the OpenRouter.ai API. To ensure reproducibility and eliminate ambiguity related to implicit platform defaults, we explicitly set all relevant generation parameters as follows:

- **Temperature:** 1.0
- **Top-p:** 1.0
- **Streaming:** disabled

The maximum token limits were determined according to the context window supported by each model as reported by OpenRouter at the time of experimentation:

- **Llama 4 Maverick 400B:** 1.05M tokens context window
- **Mixtral 8x22B Instruct:** 65.5K tokens context window

For each experiment, the effective maximum generation length was bounded by the remaining available context window after accounting for prompt tokens. Llama 4 Maverick is a high-capacity multimodal model developed by Meta, built on a Mixture-of-Experts (MoE) architecture with 128 experts and 17 billion active parameters per forward pass (400B total parameters). Mixtral 8x22B Instruct is Mistral AI's instruction-tuned MoE model, using 39 billion active parameters out of 141 billion total parameters.

Appendix C: Disaggregated Experimental Results

News Severity Classification Performance

In Table 11 we report the performance of LLAMA 4 Maverick in the news severity classification task. Results are broken down by single annotator, and information about their native language and gender is reported.

Annotator	F-Score	Language	Gender
14	0.379656	ru	m
16	0.405607	fa	f
17	0.501749	fa	m
18	0.340803	it	f
19	0.282989	it	f
20	0.460000	it	f
23	0.532380	fa	m
61	0.385902	it	m

Table 11: Annotator performance with gender and country of origin.

Downstream Evaluation of Text Generation

Table 12 reports the downstream evaluation of models in the task of generating explanations about problematic messages conveyed by news. Results are broken down by single annotator, and information about their native language and gender is reported.

The analysis of classification results along the gender axis shows that LLaMA tends to be more aligned with men than women. The average F-score is 0.445 for the former and 0.37 for the latter. A high variability between groups is observable as well. Annotations of people with Farsi as their first language seem to be more aligned with LLaMA than Italian (0.47 *versus* 0.36). Women 2.4 *versus* 3.28

Annotator ID	Evaluation	Language	Gender
15	3.14	ru	f
16	3.64	fa	f
17	3.58	fa	m
18	2.76	it	f
19	2.16	it	f
21	1.64	it	f
23	3.38	fa	m
24	2.74	it	f
61	2.90	it	m

Table 12: Average Evaluation per Annotator