International Journal of Communication 19(2025), 565–591 1932–8036/20250005

The Thin Line Between Conspiracy Theories and Opinion: Why Humans and AI Struggle to Differentiate Them

PAULA CARVALHO[1]
INESC-ID Lisboa, Portugal
Universidade de Aveiro, Portugal

DANIELLE CALED
INESC-ID Lisboa, Portugal

MÁRIO J. SILVA
INESC-ID Lisboa, Portugal
Universidade de Lisboa, Portugal

This study addresses the challenges in identifying online conspiracy theories, both by humans and automated systems. It relies on a corpus comprising conspiracy, news, and opinion articles gathered from the Portuguese blogosphere. Each article underwent evaluation through (i) InfoRadar, a multidimensional article characterization tool; (ii) ChatGPT 3.5, which involved an in-depth content analysis for key classification features; and (iii) a survey in which a group of online readers assessed various indicators for article categorization and credibility assessment. The mixed-methods approach highlights the difficulties faced by both humans and machines in differentiating conspiracy from opinion articles and provides valuable insights for distinguishing these categories. This research not only enhances our understanding of credibility perception in content marked by information disorder but also offers insights for developing transparent and explainable tools for critically assessing conspiracy theories.

Keywords: conspiracy theories, conspiracy detection, generative AI, information disorder, opinion

Paula Carvalho: pcc@ua.pt
Danielle Caled: dcaled@gmail.com
Mário J. Silva: mjs@inesc-id.pt
Date submitted: 2023-09-14

[1] This work was supported by the Fundação para a Ciência e a Tecnologia (FCT), under projects EXPL/LLT-LIN/1104/2021, UIDB/50021/2020, and SFRH/BD/145561/2019. The research was also co-funded by COMPETE 2020, Portugal 2020, and FEDER, under project POCI-05-5762-FSE-000217.

The information disorder landscape on media platforms encompasses a wide range of deliberately manipulated content, including conspiracy theories (CTs; Wardle & Derakhshan, 2017). These narratives offer alternative explanations for historical or ongoing events and often involve the belief in covert actions by influential groups to further their interests at the expense of the common good (Uscinski, 2018). Unlike straightforward false information, CTs may not be entirely false but often involve speculative elements and interpretations that are difficult to verify. This intertwining of factual information about known events with speculative motives of alleged conspirators makes it challenging to assess the overall credibility of conspiracy claims (Brotherton & Son, 2021). Consequently, most misinformation detection approaches, particularly binary classification solutions that attempt to distinguish between true and misleading information, often fail to effectively address these intricate narratives.

The challenges in identifying CTs extend to humans, as these theories often blur the line between fact and opinion. Readers frequently find it difficult to differentiate between opinion- and fact-based claims (Mitchell, Gottfried, Barthel, Sumida, & Mitchell, 2018). Additionally, they tend to label statements they subjectively agree with or believe as facts and, to a lesser extent, as opinions (Brotherton & Son, 2021; Walter & Salovich, 2021). Furthermore, psychological factors like *denialism* (i.e., the predisposition to reject authoritative information) and *conspiracy thinking* (i.e., the predisposition to attribute major events to conspiracies) heavily influence how individuals adhere to these narratives (Uscinski et al., 2020). Moreover, partisan motivations, particularly those linked to populism, can heavily influence individuals' reactions to CTs (Hameleers & van der Meer, 2021; Imhoff et al., 2022; van Prooijen et al., 2022; van Prooijen & Douglas, 2018).

The primary goal of this study is to thoroughly examine how humans and automated systems identify CTs disseminated through digital media platforms, particularly when they are intertwined with objective reporting and opinionated content. In pursuit of this goal, our investigation assesses the accuracy of different computational methods in distinguishing conspiracy from news and opinion articles, which inherently represent factual and subjective content, respectively. Specifically, we used InfoRadar,[2] a multidimensional article characterization tool developed to combat misinformation in Portuguese (Caled, Carvalho, Sousa, & Silva, 2024), along with ChatGPT 3.5 (OpenAI, 2023), for automated evaluations on a curated corpus of Portuguese digital media articles spanning *conspiracy*, *opinion*, and *news* categories. In addition, we conducted a detailed inductive content analysis (ICA) of the outputs generated by ChatGPT 3.5 to investigate the specific features used by the system in evaluating the credibility of the articles, considering their label (*conspiracy*, *opinion*, or *news*) in the corpus. Along with the ICA, we present the results of a survey among online readers using the same corpus. Beyond focusing solely on facets connected to the presentation of CTs, the survey enables the capture of an extensive assessment of various indicators recognized as pertinent in appraising article credibility (Molina, Sundar, Le, & Lee, 2021).

Overall, our study aims to deepen our understanding of the challenges faced by both humans and computational systems in identifying CTs. It explores the complexities arising from overlapping features across different types of articles and investigates how these narratives manifest within the digital media landscape, with a specific focus on the under-researched context of Portugal. By addressing this critical

[2] InfoRadar is available at: https://inforadar.inesc-id.pt

research gap, we seek to uncover persistent obstacles in CT identification, enrich our comprehension of the diverse manifestations of CTs, and contribute to the development of more accurate and transparent misinformation detection solutions.

Literature Review

The research landscape surrounding CTs within digital media has undergone substantial growth in recent years, particularly across fields such as social psychology, sociology, political science, communication, and media studies (Butter & Knight, 2020; Douglas et al., 2019; Mahl, Schäfer, & Zeng, 2023; Uscinski, 2018). However, as noted by Mahl et al. (2023), much of this research has focused primarily on major social media platforms, potentially limiting our understanding of their broader manifestation. The authors also highlight that many studies have primarily examined English content, potentially missing the nuances of CTs in different languages and pragmatic contexts. Additionally, they stress that existing research often focuses on individual conspiracy subjects, which, while providing valuable insights into conspiracy narratives, may not fully capture the complex interactions between various CTs within the digital context.

Recent studies suggest that CTs exhibit specific linguistic and discursive features that reinforce social identities and ideological divisions within conspiracy discourse (Chen, Chen, Zhang, Meng, & Shen, 2023; Demata, Zorzi, & Zottola, 2022). For example, these narratives often involve repetition and intertextuality (Campolong, 2022), use national identity language for in-group favoritism and out-group derogation (Chen et al., 2023) and employ polarized discourse to accentuate ideological divisions (Marko, 2022). Furthermore, the literature has emphasized that conspiracy theorists effectively engage their audience by tapping into negative emotions, such as anger, often directed toward established political systems and actors (Fong, Roozenbeek, Goldwert, Rathje, & van der Linden, 2021; Jolley & Paterson, 2020).

Although identifying the linguistic, discursive, and contextual features of conspiracy narratives is essential for developing explainable AI systems (Athira, Kumar, & Chacko, 2023), these features are typically analyzed in isolation without considering other content types. This narrow focus limits our understanding of how CTs may share common traits with other content types, such as opinion articles, to which they are often mistakenly attributed (Caled et al., 2024).

Research on automated CT detection remains an emerging field, and it is important to acknowledge the limited scope of existing studies. For instance, Tangherlini, Shahsavari, Shahbazi, Ebrahimzadeh, and Roychowdhury (2020) used machine learning (ML) to analyze CT narrative structures, while Shahsavari, Holur, Wang, Tangherlini, and Roychowdhury (2020) applied narrative theory and ML to uncover foundational structures in CT narratives across social media and news reports. These studies reveal how these narratives often connect seemingly unrelated knowledge domains. In contrast, Giachanou, Rosso, and Crestani (2023) proposed *ConspiDetector*, a model based on a convolutional neural network (CNN) that combines word embeddings with psycho-linguistic characteristics extracted from a collection of tweets posted by conspiracy and anticonspiracy propagators. The authors demonstrate that psycho-linguistic attributes, including personality traits, emotions, sentiment, and linguistic patterns, play a crucial role in distinguishing between individuals who promote a CT and those

who oppose it. Notably, one observation from this study is that conspiracy propagators tend to use more profanity than their anticonspiracy counterparts.

Despite the significance of these studies for CT detection, there remains a significant gap in understanding how human readers and automated systems identify specific conspiracy features. Current approaches often prioritize the assessment of conspiracy cues from either the perspective of humans (e.g., Lischka, 2024) or machines (e.g., Giachanou, Rosso, & Crestani, 2019), rather than integrating both perspectives simultaneously. This dual approach is essential for comprehending how both humans and machines discern CTs and for identifying the primary challenges in accurately identifying relevant features.

In this study, we aim to address some of the previously mentioned gaps by analyzing a collection of full-length articles from the Portuguese blogosphere covering various topics. Ultimately, the study seeks to answer the following research questions:

RQ1: To what extent do humans and machines face challenges in distinguishing conspiracies from news and opinion articles?

RQ2: What are the primary distinguishing features that set apart conspiracy from news and opinion articles?

By broadening the analysis to include not only conspiracy articles but also opinion pieces and news stories, this study enhances our understanding of the complex intersections between these content types. This is particularly important because traditional disinformation detection systems often treat these categories in isolation, overlooking the nuanced ways they overlap and influence each other, thus complicating their identification. Furthermore, our study advances the theoretical understanding of how content from different genres is perceived by both humans and machines, while offering a fresh perspective on the most relevant credibility indicators for each category. Specifically, by identifying the key characteristics for recognizing CTs, this research will contribute to the development of more accurate, transparent, and explainable systems. In turn, these systems could empower users to critically access and effectively evaluate online content.

Materials and Methods

Data Selection

We gathered a corpus consisting of 81 articles from a combination of 23 Portuguese mainstream (M) and non-mainstream (NM) media outlets (Table 1). In detail, the corpus comprises 27 news articles, 27 opinion pieces, and 27 conspiracy narratives.[3] Apart from a single news article reporting an event that occurred in 2000, all articles were published during the period spanning from 2020 to 2022, aiming to maximize coverage of events related to the topics. Furthermore, this time frame aligns with the collection period of the MINT corpus, which was used to train InfoRadar (Caled, Carvalho, & Silva, 2022).

[3] The data set is available at: https://github.com/dcaled/news_opinion_conspiracy_dataset/

Table 1. Distribution of the Articles' Sources by Category.

Category	Media Outlet	M/NM	# Articles	#Tokens
News	Diário de Notícias	M	8	
	CNN Portugal	M	1	
	Expresso	M	1	
	Jornal de Negócios	M	3	
	Jornal de Notícias	M	1	
	Observador	M	3	
	Público	M	5	
	RTP Notícias	M	2	
	TSF	M	2	
	Visão	M	1	
Subtotal			27	118,921
Opinion	Diário de Notícias	M	3	
	Esquerda Net	NM	1	
	Expresso	M	3	
	Jornal Inevitável	NM	1	
	Jornal Médico	NM	2	
	Observador	M	1	
	Página Um	NM	1	
	Público	M	1	
	Rádio Renascença	M	2	
	Sapo Atualidade	M	1	
	Sic Notícias	M	2	
	Sol	M	2	
	Visão	M	7	
Subtotal			27	164,716
Conspiracy	Casa das Aranhas	NM	3	
	O Diário de um ET	NM	8	
	Portugal Misterioso	NM	3	
	Resisitir.info	NM	8	
	O Evento	NM	5	
Subtotal			27	374,306
Total			**81**	**657,943**

Conspiracy articles were curated from five websites previously recognized as disseminators of CTs (Caled et al., 2022). Each article underwent validation by the research team, adhering to Uscinski's (2018) definition of conspiracy theories: narratives explaining historical, ongoing, or future events by attributing primary causality to a covert group of powerful individuals—the conspirators—acting in secrecy for personal gain at the expense of collective welfare. To ensure a diverse range of topics, we selected articles covering health- and science-related conspiracy subjects, such as climate change, population demographics, and the COVID-19 pandemic. We also included articles concerning geopolitical events, such as the Russia–Ukraine war, and specific incidents involving international and national figures, like the Travis Scott concert incident

at the 2021 Astroworld Festival in Houston and the controversial 1980 Camarate (Lisbon) air crash. It is important to note that these topics are merely indicative, as CTs often encompass multiple narratives simultaneously rather than focusing on a single topic (Mahl et al., 2023).

We started by assigning three keywords to each selected CT article, reflecting the main themes or entities they addressed. These keywords were then used to retrieve semantically related news and opinion pieces from mainstream and non-mainstream media sources, including newspapers, TV, radio broadcasts, and magazines. News articles were sourced from widely read Portuguese mainstream digital media, while opinion pieces included both mainstream and non-mainstream outlets to ensure diverse perspectives and reduced bias. As pointed out by Nekmat (2020), non-mainstream outlets often lack established reputation, brand recognition, and traditional structural and editorial features, which may influence their perceived trustworthiness.

Assessment Tools

We selected InfoRadar as our primary tool because it is specifically tailored for Portuguese and designed to automatically classify diverse content types, including news, opinion pieces, and CTs (Caled et al., 2024). ChatGPT was used because of its success in detecting fake news (Caramancion, 2023) and its potential in identifying CTs (Pustet, Steffen, & Mihaljević, 2024). Both tools leverage cutting-edge Transformer technology and offer user-friendly features and APIs for effective analysis and interpretation.

InfoRadar

InfoRadar provides a multidimensional indicator that automatically scores a submitted article under the following categories: *hard news*, *soft news*, *opinion*, *satire*, and *conspiracy theories*. To obtain the multidimensional indicator, a model was trained on the MINT corpus, which refers to a comprehensive collection of 20,278 articles sourced from 33 Portuguese media outlets over the course of a year (Caled et al., 2022). The corpus was categorized into these five distinct categories, serving as the basis for training InfoRadar's classifiers and enabling the system to generate scoring metrics (Caled et al., 2024). The model was built on a pretrained multilingual BERT model and fine-tuned with a linear layer on top of the pooled output (Devlin, Chang, Lee, & Toutanova, 2019).

To use InfoRadar, we entered the article's URL, and it generated a multidimensional indicator and metric scores for categorization and credibility assessment. The system includes five assessment panes, as illustrated in Figure 1 (Caled et al., 2024):

1. *Article content and metadata*: displays the input URL, a thumbnail of the main image associated with the article, and the headline and body text.
2. *Article source*: provides information on the source's verification status, and it may include additional details when accessible, such as the official organization name, geographical location, and registration date with the Portuguese Regulatory Authority for the Media.
3. *Multidimensional indicator*: presents a graphical representation of the five-dimensional indicator for article category prediction. A red pointer indicates the classifier's confidence score in each

category, identified as high (score > 0.75), medium-high (0.50 < score < 0.75), medium-low (0.25 < score < 0.50), or low (score < 0.25).

4. *Explanatory metrics*: provides information from five explanatory metrics often highlighted in the literature as evidence of deviation from normative journalism (Molina et al., 2021). These include *sentiment*, *subjectivity*, *headline sensationalism*, *headline representativeness*, and the use of *unfamiliar words*. For each metric, InfoRadar (a) computes a score for the input article, and (b) determines the article's percentile rank, positioning the analyzed article relative to each <collection, metric> pair in the reference corpus.
5. *Explainer*: presents a short summary generated based on the information provided by the multidimensional indicator and explanatory metrics.

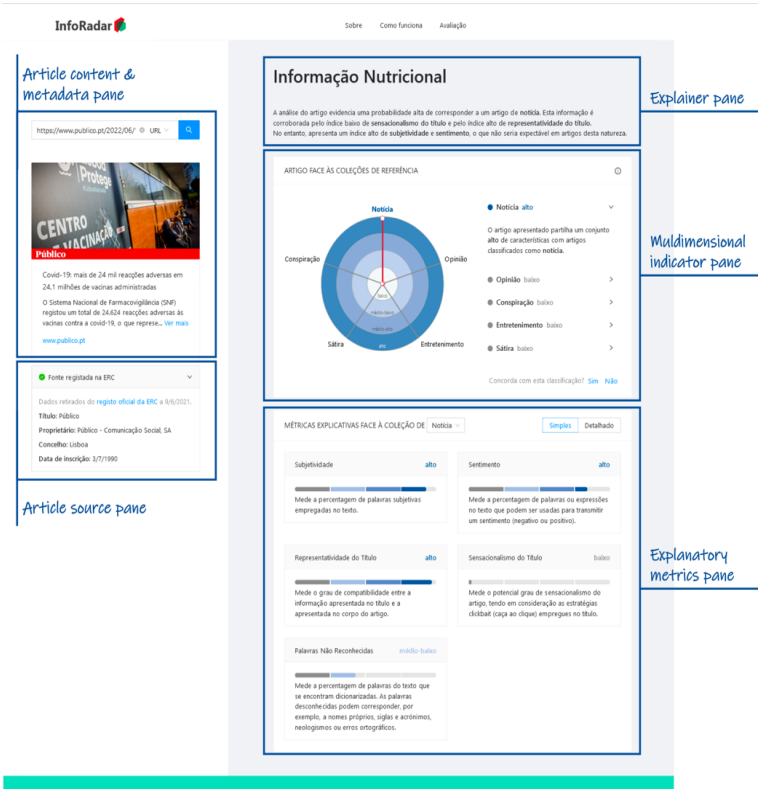


Figure 1. InfoRadar assessment panel. Screenshot from an illustrative article analyzed by InfoRadar (Caled et al., 2024).

We used InfoRadar to obtain automated assessments of the likelihood of each article in our corpus falling into specific categories indicated by the multidimensional indicator. Each article was assigned a confidence score indicating its likelihood of belonging to each of the five categories analyzed. The sum of the scores for a given article is 100%. Hence, confidence scores greater than 0.50 indicate that the selected class has higher confidence than all the other classes combined. We focused on categories assigned to articles with a confidence score of 0.50 or higher. Furthermore, we considered the explanatory metrics

assigned to each article. The system categorizes these metrics on a percentile scale of low, medium-low, medium-high, and high, which we represent as numerical values of 1, 2, 3, and 4, respectively.

ChatGPT 3.5

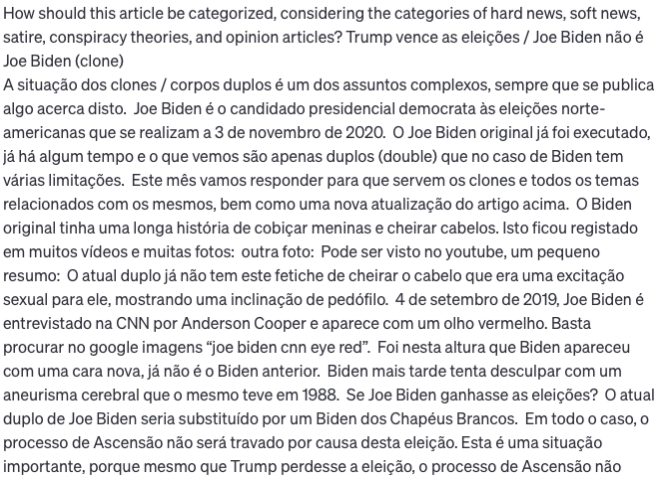
The evaluation of the corpus using ChatGPT was conducted by taking the headline plus the body text of each article as input to ChatGPT's text-based interface, as illustrated in Figure 2 (OpenAI, 2023). Two specific chat prompts were formulated as questions:

Prompt 1: How should this article be categorized, considering the categories of hard news, soft news, satire, conspiracy theories, and opinion articles?

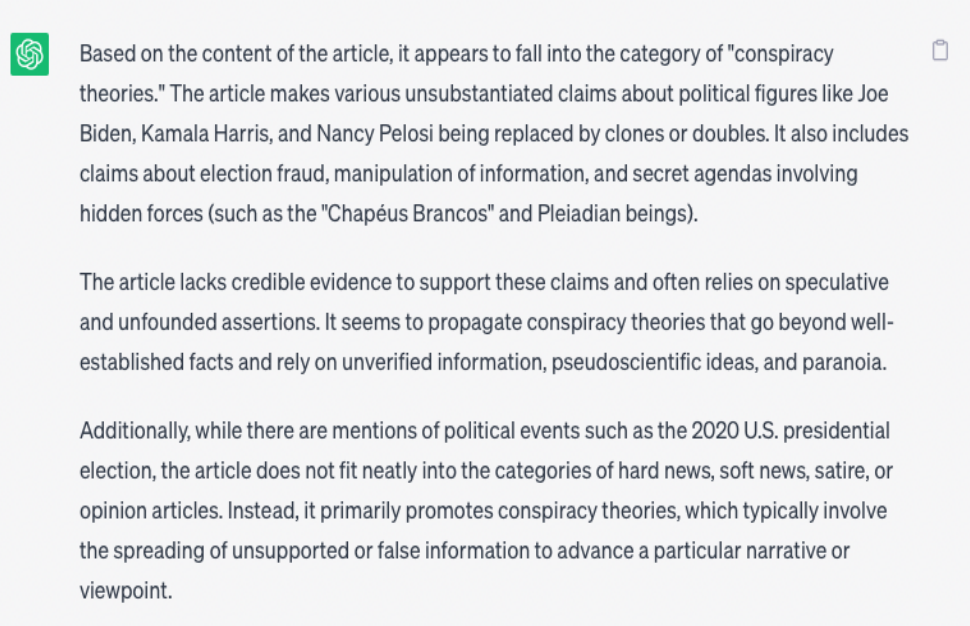
Prompt 2: What are the main textual indicators that could support the analysis of this article as news, opinion, or conspiracy theory?

Prompt 1 includes all InfoRadar's categories to ensure a fair comparison between the two tools. In contrast, to reduce bias in the results, Prompt 2 deliberately avoids mentioning specific indicators or characteristics (such as those used in InfoRadar or the survey presented to online readers). Although we recognize that this approach may limit direct comparisons across different applications, we believe it helps prevent question framing from influencing responses and allows for deeper insights into the articles being evaluated.

We observed that the generated answers may vary across different runs or interactions with the chatbot because of the probabilistic nature and inherent randomness of the system. Although we registered some variation in the explanations provided, the main results from the chatbot remained consistent after multiple runs. We considered only the responses generated by the initial run of the tool, executed between March 27 and April 4, 2023. When articles exceeded the maximum text length allowed for ChatGPT 3.5 (4,096 tokens), we truncated them to fit within the limit.


How should this article be categorized, considering the categories of hard news, soft news, satire, conspiracy theories, and opinion articles? Trump vence as eleições / Joe Biden não é Joe Biden (clone)
A situação dos clones / corpos duplos é um dos assuntos complexos, sempre que se publica algo acerca disto. Joe Biden é o candidado presidencial democrata às eleições norte-americanas que se realizam a 3 de novembro de 2020. O Joe Biden original já foi executado, já há algum tempo e o que vemos são apenas duplos (double) que no caso de Biden tem várias limitações. Este mês vamos responder para que servem os clones e todos os temas relacionados com os mesmos, bem como uma nova atualização do artigo acima. O Biden original tinha uma longa história de cobiçar meninas e cheirar cabelos. Isto ficou registado em muitos vídeos e muitas fotos: outra foto: Pode ser visto no youtube, um pequeno resumo: O atual duplo já não tem este fetiche de cheirar o cabelo que era uma excitação sexual para ele, mostrando uma inclinação de pedófilo. 4 de setembro de 2019, Joe Biden é entrevistado na CNN por Anderson Cooper e aparece com um olho vermelho. Basta procurar no google imagens "joe biden cnn eye red". Foi nesta altura que Biden apareceu com uma cara nova, já não é o Biden anterior. Biden mais tarde tenta desculpar com um aneurisma cerebral que o mesmo teve em 1988. Se Joe Biden ganhasse as eleições? O atual duplo de Joe Biden seria substituído por um Biden dos Chapéus Brancos. Em todo o caso, o processo de Ascensão não será travado por causa desta eleição. Esta é uma situação importante, porque mesmo que Trump perdesse a eleição, o processo de Ascensão não

(a) Reply to Prompt 1 (the submitted article is truncated in the image).



(b) Reply to Prompt 2.

Figure 2. Illustration of Prompt 1, along with ChatGPT's output. Screenshots from ChatGPT 3.5 (OpenAI, 2023).

Inductive Content Analysis

The responses to Prompt 2 of ChatGPT were imported into MAXQDA (VERBI Software, 2021), where an ICA was performed on the data. ICA involves the iterative development of codes and subcodes, based on the data, as outlined by Vears and Gillam (2022). The coding process was led by the first author, with the annotation system and final annotations discussed and harmonized jointly with the other authors. This collaborative approach aimed to reconcile potential differences and strengthen the robustness of our analysis.

We began by reading and familiarizing ourselves with the text to gain a comprehensive understanding. In the first round of analysis, we identified broad categories that were potentially relevant. In the second round, we further refined the analysis by developing subcategories and fine-grained codes within those broad categories. The coding process was iterative, with comparisons between texts helping to ensure that important codes were not overlooked. We refined the fine-grained subcategories and removed any overlaps or redundancies to improve clarity. Finally, we synthesized and interpreted the coded data. Following discussions and joint harmonization among the authors, this process yielded a total of 15 codes and 48 subcodes, organized into three primary categories: news, opinion, and conspiracy. Table 2 offers a comprehensive overview of our code system, detailing the frequency of occurrences for each subcode. Additionally, Figure 3 illustrates an instance where ChatGPT's response has been annotated using this coding system, as implemented in MAXQDA.

Table 2. Code System and Number of Responses Assigned to Each Subcode.

	Code	Subcode	#
News	Structure	Inverted pyramid	1
	Headline	Clear, concise, and factual	27
	Language	Clear and concise	23
		Objective	27
	Timeliness	Present and immediate past	4
		Present and immediate future	23
	Sources	Official sources and authorities	21
		Factual data and statistics	4
		Scientific research and studies	7
Opinion	Headline	Suggestive and provocative	6
	Language	1st Person personal pronouns	10
		Subjective lexicon	21
		Metaphor and analogy	2
		Rhetorical questions	2
		Irony	1
		Hyperbole	3
	Discourse and argumentation	One-sided perspective	8
		Persuasion tactics	6
		Lack of factual reporting and empirical evidence	6
		Personal views and opinions	22
	Sources	Lack of references and quotes from experts	10
	Metadata	Author identity	9
		Section identification	3
Conspiracy	Themes	Hidden agenda or secret plan	16
		Supra-natural and malevolent forces	5
		Powerful groups or secret societies	15
		New world order	5
	Language	Emotionally charged	17
		Vague and ambiguous	3
		Pejorative	1
		Hyperbole	3
		Rhetorical questions	1
	Discourse and argumentation	Lack of critical analysis	2
		Contradictions and inconsistencies	2
		Oversimplification	6
		Speculation and conjecture	7
		Vague and unsupported claims	20
		Pseudoscience claims	2
		Rejection of mainstream information	6
		Paranoia	3
		Victimization	1

Fallacies	Appeal to action	6
	Appeal to authority	1
	Appeal to emotion	6
	Appeal to fear	9
	Appeal to hidden or secret knowledge	7
Sources	Unreliable, misleading, or biased	14
	Anonymous	10

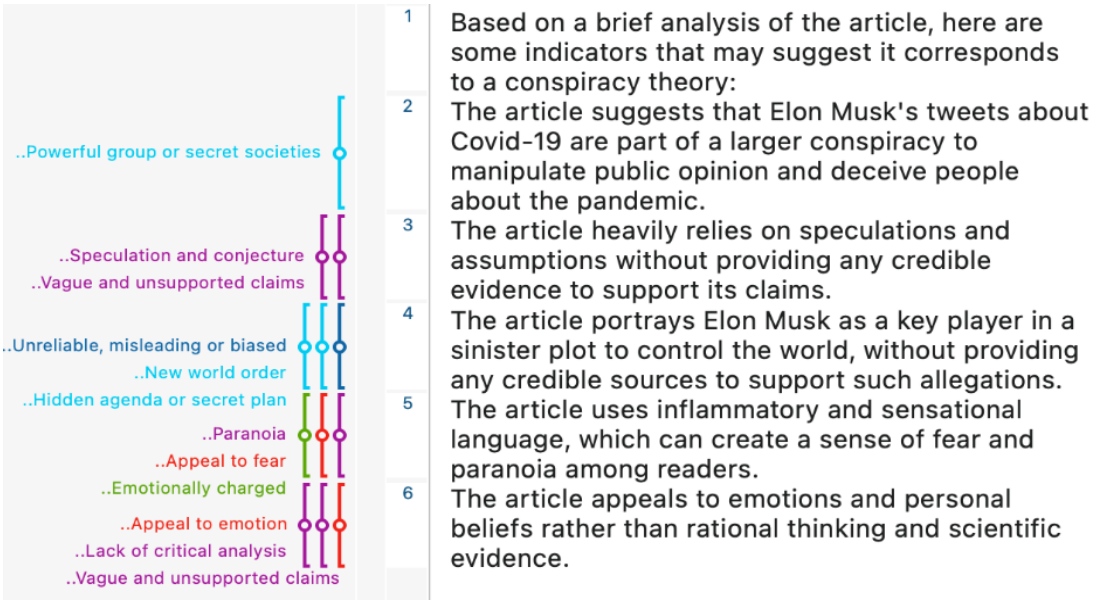


Figure 3. Illustration of an annotated response from ChatGPT using the coding system. Screenshot from the MAXQDA software interface (VERBI Software, 2021).

A group of 12 online readers, recruited by a survey company, participated in the analysis of each article in the corpus through a survey. In detail, the study included five males, six females, and one nonbinary participant, all of whom were Portuguese. As detailed in Table 3, the participants were distributed across different age groups and educational backgrounds. Before engaging with the survey questions, participants were requested to complete a sociodemographic survey. To ensure ethical compliance, they were also required to read and sign a consent form that had received prior approval from our institutional ethics committee.

Table 3. Profile of Survey Participants.

ID	Age Range	Gender	Education
1	26–40	Female	Secondary
2	41–60	Male	Master
3	41–60	Male	Secondary
4	18–25	Female	Secondary
5	18–25	Female	Bachelor
6	18–25	Male	Secondary
7	18–25	Male	Bachelor
8	18–25	Female	Master
9	26–40	Male	Secondary
10	26–40	Other	Secondary
11	18–25	Female	Secondary
12	41–60	Female	Secondary

The survey consisted of multiple-choice, dichotomous, and Likert-scale questions, organized into eight dimensions, as outlined below[4]:

1) *Article reading.* Participants were given access to the article's text-only (images, metadata, or hyperlinks excluded). This aimed to eliminate potential evaluation biases, as prior studies have indicated the substantial impact of these elements on credibility assessment (Viviani & Pasi, 2017).

2) *Article classification and overall credibility perception*. Participants were tasked with distinguishing between *news*, *opinion*, *conspiracy*, or any "*other"* category, and then providing their perception of the overall credibility of the article. This aimed to understand how readers categorized different types of content and assessed their reliability within the context of the study.

3) *Headline assessment*. Participants were prompted to evaluate the clarity and accuracy of the article's headline. This process aimed to facilitate the identification of sensationalist headlines, which may have the potential to diminish perceptions of credibility and content quality (Luo, Hancock, & Markowitz, 2020; Molyneux & Coddington, 2020).

4) *Article consistency, references, and sources.* Participants were asked to evaluate the logical coherence and cohesion of the article's narrative, the inclusion of specific references to time and location, and the presence of citations. These factors serve as key indicators for distinguishing between credible news and misleading content (Molina et al., 2021).

5) *Article objectivity.* Participants were asked to assess the subjectivity level of the article and determine the prevalence of facts or opinions within the article. This aspect was important to

[4] The survey can be assessed at: https://inforadar.inesc-id.pt/avaliacao

evaluate the reader's ability to discern between opinion and fact-based claims within the article (Walter & Salovich, 2021).

6) *Rhetorical strategies.* Participants were asked to identify specific rhetorical and discursive strategies commonly associated with information disorder, such as irony, sarcasm, and humor (Rubin, Conroy, Chen, & Cornwell, 2016). In addition, they were asked to identify classical fallacies, including (i) personal attack (or *ad hominem;* Tindale, 2007); (ii) appeal to fear (or *ad baculum argument*; Tindale, 2007); and (iii) call to action (Carvalho, Caled, Silva, Batista, & Ribeiro, 2024).

7) *Conspiracy narrative.* Participants were asked to identify elements commonly found in conspiracy narratives, such as references to secret societies or groups, suggestions of powerful or malicious forces behind events, the presence of opposing factions, and the author's explicit intention to reveal disruptive or threatening truths (Douglas et al., 2019).

8) *Sentiment and emotion.* Participants were asked to assess sentiment polarity, intensity, and primary emotions in the article, as conspiracy beliefs are often associated with psychological states linked to negative emotions (Douglas et al., 2019).

Responses related to article categorization (Dimension 2) allow us to compare assessments made by humans, InfoRadar, and ChatGPT (RQ1). The information from the other dimensions provides valuable insights for addressing RQ2.

Results

This section presents the results of the corpus assessment, including the outcomes provided by InfoRadar and ChatGPT, along with the outcomes from the online readers survey. We considered the total number of survey responses (i.e., a total of 972 responses, corresponding to 324 responses per article category). To assess task complexity and subjectivity, we computed Krippendorff's alpha coefficient for estimating inter-annotator agreement (IAA; Krippendorff, 2007). Additionally, we showcased the outcomes obtained from the ICA of the ChatGPT output for Prompt 2. To enhance readability, this section is organized according to the research questions guiding our study.

Distinguishing Conspiracy From News and Opinion Articles

Figure 4 illustrates the classification outcomes for news, opinion, and conspiracy articles within the corpus, as determined by InfoRadar, ChatGPT 3.5, and the online readers who participated in our study. The results reveal that while news and opinion articles are generally distinguishable, conspiracy articles tend to be mistaken as opinions by both human readers and ML tools.

Interestingly, while not specifically designed for misinformation detection like InfoRadar, ChatGPT 3.5 exhibits a slightly superior performance across all categories, including *conspiracy*. Human readers face distinct challenges in identifying opinion articles as opposed to automated systems, demonstrating comparatively lower proficiency in this specific scenario. However, humans also show a higher proficiency

in detecting conspiracy articles when compared with both automated solutions. Furthermore, within the pool of misclassified articles, computational tools tend to categorize conspiracy articles as opinion pieces, whereas readers classify them as news.

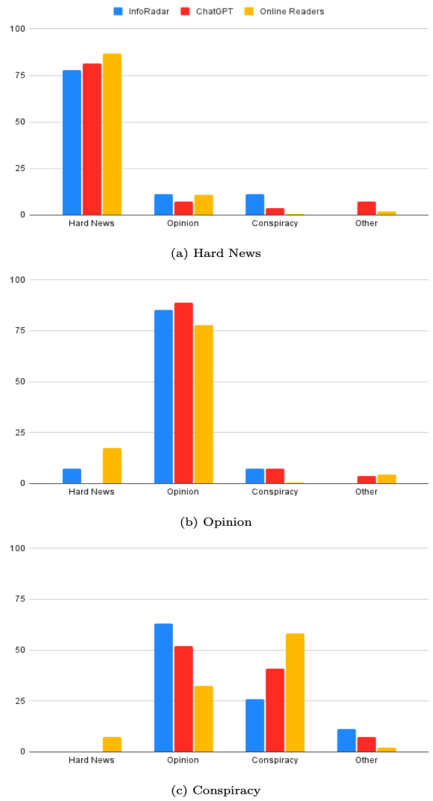


Figure 4. Compilation of responses (%) from InfoRadar, ChatGPT, and online readers on the likelihood of the article belonging to the categories of news, opinion, or conspiracy. The "Other" category includes "soft news" and "satire" as predicted by InfoRadar, as well as any other categories identified by ChatGPT and readers.

Despite these challenges, the IAA results consistently demonstrate a substantial level of consensus among online readers in terms of category assignments ($\alpha \geq 0.61$). However, the agreement diminishes to 0.47 between automated solutions, highlighting the tendency for human consensus to surpass that of automated systems (Table 4). An analysis of human reader responses in relation to each system indicates a discernible inclination toward stronger agreement with the outputs from ChatGPT than with those from InfoRadar.

Table 4. Krippendorff''s Alpha (α) Results for Inter-Annotator Agreement (IAA) on Category Assignment Between Online Readers and ML Tools.

IAA	A
Online readers	0.614
Online readers x InfoRadar	0.361
Online readers x ChatGPT	0.507
InfoRadar x ChatGPT	0.473

Key Differences Between Conspiracy, News and Opinion Articles

Table 5 presents the average metric scores generated by InfoRadar for conspiracy articles, contrasting these scores with those for articles classified as news and opinion.

Table 5. InfoRadar Scores (Average) for Each Category in the Corpus.

Metric	News	Opinion	Conspiracy
Subjectivity	1.56	3.04	2.26
Sentiment	1.67	2.74	2.56
Headline representativeness	3.52	3.67	3.52
Headline sensationalism	1.85	2.63	2.67
Unfamiliar words	1.44	1.19	2.19

Conspiracy articles exhibit a moderate level of subjectivity, higher than news articles but slightly lower than opinion pieces. As expected, they contain more sentiment words than news articles but fewer than opinion articles, possibly because of the use of subtle emotional cues not fully captured by sentiment lexicons. Additionally, conspiracy articles feature a higher number of unfamiliar words compared with news, though they are slightly fewer than opinion pieces. When it comes to headlines, both conspiracy and opinion articles tend to be more sensationalist compared with news articles. Despite this sensationalism, the average headline representativeness scores are quite similar across all article types in our corpus. Overall, news articles exhibit the lowest levels of subjectivity and unfamiliar words, maintaining a more objective tone than both opinion and conspiracy pieces. In contrast, conspiracy and opinion articles share more characteristics, indicating that the explanatory metrics currently provided by InfoRadar may not effectively help readers distinguish between these two categories.

Figure 5 presents the outcomes of the Likert-scale survey questions, capturing readers' perspectives on various factors that aid in distinguishing between different article categories. Specifically, the results were computed based on readers' mean ratings for headline representativeness, headline sensationalism, reliability of cited sources, objectivity, and sentiment intensity.

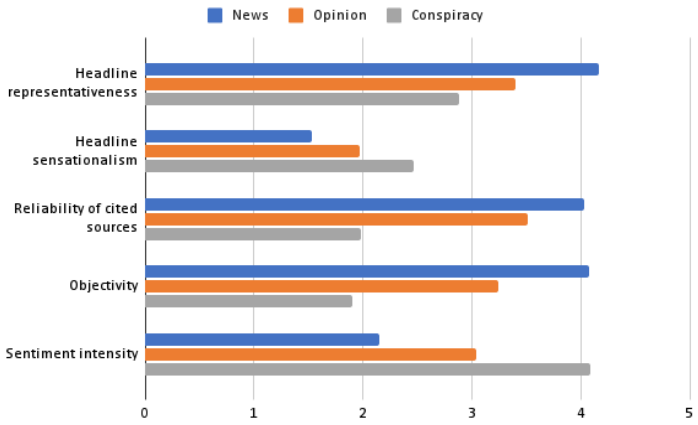


Figure 5. Readers' average ratings of headline representativeness, headline sensationalism, reliability of cited sources, objectivity, and sentiment intensity.

News headlines were rated as the most representative of their content, and opinion headlines received moderately positive ratings (above 3), while conspiracy headlines were viewed as less representative (below 3) and potentially misleading. All categories received negative (below 2) ratings for sensationalism or clickbait, with news headlines exhibiting the lowest level of sensationalism, aligning with InfoRadar's scores. Additionally, readers found the language in news articles to be more objective than in opinion and conspiracy articles. While opinion pieces received favorable (above 3) ratings for objectivity, conspiracy articles were viewed as lacking objectivity (below 3). This contrasts with InfoRadar's analysis, which categorized opinion articles as the most subjective pieces within the *corpus*. Furthermore, readers associated higher sentiment intensity with conspiracy articles compared with opinion articles, which averaged around 3, while news articles received negative ratings in this regard. Although InfoRadar does not measure sentiment intensity, both human readers and InfoRadar appear to recognize more sentiment and emotions in opinion and conspiracy articles than in news pieces, as expected. Finally, online readers generally assign greater credibility to sources cited in news articles compared with those cited in opinion pieces. In contrast, sources in conspiracy articles were perceived as unreliable (average rate of 2).

Delving deeper into the results from the readers' evaluations, we found that primary emotions were more prevalent in conspiracy articles, identified in approximately 75% of survey responses (see Table 6). Readers noted fear, disgust, and anger as the dominant emotions in these articles. In contrast, emotions in opinion articles were detected in 49% of the responses and evoked a broader range of negative emotions, including sadness, fear, and disgust.

Regarding the use of rhetorical devices and discursive strategies often associated with deceptive content, the survey results revealed that conspiracy articles exhibit a significantly higher occurrence of appeals to fear, calls to action, and personal attacks. Notably, personal attacks were identified in more than 80% of the survey assessments. In contrast, these strategies were sparse in news but more pronounced in opinion articles, where calls to action were evident in roughly half of the assessments.

Table 6. Proportion (%) of Specific Traits Identified by Category in the Corpus.

Dimension	Feature	News	Opinion	Conspiracy
Emotion	None	0.82	0.51	0.25
	Anger	0.00	0.03	0.22
	Disgust	0.01	0.10	0.23
	Enjoyment	0.02	0.00	0.00
	Sadness	0.03	0.20	0.03
	Fear	0.08	0.11	0.24
	Surprise	0.05	0.04	0.03
Rhetorical Strategies	Appeal to fear	0.16	0.25	0.59
	Call to action	0.10	0.45	0.51
	Personal attack	0.15	0.35	0.81
	Irony and sarcasm	0.01	0.12	0.20
Conspiracy Elements	Secret society	0.03	0.07	0.60
	Malevolent forces	0.09	0.14	0.76
	Threatening truths	0.02	0.05	0.44
	Us vs. Them	0.09	0.25	0.35
Other textual and discursive elements	Citations	0.88	0.50	0.51
	Consistency	0.85	0.80	0.36
	Time and space	0.75	0.44	0.37
	Facts	0.90	0.54	0.22
	Opinions	0.06	0.41	0.70

Furthermore, while elements typically associated with CTs were significantly more prevalent in conspiracy articles compared with other categories, some, such as the polarization of *in-groups* and *out-groups*, were also identified in opinion articles manifesting in 25% of the survey assessments.

Regarding elements highlighted in the literature as key indicators for distinguishing different types of online content, the survey results confirm that news articles included significantly more citations to support facts, hypotheses, or conclusions compared with opinion and conspiracy articles. As expected, readers observed that news articles predominantly conveyed facts, while they were inclined toward the *opinion* class in conspiracy articles. Interestingly, the prevalence of opinion expression was perceived to be more pronounced in conspiracy articles than within opinion articles. Explicit references to time and space were also more evident in news articles. Additionally, readers perceived news and opinion articles as presenting a coherent and cohesive narrative, but frequently identified conspiracy articles as lacking coherence or cohesion.

Overall, the agreement results were modest. The recognition of emotions by online readers showed particularly low consensus ($\alpha = 0.218$), highlighting the subjectivity of this task. Identifying specific rhetorical devices yielded IAA values ranging from 0.23 (call to action) to 0.41 (personal attack), reflecting the challenges readers face in detecting these strategies. When it comes to conspiracy elements, the *Us vs. Them* distinction had the lowest consensus ($\alpha = 0.272$). Readers also struggled with identifying temporal

and geographic references (α = 0.18) and assessing the text's coherence and consistency (α = 0.25). Although there was slightly higher agreement in evaluating the reliability of cited sources (α = 0.37), it remained relatively low. Finally, a moderate level of consensus was found in distinguishing between facts and opinions (α = 0.499).

Table 7 presents the distribution of primary dimensions (identified through ICA) in ChatGPT's outputs, along with the corresponding percentages for each article category, thereby complementing the results provided by InfoRadar and the online reader survey.

Language usage seems to play a significant role in distinguishing article categories, with mentions found across all categories. However, there is a higher prevalence of distinctive language features in news and opinion articles compared with conspiracy articles. ChatGPT effectively distinguishes between different article categories by focusing on specific language characteristics. It excels at identifying subjective language in opinion articles, objective lexicon in news articles, and emotionally charged language in conspiracy articles, as detailed in Table 2, globally aligning with the results provided by InfoRadar and the survey. Again, the features associated with discourse and argumentation vary depending on the article's category. For opinion articles, the most prevalent feature is the reference to personal views and opinions. In conspiracy articles, the focus is on the presence of vague and unsupported claims.

Table 7. Distribution (%) of the Main Classes Identified in ChatGPT's Outputs per Article Category.

Class	News	Opinion	Conspiracy
Structure	0.7	0.0	0.0
Headline	19.7	5.5	0.0
Language	36.5	35.8	14.9
Timeliness	19.7	0.0	0.0
Discourse and argumentation	0.0	38.5	29.2
Sources	23.4	9.2	14.3
Metadata	0.0	11.0	0.0
Themes	0.0	0.0	24.4
Fallacies	0.0	0.0	17.3

The explicit mention of information sources is most prevalent in responses addressing news articles, which corroborates the previous results. Further analysis reveals significant differences in how sources are addressed across article categories. For news articles, ChatGPT primarily focuses on the identification of official sources and authorities. In contrast, for *conspiracy*, the most common feature is the reliance on unreliable, misleading, or biased sources, which aligns with the overall perception of online readers. For opinion articles, the dominant aspect is the absence of references to sources and quotes from experts.

Timeliness also plays a key role in categorizing and identifying news articles, a pattern observed by online readers as well. Specifically, ChatGPT highlights the relevance of timely and up-to-date information in news articles, as it recognizes the use of present and immediate future tenses in most responses.

Regarding headlines, the model consistently identifies clear, concise, and factual headlines in responses about news articles. In contrast, for opinion articles, when headlines are mentioned, ChatGPT 3.5 tends to classify them as suggestive and provocative. These findings align with *InfoRadar*, which categorizes conspiracy headlines as the most sensationalist.

Conspiracy themes are present in more than 20% of the coded segments specifically related to the conspiracy articles. The most common themes in ChatGPT's responses include general references to a hidden agenda or secret plan and mentions of a powerful group or secret society, aligning again with online reader results. Fallacies were found only in conspiracy articles, with appeal to fear and appeal to hidden or secret knowledge being the most prevalent strategies reported in ChatGPT's responses. Although appeal to fear was also found to be one of the most present strategies by online readers, personal attacks emerged as the most prevalent category by human readers but had no expression in the ChatGPT outputs.

Discussion and Main Conclusions

Our study was primarily aimed at investigating the challenges faced by both human readers and automated models in distinguishing between news, opinion, and conspiracy articles (RQ1). The results of our experiments, which encompassed both manual and automated article categorization, strongly underscore the enduring difficulties associated with identifying CTs. The relatively limited agreement observed among online readers concerning the assessment of the articles' overall credibility and content further reinforces this finding.

Although readers can identify trustworthy content, such as news articles and opinion pieces, sometimes without accurately categorizing them, state-of-the-art AI tools might occasionally mistakenly classify credible information as deceptive. This suggests that subtle or even nonrelevant features shared between news, opinion, and conspiracy articles may confound machines.

Additionally, our study reveals that both humans and machines may fail to differentiate news from opinion articles and vice-versa. Interestingly, online readers exhibit higher proficiency in identifying news articles, while AI tools perform slightly better in discerning opinion pieces. This raises special concerns about individuals' media literacy in accurately distinguishing between facts and opinions and challenges the idea of a definitive boundary between these two concepts. Prior experiments have also demonstrated difficulties in differentiating opinion-based claims from fact-based claims, particularly when they align with individuals' pre-existing beliefs (Walter & Salovich, 2021).

Detecting CTs poses the most significant challenge for both human readers and AI tools. The best-performing model (ChatGPT 3.5) achieved only 44% accuracy, indicating considerable difficulties in differentiating *conspiracy* from other categories, especially *opinion*, which they are frequently mistaken for. Our findings align with previous reports on InfoRadar, where *conspiracy* was classified as the most challenging category to detect, often being confused with opinion articles (Caled et al., 2022; Caled et al., 2024). The consistency of these results with ChatGPT suggests that this aspect does not reflect a particular limitation or bias specific to InfoRadar, but rather a shared challenge among various automated systems that should be addressed in further research.

Although there is a lack of explicit data on ChatGPT 3.5's performance in identifying CTs, recent research has highlighted its potential in related tasks, such as distinguishing between accurate and false claims (Hoes, Altay, & Bermeo, 2023). Although the task undertaken in our experiments inherently holds greater complexity because of the nuanced characteristics of CTs and their potential intersections with factual data and personal viewpoints, our findings point to the difficulties that generative AI may encounter when attempting to address misinformation adeptly.

The examination of survey outcomes, coupled with the ICA of ChatGPT output for Prompt 2, has furnished significant insights into the distinguishing attributes separating *conspiracy* from the other categories examined in this research, chiefly opinion articles (RQ2).

As highlighted by Caled et al. (2022), conspiracy and opinion articles may exhibit linguistic and discursive characteristics in common, such as subjectivity and the use of fallacious arguments, making their differentiation complex. However, a closer examination of our findings uncovers nuanced expressions and interpretations of these traits. In relation to subjectivity, it is important to note that the distinguishing factor between conspiracy and opinion articles does not rest solely on the usage of subjectivity. Instead, it hinges on the use of emotionally charged language. In our corpus, CTs frequently manifest negative emotions, such as fear and anger, which are consistent with earlier research associating these emotions with conspiracy beliefs (Mitra, Counts, & Pennebaker, 2021; van Prooijen & Douglas, 2018). Our analysis identifies anger as a key distinguishing feature between conspiracy and opinion articles. This is consistent with the findings of Jolley and Paterson (2020), who note that CTs often direct anger toward a scapegoat, justifying aggressive behavior toward political systems and figures. Fong et al. (2021) also emphasize anger as a pivotal emotional trait, distinguishing language usage between individuals engaging with conspiratorial content and those focusing on scientific discourse. More recently, Korenčić et al. (2024) further confirm that anger is a key factor in differentiating between conspiracy and critical narratives.

Emotionally charged language often aligns with various discursive strategies, such as emotional appeals invoking fear, calls to action, and personal attacks. Importantly, these strategies are not only present in CTs but also manifest in other types of information disorder like online hate speech (Carvalho et al., 2024; Lee, 2022). Moreover, these aspects extend into the discourse held by populist political parties and their adherents (Wodak, 2020). This underscores the need for future research to explore how these components interplay and reinforce one another, thereby contributing to the escalation of social radicalization and polarization. Interestingly, while AI tools detected these strategies primarily in conspiracy articles, online readers also identified them in opinion articles. This implies that the subtleties underlying these persuasive tactics might be identified differently by humans and machines. Alternatively, it suggests that these strategies are not exclusive to misinformation or harmful narratives; they may also find legitimate use in opinion pieces.

Another significant discovery is that conspiracy articles are considered to be more opinionated than actual opinion pieces. This aspect aligns with prior studies that suggest conspiratorial content typically contains less factual information but a more pronounced presence of emotionally charged, threat-related details than nonconspiratorial content (Meuer, Oeberst, & Imhoff, 2023). Our research underscores this contrast, even when comparing conspiracy narratives with inherently subjective forms like opinion articles.

This perception might be attributed to factors like lack of concrete evidence and reliance on speculative connections, thereby eroding objectivity and grounding in facts.

Additionally, our research reveals that, unlike news and opinion articles, many conspiracy articles are viewed as lacking discursive coherence and consistency, which is in line with previous research findings (Miani, Hills, & Bangerter, 2022). Moreover, these narratives often feature vague assertions and citations from sources identified as unreliable, misleading, or biased, contrasting particularly with news articles. Specific elements of CTs, including references to malevolent and untrustworthy forces and the proposition of threatening truths, further amplify the perception of lacking credibility. In contrast, opinion articles inherently embrace subjectivity and personal perspectives. Readers anticipate a degree of subjectivity in opinion pieces, which prevents them from solely assessing these articles based on factual accuracy.

Practical Implications

Overall, the findings suggest that most credibility indicators described in literature and implemented in automated misinformation systems, such as InfoRadar, may be insufficient for distinguishing CTs from other types of content, especially opinion pieces. Our analysis emphasizes the importance of incorporating more granular information, such as emotions, which can aid in identifying not only misinformation, particularly CTs but also in differentiating facts from opinions—a task that proves challenging for both humans and machines. Integrating such information would lead to more accurate and transparent systems, better supporting users in critically assessing the content they encounter daily.

Additionally, our research also suggests that while ChatGPT has limitations in predicting CTs, it can still play a significant role in combating misinformation. This potential could be strengthened by integrating ChatGPT with existing explainable systems, such as InfoRadar, which also struggles with detecting CTs. In this combined approach, ChatGPT's responses could offer valuable insights into the categories assigned by InfoRadar and provide additional cues that either support or challenge InfoRadar's predictions about misinformation. This collaboration would help users make more informed judgments and promote critical thinking. Furthermore, developing collaborative frameworks that integrate AI combined with human-in-the-loop evaluation could yield innovative methodologies for combating misinformation.

Limitations and Future Research

Further research is needed to address the limitations of the current study, including the small participant sample and the focus on a single country and language. Future studies should expand the scope to compare human and AI performance across different languages, regions, tools, and contexts to improve generalizability. Additionally, a more diverse corpus, representing a wider range of media types and genres, should be incorporated to capture the multifaceted nature of misinformation. Investigating the accuracy of both machine and human assessments across various topics, particularly comparing objective subjects like science with ideologically driven topics like geopolitics, is also crucial. As the current AI models (ChatGPT 3.5 and InfoRadar) were trained only up until 2021, future research should explore the performance of updated models to assess improvements in detecting CTs. Ultimately, a

comprehensive investigation that combines both systems is crucial to fully evaluate their performance and influence on human assessment.

References

Athira, A. B., Kumar, S. M., & Chacko, A. M. (2023). A systematic survey on explainable AI applied to fake news detection. *Engineering Applications of Artificial Intelligence, 122,* 1–13. doi:10.1016/j.engappai.2023.106087

Brotherton, R., & Son, L. (2021). Metacognitive labeling of contentious claims: Facts, opinions, and conspiracy theories. *Frontiers in Psychology, 12*, 1–13. doi:10.3389/fpsyg.2021.644657

Butter, M., & Knight, P. (2020). *Routledge handbook of conspiracy theories*. New York, NY: Routledge.

Caled, D., Carvalho, P., & Silva, M. J. (2022). MINT-Mainstream and independent news text corpus. In V. Pinheiro, R. Amaro, C. Scarton. F. Batista, D. Silva, C. Magro, & H. Pinto (Eds.), *Proceedings of International Conference on Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022* (pp. 26–36). Lisbon, Portugal: Springer. doi:10.1007/978-3-030-98305-5_3

Caled, D., Carvalho, P., Sousa, F., & Silva, M. (2024). DOMAIN: Explainable credibility assessment tools for empowering online readers coping with misinformation. *ACM Transactions on the Web, 19*(1), 1–31. doi:10.1145/3696472

Campolong, K. (2022). "These cameras won't show the crowds": Intradiscursive intertextuality in Trumpian discourse's crowd size conspiracy theory. In M. Demata, V. Zorzi, & A. Zottola (Eds.), *Conspiracy theory discourses* (pp. 421–442). Amsterdam, The Netherlands: John Benjamins. doi:10.1075/dapsac.98.18cam

Caramancion, K. (2023). Harnessing the power of ChatGPT to decimate mis/disinformation: Using ChatGPT for fake news detection. In S. Chakrabarti, A. Sakib, S. Chattopadhyay, S. Poddar, A. Bhattacharya, & M. Gangopadhyaya (Eds.), *Proceedings of 2023 IEEE World AI IoT Congress* (pp. 42–46). New York, NY: IEEE. doi:10.1109/AIIoT58121.2023.10174450

Carvalho, P., Caled, D., Silva, C., Batista, F., & Ribeiro, R. (2024). The expression of hate speech against Afro-descendant, Roma, and LGBTQ+ communities in YouTube comments. *Journal of Language Aggression and Conflict, 12*(2), 171–206. doi:10.1075/jlac.00085.car

Chen, A., Chen, K., Zhang, J., Meng, J., & Shen, C. (2023). When national identity meets conspiracies: The contagion of national identity language in public engagement and discourse about Covid-19 conspiracy theories. *Journal of Computer-Mediated Communication, 28*(1), 1–12. doi:10.1093/jcmc/zmac034

Demata, M., Zorzi, V., & Zottola, A. (Eds.). (2022). *Conspiracy theory discourses*. Amsterdam, The Netherlands: John Benjamins. doi:10.1075/dapsac.98

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the ACL-HLT* (Vol. 1, pp. 4171–4186). Cambridge, MA: Association for Computational Linguistics. doi:10.18653/v1/N19-1423

Douglas, K., Uscinski, J., Sutton, R., Cichocka, A., Nefes, T., Ang, C., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology, 40*(1), 3–35. doi:10.1111/pops.12568

Fong, A., Roozenbeek, J., Goldwert, D., Rathje, S., & van der Linden, S. (2021). The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter. *Group Processes & Intergroup Relations, 24*(4), 606–623. doi:10.1177/1368430220987596

Giachanou, A., Ghanem, B., & Rosso, P. (2023). Detection of conspiracy propagators using psycho-linguistic characteristics. *Journal of Information Science, 49*(1), 3–17. doi:10.1177/0165551520985486

Giachanou, A., Rosso, P., & Crestani, F. (2019). Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 877–880). New York, NY: Association for Computing Machinery. doi:10.1145/3331184.3331285

Hameleers, M., & Van der Meer, T. G. (2021). The scientists have betrayed us! The effects of anti-science communication on negative perceptions toward the scientific community. *International Journal of Communication, 15*, 4709–4733.

Hoes, E., Altay, S., & Bermeo, J. (2023). *Leveraging ChatGPT for efficient fact-checking*. Retrieved from https://osf.io/preprints/psyarxiv/qnjkf

Imhoff, R., Zimmer, F., Klein, O., António, J. H., Babinska, M., Bangerter, A., . . . Van Prooijen, J. W. (2022). Conspiracy mentality and political orientation across 26 countries. *Nature Human Behaviour, 6*(3), 392–403. doi:10.1038/s41562-021-01258-7

Jolley, D., & Paterson, J. (2020). Pylons ablaze: Examining the role of 5G covid-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology, 59*(3), 628–640. doi:10.1111/bjso.12394

Korenčić, D., Chulvi, B., Casals, X. B., Toselli, A., Taulé, M., & Rosso, P. (2024). What distinguishes conspiracy from critical narratives? A computational analysis of oppositional discourse. *Expert Systems, 41*(11), 1–21. doi:10.1111/exsy.13671

Krippendorff, K. (2007). *Computing Krippendorff's Alpha Reliability'*. Retrieved from https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf

Lee, C. (2022). Covid-19 conspiracy theories as affective discourse. In M. Demata, V. Zorzi, & A. Zottola (Eds.), *Conspiracy theory discourses* (Vol. 98, pp. 215–238). Amsterdam, The Netherlands: John Benjamins. doi:10.1075/dapsac.98.10lee

Lischka, J. (2024). Credibility cues of conspiracy narratives: Exploring the belief-driven credibility evaluation of a YouTube conspiracy video. *Information, Communication & Society,* 1–20. Advance online publication. doi:10.1080/1369118X.2024.2334391

Luo, M., Hancock, J., & Markowitz, D. (2020). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research, 49*(2), 171–195. doi:10.1177/0093650220921321

Mahl, D., Schäfer, M., & Zeng, J. (2023). Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research. *New Media & Society, 25*(7), 1781–1801. doi:10.1177/1461444822107575

Marko, K. (2022). Extremist language in anti-Covid-19 conspiracy discourse on Facebook. *Critical Discourse Studies, 21*(1), 92–111. doi:10.1080/17405904.2022.2110134

Meuer, M., Oeberst, A., & Imhoff, R. (2023). How do conspiratorial explanations differ from non-conspiratorial explanations? A content analysis of real-world online articles. *European Journal of Social Psychology, 53*(2), 288–306. doi:10.1002/ejsp.2903

Miani, A., Hills, T., & Bangerter, A. (2022). Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances, 8*(43), 1–9. doi:10.1126/sciadv.abq3668

Mitchell, A., Gottfried, J., Barthel, M., Sumida, N., & Mitchell, A. (2018). *Distinguishing between factual and opinion statements in the news*. Pew Research Center. Retrieved from https://www.pewresearch.org/journalism/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news

Mitra, T., Counts, S., & Pennebaker, J. (2021). Understanding anti-vaccination attitudes in social media. *Proceedings of the International AAAI Conference on Web and Social Media, 10*(1), 269–278. doi:10.1609/icwsm.v10i1.14729

Molina, M., Sundar, S. S., Le, T., & Lee, D. (2021). "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist, 65*(2), 180–212. doi:10.1177/0002764219878224

Molyneux, L., & Coddington, M. (2020). Aggregation, clickbait and their effect on perceptions of journalistic credibility and quality. *Journalism Practice, 14*(4), 429–446. doi:10.1080/17512786.2019.1628658

Nekmat, E. (2020). Nudge effect of fact-check alerts: Source influence and media skepticism on sharing of news misinformation in social media. *Social Media + Society, 6*(1), 1–14. doi:10.1177/2056305119897322

OpenAI. (2023). *GPT-3.5 Turbo* [Large language model]. Retrieved from https://chat.openai.com/chat

Pustet, M., Steffen, E., & Mihaljevic, H. (2024). Detection of conspiracy theories beyond keyword bias in German-language Telegram using large language models. In Y.-L. Chung, Z. Talat, D. Nozza, F. M. Plaza-del-Arco, P. Röttger, A. M. Davani, & A. Calabrese (Eds.), *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)* (pp. 13–27). Cambridge, MA: Association for Computational Linguistics. doi:10.18653/v1/2024.woah-1.2

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In T. Fornaciari, E. Fitzpatrick, & J. Bachenko (Eds.), *Proceedings of the Second Workshop on Computational Approaches to Deception Detection* (pp. 7–17). Cambridge, MA: Association for Computational Linguistics. doi:10.18653/v1/W16-0802

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: Automatic detection of emerging Covid-19 conspiracy theories in social media and the news. *Journal of Computational Social Science, 3*(2), 279–317. doi:10.1007/s42001-020-00086-5

Tangherlini, T., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E., & Roychowdhury, V. (2020). An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. *PLoS One, 15*(6), 1–39. doi:10.1371/journal.pone.0233879

Tindale, C. (2007). *Fallacies and argument appraisal*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511806544

Uscinski, J. (2018). The study of conspiracy theories. *Argumenta, 3*(2), 233–245. doi:10.23811/53.arg2017.usc

Uscinski, J., Enders, A., Klofstad, C., Seelig, M., Funchion, J., Everett, C., . . . Murthi, M. (2020). Why do people believe Covid-19 conspiracy theories? *Harvard Kennedy School Misinformation Review, 1*(3), 1–12. doi:10.37016/mr-2020-015

van Prooijen, J., Cohen Rodrigues, T., Bunzel, C., Georgescu, O., Komáromy, D., & Krouwel, A. (2022). Populist gullibility: Conspiracy theories, news credibility, bullshit receptivity, and paranormal belief. *Political Psychology, 43*(6), 1061–1079. doi:10.1111/pops.12802

van Prooijen, J., & Douglas, K. (2018). Belief in conspiracy theories: Basic principles of an emerging research domain. *European Journal of Social Psychology, 48*(7), 897–908. doi:10.1002/ejsp.2530

Vears, D., & Gillam, L. (2022). Inductive content analysis: A guide for beginning qualitative researchers. *Focus on Health Professional Education: A Multi-disciplinary Journal, 23*(1), 111–127. doi:10.11157/fohpe.v23i1.544

VERBI Software. (2021). MAXQDA 2022 [Computer software]. Berlin, Germany: VERBI Software. Retrieved from maxqda.com

Viviani, M., & Pasi, G. (2017). Credibility in social media: Opinions, news, and health information: A survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 7*(5), 1–25. doi:10.1002/widm.1209

Walter, N., & Salovich, N. (2021). Unchecked vs. uncheckable: How opinion-based claims can impede corrections of misinformation. *Mass Communication and Society, 24*(4), 500–526. doi:10.1080/15205436.2020.1864406

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27, pp. 1–107). Strasbourg, France: Council of Europe.

Wodak, R. (2020). *The politics of fear: The shameless normalization of far-right discourse* (2nd rev. ed.). London, UK: Sage Publications.