Funded by the European Union

UK Research and Innovation

Deliverable number: D3.2

Technical report on the state of the art of political disinformation in social networks and the press

An assessment of the definitions, risks, models, datasets and challenges of political disinformation detection in social networks

Version 1.

Project Details

Project Acronym:	HYBRIDS
Project Title:	Hybrid Intelligence to monitor, promote and analyse transformations in good democracy practices
Grant Number:	101073351
Call	HORIZON-MSCA-2021-DN-01
Topic:	HORIZON-MSCA-2021-DN-01-01
Type of Action:	HORIZON-TMA-MSCA-DN
Project website:	https://hybridsproject.eu/
Coordinator	Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)-Universidade de Santiago de Compostela (USC)
Main scientific representative:	Prof. Pablo Gamallo Otero, pablo.gamallo@usc.es
E-mail:	citius.kmt@usc.es, info@hybridsproject.eu
Phone:	+34 881 816 414

Deliverables Details

Number:	D3.2
Title:	Technical report on the state of the art of political disinformation in social networks and the press
Work Package	WP3: Democracy Threats
Lead beneficiary:	UNICAEN
Deliverable nature:	R- Document, report
Dissemination level:	PU-Public
Due Date (month):	30/04/2024 (M16)
Submission Date (month):	30/04/2024 (M16)
Keywords:	Disinformation, hate speech, deepfakes, hyper-partisanship, automated factchecking, deep learning, natural language processing, artificial intelligence

Abstract

This technical report presents an overview of the threats posed by political disinformation strategies in social networks and the press and it also discusses the efforts to deal with them from the computer science perspective. Topics like hate speech, deepfakes, fact-checking and disinformation made with generative models are assessed. We present the strengths and limitations of state of the art detection models and existing datasets, highlighting possible research directions.

Deliverable Contributors

	Name	Institution	E-mail
Contributing Authors	Søren Fomsgaard	UNICAEN	soren.fomsgaard@unicaen.fr
	Michele Joshua Maggini	CiTIUS-USC	michele.cafagna@um.edu.mt
	Rabiraj Bandyopadhyay	GESIS	rabiraj.Bandyopadhyay@gesis.org
	Rafael Martins Frade	NEWTRAL	rafael.martins@newtral.es
	Paloma Piot Pérez-Abadín	UDC	paloma.piot@udc.es
	Rrubaa Panchendrarajan	QMUL	r.panchendrarajan@qmul.ac.uk
Reviewers	Gaël Dias	UNICAEN	gael.dias@unicaen.fr
	Pablo Gamallo Otero	CiTIUS, USC	pablo.gamallo@usc.es

History of Changes

Version	Date	Changes to previous version	Status
0.1	15/04/2024	First Draft	Draft
0.2	22/04/2024	Consortium Internal review	Review
1.0	30/04/2024	Approved version to be submitted	Final

List of Acronyms

AMG	Arbitraty Misinformation Generation
AIGC	Artificial Intelligence Generated Content
CNG	Controllable Misinformation Generation
GAN	Generative Adversarial Networks
GenAI	Generative Artificial Intelligence
HG	Hallucination Generation
ML	Machine Learning
NLP	Natural Language Processing
VAE	Variational Autoencoders

Contents

1 Introduction 6

2 Definitions 7
- 2.1 Disinformation . . . 7
- 2.2 Hate Speech . . . 8
- 2.3 Implicit Hate Speech . . . 8
- 2.4 Deepfakes . . . 9
- 2.5 Hyperpartisan News . . . 9

3 Real and Potential Threats 11
- 3.1 Disinformation . . . 11
- 3.2 Hate Speech in Political Discourse . . . 12
- 3.3 Implicit Hate Speech . . . 13
- 3.4 Deepfakes . . . 14
- 3.5 Hyperpartisan News . . . 14
- 3.6 Fact-checking for Combating Political Disinformation . . . 15

4 Models and Datasets 17
- 4.1 Hate Speech in Political Discourse . . . 17
- 4.2 Implicit Hate Speech . . . 18
- 4.3 Deepfakes . . . 19
  - 4.3.1 Audio . . . 19
  - 4.3.2 Image and video . . . 20
- 4.4 Hyperpartisan News . . . 21
- 4.5 Automated Fact-checking . . . 22

5 Conclusions 24

1 Introduction

In today's digital age, politics and the online world are deeply intertwined. This report aims to shed light on how false information spreads across social networks and traditional media, particularly in the realm of politics.

Political disinformation involves spreading false or misleading information to influence public opinion or cause confusion. It's a significant challenge because it clouds people's understanding of what's true and what's not.

A key aspect of this problem is hate speech, where individuals use hurtful language to target others based on their identity or beliefs. Hate speech can fuel division and harm in society, making it a critical issue to address.

Another concern is the rise of deep fake news, where manipulated videos or images portray individuals saying or doing things they never did. This deceptive practice undermines trust and complicates efforts to verify information online.

Additionally, hyperpartisan news exacerbates the spread of disinformation by presenting biased viewpoints that cater to specific political leanings. This further polarizes society and contributes to the dissemination of false information.

Despite these challenges, fact-checking serves as a vital tool in combating disinformation by verifying the accuracy of claims. However, effective fact-checking requires reliable data and robust methodologies to navigate the complexities of online misinformation.

Through this report, we aim to explore these issues in depth and identify strategies to address the spread of political disinformation. By fostering greater awareness and understanding, we can work towards promoting a more informed and resilient society in the digital age.

2 Definitions

The following section will introduce disinformation and some of its key concepts in the political domain. Treating the definitional issues related to various forms of current state of the art political disinformation will aid in grounding the analysis of the threats they pose in section 3.6. The key areas of impact of disinformation that we identify are: deepfakes, hate speech and hyperpartisanism.

2.1 Disinformation

There are no universally agreed-upon scientific definitions or methods for applying the concept of dis-, mis- and malinformation in practice, there are (many) common sense definitions [Baines and Elliott, 2020]. A general definition of disinformation is information simply "intentionally misleading information" [Fallis, 2014]. Broadly speaking, disinformation can be distinguished from misinformation and malinformation given that misinformation is false information (in the sense of being incorrect), whereas malinformation is true or correct information that is spread in an incorrect context and thus constitute "reconfigurations of the truth" [Baines and Elliott, 2020]. This means that disinformation $\subseteq$ misinformation[1]. Next to the differences that tend to arise in the conceptual modeling of disinformation between disciplines, the other main challenge with defining disinformation and its related concepts is that it is difficult to devise such definitions and taxonomies that are conceptually sound and yet parsimonious, and at the same time scientifically operationalizable and valid [Baines and Elliott, 2020].

Disinformation can be approached from a variety of fields, including (at least) philosophy, cognitive and social psychology, communications studies, linguistics, political science, sociology, and computer science and information theory. In addition to differences between common sense and technical definitions of information and disinformation [Baines and Elliott, 2020], each field also tends to its own understanding of the concept of information [Søe, 2014] This plurality of perspectives is necessary because disinformation is a complex phenomenon. This is because analyzing disinformation content and the nature of its production, dissemination, and their negative impacts both at the level of the individual and the level of society requires cannot be answered adequately from within any single discipline. Within the scope of this report, we restrict ourselves primarily to the computational, communicative and linguistic perspectives on the state of the art of political disinformation.

Due to the disciplinary wealth of perspectives, and the diversity of the types of disinformation (more on this in a moment), it is notoriously difficult to define disinformation in such a way as to cover all the contexts in which it appears. For the purpose of defining the state of the art of political disinformation, however, we can modify a recent comprehensive taxonomy given by [Chen and Shu, 2024]. Keeping in mind that disinformation is intentionally misleading information, common types of disinformation include false news[2], deepfakes, rumors, conspiracy theories, clickbait, misleading claims, cherry-picking in the political domain[3]. Adding to this we can distinguish

[1] In a strict sense, misinformation also includes satire.

[2] As has been pointed out, 'fake news' has itself become a polarized and contested term [Molina et al., 2021; Vosoughi et al., 2018]. For this reason we use the phenomenon as 'false news' instead.

[3] Most of these types apply to other domains such as health, science, finance, etc. (cf. [Chen and Shu, 2024]).

between these types based on the dissemination (one-to-one or one-to-many [Buchanan et al., 2021], modality (textual, auditive, visual, multimodal), genre (social media post, news/scientific article, etc.) and communicative and epistemic errors[4] committed that each type can instantiate.

2.2 Hate Speech

There is no formal definition of **hate speech**, but previous works [Davidson et al., 2017; Founta et al., 2018; ElSherief et al., 2018a,b; Mathew et al., 2021; Silva et al., 2021; Das et al., 2023] deepened on this topic, defining it as "language characterized by offensive, derogatory, humiliating, or insulting discourse [Founta et al., 2018] that promotes violence, discrimination, or hostility towards individuals or groups [Davidson et al., 2017] based on attributes such as race, religion, ethnicity, or gender [ElSherief et al., 2018a,b; Das et al., 2023]". Under this definition, which aligns very well with the United Nations one [Nations, 2023], we frame our hate speech definition by differentiating hate speech from non-hate and offensive speech.

When defining hate speech in the political realm, we can build upon the previously proposed definition by emphasizing its political nature and the resulting impacts or consequences. Then, based on the provided definition of hate speech, we can extend this to define **political hate speech** as "language that includes offensive, derogatory, humiliating, or insulting discourse specifically targeting individuals or groups based on their political beliefs, affiliations, or ideologies". Political hate speech aims to promote hostility, discrimination, or violence against individuals or groups due to their political views, involvement in political activities, or association with particular political parties or figures. This type of speech can contribute to polarization, division, and conflict within political discourse and society at large.

2.3 Implicit Hate Speech

As pointed out in 2.2, hate speech can be defined as offensive, derogatory and humiliating discourse that is used to incite violence and discrimination against people belonging to a particular race, religion, ethnicity, or gender. In this section we will provide definition of another type of hate speech called **implicit hate speech** which are notoriously difficult for models to detect.

Implicit hate speech as defined in [ElSherief et al., 2021] is "coded or indirect language that disparages a person or group on the basis of protected characteristics like race, gender or cultural identity". Implicit hate speech is difficult to detect because of the lack of keywords that can help the model identify implicit texts unlike more overt forms of hate-speech [Waseem et al., 2017; Wiegand et al., 2019] and hence classifiers trained on more overt forms of hate speech don't perform well when presented with the task of detecting implicit hate speech [Caselli et al., 2020]. Since this kind of speech is difficult to detect, extremist groups can use this language to incite violence while denying accountability [Dénigot and Burnett, 2020], and hence efforts are required to detect implicit hate speech, from curating of datasets to creating robust models.

[4] E.g., violations of Gricean Maxims on the one hand and epistemic errors like total fabrication, being unverifiable, vagueness, or ambiguity on the other.

2.4 Deepfakes

The term “deepfake” is a popular term, derived from ‘deep learning’ and ‘fake’ (media) content [Shoaib et al., 2023; Danry et al., 2022]. It refers to a broad category of deceptive and inauthentic digital content that has been deliberately crafted to manipulate its audiences across modalities, typically (though not exclusively) as part of larger disinformation campaigns on online social media platforms in political contexts.

2.5 Hyperpartisan News

When addressing hyperpartisanship, we should differentiate between the vastness of political and computer science literature. In the former, hyperpartisanship is described in its intrinsic nature, considering its political traits and societal implications. The latter case implies simplifying the concept to fit the necessity of the detection task with Machine Learning or Deep Learning methods. For instance, only some aspects of the issue are considered, i.e.: content style, and source, while other traits are not evaluated when attempting the classification. That could implicate limiting the process of tackling hyperpartisan news by misunderstanding or misinterpreting the phenomenon. Indeed, hyperpartisan news detection has been categorized under the fake news domain [Dumitru and Rebedea, 2021], although a hyperpartisan article does not always coincide with fabricated and untruthful facts. Another misconception consists of considering an article as extremely polarized only because the newspaper or the journalist has a clear political preference. In other words, despite its political leaning or ideology, left or right-wing journals will not always publish only hyperpartisan articles. Unfortunately, in some articles [Azizov et al., 2023; Ko et al., 2023], the assignation of the hyperpartisan golden label is based on the news outlets’ political leaning inferred by experts, like the case of Allsides[5]. Thus, giving a comprehensive and exhaustive definition to hyperpartisanship is a difficult challenge, especially considering that sometimes this concept overlaps with similar ones, at least in Computer Science.

Given the notions of ethos, logos and pathos, Weiai Wayne Xu and Kim [2020] found three factors determining the odds of sharing news: transparency, content style and moral framing. Hyperpartisanship involves each of them and at different levels: the media propagating the news outlet that usually defines itself as alternative [Ernesto de León and Adam, 2024; Kristoffer Holt and Frischlich, 2019], the political agenda of the news publisher and its coverage bias [Lyu et al., 2023; Gangula et al., 2019; Yang et al., 2016], the stylistic traits that shape the content, the propagandistic aim displaying itself through specific biases [Maggini et al., 2024; Gangula et al., 2019], and the audience it is addressing [Yang et al., 2016]. Thus, different factors contribute to making an article hyperpartisan.

Authors like Garg and Sharma [2022]; Dumitru and Rebedea [2021] highlight the importance of the intentionality of news to be misleading or distorting the facts as a feature of hyperpartisanship. Furthermore, the presence of the intent could result in considering hyperpartisanship in either the misinformation or disinformation category. The primary distinction between these two close fields is the intentionality with which an article conveys doubtful truth and propagandistic content. The problem of investigating the vicious intent is related to the perception of news as

[5] https://www.allsides.com/unbiased-balanced-news

spreading it. This fact is subjective and depends on the readers' political orientation and level of conflictuality with the content.

Currently, researchers employ various methodologies, ranging from traditional machine learning (manually crafted features, linguistic patterns, sentiment analysis, rule-based systems) to deep learning methods (Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory), Transformers and hybrid approaches. To streamline the detection process, working with few labels or binary classification would be the best trade-off. That approach will avoid resource consumption and score better results [Aksenov et al., 2021]. Interestingly, hyperpartisan news employs the same formal textual structure as standard news [Kiesel et al., 2019], but emphasizes a one-sided political agenda [Lyu et al., 2023; Barnidge and Peacock, 2019], often introducing anti-system narratives [Barnidge and Peacock, 2019], thus labeled as *alternative*, highlighting the contrast with mainstream sources and resulting in a division counterposing two contrasting subjects, *us* against *them*, with a different perception of what is the truth. To underscore this dichotomy, SemEval 2019 Task 4 introduced two classification labels: mainstream and hyperpartisan. Due to these traits, hyperpartisanship has been considered to be close to conspiracy theories [Ernesto de León and Adam, 2024].

3 Real and Potential Threats

As it stands, the potential risks and benefits of frontier AI, meaning cutting-edge developments in the research and innovation of artificial intelligence applications, are many and are only expected to increase. Frontier AI includes current advanced algorithms and large models, and generative AI (GenAI) [Shoaib et al., 2023]. As such, frontier AI can be considered the state of the art in terms of political disinformation. In the context of the production and dissemination of disinformation in social and legacy media alike, the main threat of frontier AI in the domain of political disinformation, at the time of this writing, is to generative AI and its applications in the visual, textual, auditive and hybrid modalities. From a birds-eye view, the largest threat of GenAI is its potential to undermine trust in its many forms given that “without trust, conspiracy theories flourish, scientific consensus is questioned, and social polarization deepens” [Shoaib et al., 2023].

One approach to conceptualizing and concretizing the threats implied by the potential for abuse in technology in general is threat modeling, a methodology with roots in informational security (infosec). Recently, authors working on mapping and pre-empting the potential threats of frontier AI have used threat modeling as a way to analyze the threats in the specific case of generative AI and its potential application in producing disinformation content in the textual domain [Crothers et al., 2023; Buchanan et al., 2021]. The threat model proposed by Crothers et al. [2023] covers four main types in the textual modality, of which three are relevant in the political domain: exploitation of AI Authorship (including article or opinion piece submission at scale),online Influence Campaigns (including propaganda, astroturfing and information warfare), and spam & harassment (direct messages and comment submissions at scale) [Crothers et al., 2023, p.9].

3.1 Disinformation

The potential social impact of disinformation campaigns can be analyzed through methods such as threat modeling. Assessing the real social impacts of disinformation is a complex task. There are two reasons for this. First, assessing the impact of a disinformation campaign requires answering open questions that are difficult to quantify. For example, in disinformation detection research the most fundamental question is precisely: “what makes disinformation effective?” [Buchanan et al., 2021]. Evaluating the effectiveness of a piece of disinformation used in a larger campaign (e.g., a deceptive social media post about fabricated events) requires analyzing its psychological, social, linguistic impacts and conditions for success at the level of the individual, and its political and social impacts at the level of the collective (communities, society at large). Second, in answering questions such as these, researchers depend on incomplete information about the identity and means of threat actors. One reason for this is that the situation, in which bad actors produce disinformation ’one step ahead’ of the researchers and journalists who work on detecting it, is fundamentally adversarial. One example of this is the prevalence of dark numbers of inauthentic accounts, such as trolls and bots, that are active on social media platforms. Similarly, as a form of manipulation and deception (see 2.5, disinformation works in epistemically asymmetrical situations where the receiver, audience, or target is unaware of its character.

A common answer to the fundamental question of what makes disinformation effective from the perspectives of psychology and policy, is that it tends to confirm pre-existing views and reinforce socio-political division which already exist [Buchanan et al., 2021]. The more complex questions of how disinformation is effective is best answered by focusing on specific disinformation campaigns.

Disinformation in the political domain poses general threats, both directly and indirectly. It can directly impact and endanger political candidates and interfere with electoral campaigns via narrative manipulation and astroturfing, and it indirectly harms societies at large through the erosion of public trust [Fallis, 2014; Shoaib et al., 2023; Pawelec, 2022], i.e., that the undermining of general trust in online information and its sources among the public as well as expert sources. It is expected that a general decrease in the trust of content, platform, and users of our information environments might lead to a gradual destabilization of electoral campaigning and reporting, and a larger degree of partisan polarization[6] in the political domain as a consequence of this. It is worth mentioning that the proliferation of disinformation poses a potential threat at the collective, societal level outside the political, such as threatening public health by promoting skepsis of vaccines or expert advice in the health domain, undermining trust in emergency-response efforts [Graham and Bogle, 2022], and financial and commercial manipulation via false promotion of products and services.

As is the case with other forms of frontier AI, GenAI poses a state-of-the-art threat in the context of political disinformation. Although several types of generative model architectures exist, the current state-of-the-art of GenAI in many applications is currently held by generative adversarial networks Goodfellow et al. [2014] produce text, audio and video, as the name suggests, after being trained in an adversarial learning regime whereby a discriminator and generator iteratively improve one another. As such, generative models are inherently adversarial, and this fact has led researchers to describing the current situation where generative models are becoming increasingly democratized, and available to threat actors, as a ‘cat-and-mouse game’ with the odds in favor of those that seek to generative disinformation content in various contexts online and the researchers that seek to detect it [Bontcheva et al., 2024; Shoaib et al., 2023].

In the textual domain, recent work has explored various strategies that could be employed by threat actors to generate disinformation content in a political context. Buchanan et al. [2021] show that OpenAI’s ChatGPT-3 model can be prompted to generate text that expresses climate denialist sentiments, politically confrontational (false) headlines, (hyper)partisan headlines and articles, and conspiratorial tweets, which could be used in disinformation campaigns to carry out narrative reiteration, elaboration, manipulation, seeding and wedging. Chen and Shu [2024] show that ChatGPT-3-5, Llama2-7b, Vicuna-7b can be used to generated misinformation (uintentionally) through hallucinated generation (HG), and (intentionally) through different forms of arbitrary (AMG) and controllable (CMG) misinformation generation.

3.2 Hate Speech in Political Discourse

In recent years, the use of hateful and divisive language in political discourse has become a pressing issue with far-reaching consequences. Hate speech, particularly within the realm of politics, poses significant threats to democratic values, social cohesion, and individual well-being.

[6] Sometimes referred to as ’wedging’ or ’deepening’ of socio-political divides [Buchanan et al., 2021].

Political discourse, which ideally should foster healthy debate and exchange of ideas, is increasingly damaged by hate speech. Hate speech refers to language that seeks to attack, intimidate, or incite violence against a particular group based on characteristics such as race, ethnicity, religion, gender, or political beliefs. When employed within politics, hate speech can undermine the foundations of democracy by polarizing societies, eroding trust in institutions, and diminishing respect for diverse viewpoints.

Several works have studied the presence of hate speech in political discourse. Agarwal et al. [2021] have analysed 2.5 million tweets to identify hate speech against members of the parliament and they characterised hate across multiple dimensions of time, topics and members' demographics. Other works acknowledge the presence of hate speech in the political discourse in social media. Solovev and Pröllochs [2022] analyzed how the amount of hate speech in replies to posts from politicians on Twitter depends on personal characteristics, such as their party affiliation, gender, and ethnicity.

The phenomenon of hate speech in political discourse is not confined to specific regions but is a global concern affecting societies worldwide. From the United States [Grimminger and Klinger, 2021] to Europe [Grapă and Mogoș, 2023], India [Masud and Charaborty, 2023], and beyond, instances of inflammatory rhetoric and discriminatory language in political settings have been observed. Each region may have its unique socio-political dynamics and cultural contexts influencing the nature and targets of hate speech. However, the underlying impact remains consistent: polarization, division, and the erosion of trust within communities. Recognizing the global scope of this issue underscores the importance of developing comprehensive strategies and leveraging technologies to address hate speech effectively on an international scale.

3.3 Implicit Hate Speech

As defined in 2.3, we saw that implicit hate is defined as a coded language that is used to incite violence against groups belonging to race, religion, ethnicity or gender. These type of language can be very hard to detect by a model because of the absence of linguistic signals that can aid the model in helping to classify the corresponding text. Hartvigsen et al. [2022]; Jurgens et al. [2019]; ElSherief et al. [2021] have started focusing on other constructs of hate speech like sarcasm, circumlocution which requires the models to make sense of the text rather than focusing on tokens related to overt hate speech [Ocampo et al., 2023a; Waseem et al., 2017]. Attempts have been made to detect implicit hate speech in both curation of datasets and development of models. ElSherief et al. [2021]; Sap et al. [2020] released 2 datasets (LatentHatred and SBIC) that aid in detecting implicit hate speech. Hartvigsen et al. [2022] also released a dataset named ToxiGen which was machine generated using GPT-3 model [Brown et al., 2020]. With the rise of LLMs like GPT-4 OpenAI et al. [2024] and open source alternatives like Llama-2 [Touvron et al., 2023] and Gemma [Google, 2024], attempts have been made to detect implicit hate speech using both open-source and closed source models. Huang et al. [2023] proposed a prompting framework inspired by chain-of-thought [Wei et al., 2023] prompting method to tune large language models to generate explanations for the classification of a text to hateful or non-hateful. Kim et al. [2022]; Ocampo et al. [2023b] have also tried to build robust classifiers using contrastive learning and adversarial examples respectively. But the there seems to be gaps in incorporating quality metrics as well as mitigation of biases that exist in hate speech classification systems [Santurkar et al., 2023]. We will elaborate about models and datasets in Section 4.2.

3.4 Deepfakes

As mentioned previously in section 2.5, deepfakes concern various forms of AI-generated content (AIGC) deception [Shoaib et al., 2023]. Insights from recent work in deception detection research indicate that humans have roughly the same amount of difficulty (and corresponding poor success rate) with detecting deceptive content across modalities (text, audio, video) [Hancock and Bailenson, 2021]. The main threat of deepfakes has been described as 'the epistemic threat that people can be lead to acquire false beliefs' [Fallis, 2021; Hancock and Bailenson, 2021], which in turn has detrimental down-stream effects on both on individuals and societies.

The state of urgency in which we find ourselves at present has cannot be understated. Extending on the notion of the "infodemic" [Baines and Elliott, 2020] during the spread of health d/misinformation during the COVID-19 pandemic, Fallis [2021] have raised worries that the general epistemic threat of deepfakes is setting us on a path towards an outright "infopocalypse".

3.5 Hyperpartisan News

Algorithms permeate various aspects of our society, exerting differing degrees of influence. Within political processes, they serve as a regulatory force, controlling the flow and consumption of information [Wagner et al., 2021]. The proliferation of hyperpartisan news in recent years has raised significant concerns regarding political stability. The emergence of alternative media platforms exacerbates risks to democracy [McCoy and Somer, 2019], as they often propagate divisive content [Kristoffer Holt and Frischlich, 2019]. Addressing this issue, along with related phenomena such as misinformation and disinformation, requires leveraging machine learning and deep learning methods due to their effectiveness and scalability.

Hyperpartisan news outlets frequently deviate from the objectivity and factual standards expected of professional journalism. Instead, they report events with a pronounced bias towards specific ideologies and political parties. These biases, often compounded with other rhetorical techniques, serve to disseminate propagandistic content, which can mislead and manipulate audiences.

As individuals and political parties strive to undermine their adversaries and secure an edge, hyperpartisanship often fuels the dissemination of misinformation and propaganda. This trend ultimately erodes trust in the political process and exacerbates political divisions within society.

Since news is a form of content, it carries a narrative that significantly shapes one's social identity. Social identity theory posits an inherent division, both internal and external, among members of various groups, often favoring those who are similar over those who are different [Novoa et al., 2023]. Given that different coalitions are grounded in distinct beliefs, we can view hyperpartisanship as a form of epistemological polarization. The digital landscape, particularly with the advent of social media [Barnidge and Peacock, 2019], has facilitated the proliferation of hyperpartisan news, thereby exacerbating polarization among audiences. This proliferation has led to the emergence of echo chambers, enclosed epistemic communities where individuals reinforce shared perceptions of reality, typically aligned with specific political ideologies or issues. Within these echo chambers, exposure to differing viewpoints is limited, as communication tends to reinforce preexisting assumptions [Ross Arguedas et al., 2022], even in the face of contradictory evidence. Consequently, those who uphold the dominant narrative within these homogenous social clusters gain influence. Such insulated social environments perpetuate polarization through a feedback loop that reinforces entrenched opinions [Hobolt et al., 2023]

When this phenomenon permeates the structure of the government, it has the potential to cause its collapse, as political parties become unable or unwilling to compromise.

3.6 Fact-checking for Combating Political Disinformation

As we already discussed, political disinformation is a growing concern with a greater impact on society [Allcott and Gentzkow, 2017]. At the same, several fact-checking organizations and platforms were introduced over the past decades to fight against political disinformation. Several studies reveal the impact of fact-checking on individuals in identifying the truthfulness of political statements, and the corresponding impacts during political movements [York et al., 2020; Wintersieck et al., 2021]. While fact-checking plays a key role in restoring the credibility of information and social trust, the task remains challenging due to the wide range of threats we already discussed. In this section, we discuss the role of fact-checking in combating disinformation in the political domain.

Political claim extraction: analysing the truthfulness of political discourse begins with extracting claims made by political entities that have the potential to mislead the public. Various research focuses on this direction by extracting check-worthy claims and prioritizing them for the verification process. This includes claim extraction from political debates, transcripts of proceedings, social media posts about political movements, manifestos, and newspaper reports [Hassan et al., 2017; Blokker et al., 2020; Beltrán et al., 2021; Patwari et al., 2017].

Political evidence gathering: followed by claim extraction, gathering accurate evidence plays a key role in determining the veracity of political claims. Various sources have been used in the literature for extracting political evidence for fact-checking. For example, Long et al. [2017] uses the speaker's credit history and metadata of the speaker as evidence. However, this information may not be always available, hence several works utilize external resources such as Wikipedia and Google search [Yasser et al., 2018] for evidence gathering. Further, integrating multiple sources such as external reports with the credit history, and metadata of the speaker [Karimi et al., 2018] and utilizing the stance towards the speaker's statements [Hassan and Lee, 2020] have shown to be serving as powerful evidence.

Political false news detection: given political claims and evidence related to them, the false news detection task identifies the veracity of the claims. This may vary from a binary classification (true or false news) to a multi-class classification. The challenge escalates when false news is mixed with true information, for example, a false claim referring to a true event. This demands identifying the degrees of truthfulness (e.g. half true) in false news detection. For example, The

PolitiFact[7] fact-checks political claims and labels the accuracy of the claims into six levels, *True, Mostly True, Half True, Mostly False, and Pants on Fire*.

Political rumour detection: rumour is defined with different definitions in the literature. According to the Oxford Dictionary, rumour is a piece of information passed from one person to another which may or may not be true [Patel et al., 2022]. While various works treat rumour as false information, the fact-checking domain characterizes a rumour as an unverified claim at the time of reporting. Detecting rumour is often modeled as a binary classification indicating whether a statement is a rumour or non-rumour. Subsequently, identified rumours can undergo further scrutiny for veracity classification, determining whether a rumour is true or false or unverified [Zubiaga et al., 2018]. Social platforms serve as a major source of spreading political rumours. The content circulated on social platforms is often noisy and less semantic and spreads quickly through evolving networks. Such rapid propagation of disinformation leads to rumours easily circulated on social platforms. Existing research on rumour detection intends to analyze the stream of social media posts or analyze posts from selected user accounts [Hussaini et al., 2022] for identifying rumours. Further, the multimodal information present in social media posts such as user details, hashtags, and URLs mentioned are also used as sources of information for accurate rumour detection [Tam et al., 2019]. In addition to the analysis of content posted in social platforms, the characteristics of the social graphs that depict the information spread can also be analyzed to identify patterns of spread of rumours [Nguyen et al., 2024, 2023].

Real-time monitoring: verifying the truthfulness of political claims requires timely identification of the veracity of the claims. Various researchers have attempted to develop real-time monitoring systems to cater to this requirement. For example, ClaimBuster [Hassan et al., 2017] verified the real-time fact-checking pipeline developed by performing case studies on live political debate broadcasts on television and by monitoring real-time social media posts. Similarly, ClaimHunter developed by Beltrán et al. [2021] monitors social media posts of selected accounts and retrieves them in real time for the verification process.

Feedback-loop: the development of a fact-checking pipeline in political discourse requires human-annotated data and feedback from journalists to build accurate models. This is often done by annotating the transcripts of political debates and social media posts with the help of journalists [Hassan et al., 2017]. Further, semi-automated annotation with human intervention [Blokker et al., 2020] has also been explored in the literature for creating training data for political fact-checking. Apart from the annotation tasks, research studies have attempted hybrid solutions with humans in the loop to improve models' performance. For example, ClaimHunter developed by Beltrán et al. [2021] gathers real-time claims, notifies journalists through Slack channels, and obtains their feedback for improving automated fact-checking performance.

[7] http://www.politifact.com/

4 Models and Datasets

4.1 Hate Speech in Political Discourse

In the realm of addressing hate speech in political discourse, the availability of datasets and the development of machine learning models play a pivotal role in analysis and mitigation efforts. This chapter explores the landscape of hate speech datasets and the utilization of machine learning models to identify and combat this phenomenon. By examining existing resources and technological advancements, we aim to uncover how data-driven approaches can contribute to a deeper understanding of hate speech dynamics and inform effective strategies for promoting respectful political dialogue, but also highlighting the potential issues with the current datasets and models.

Despite the progress made in hate speech in the political discourse using datasets and models, challenges persist in ensuring algorithmic fairness, addressing linguistic nuances, and adapting to evolving forms of hate speech in this field.

Several datasets have been developed to identify hate speech, particularly within the realm of politics. One notable dataset focuses on the 2020 US elections [Grimminger and Klinger, 2021]. Grimminger and Klinger specifically targeted incidents of hatred in the political sphere during this time, gathering data from Twitter using keywords related to presidential candidates, campaign slogans, voter affiliations indicated by hashtags, and even politicians' nicknames. They meticulously annotated this dataset using binary classification.

Other projects have sought to encompass hate speech across various topics, including politics [Toraman et al., 2022]. Toraman et al. compiled a dataset by utilizing a broad range of keywords and hashtags spanning different areas such as religion, gender, racism, politics, and sports. For political content, they collected data using keywords like democratic party, republican party, government, white house, president, Trump, Biden, or minister. Their dataset identified 1610 instances of hate speech in English tweets related to politics and 7657 instances in Turkish messages.

Another study by Rezvan et al. focused on capturing various types of hate speech in the political domain [Rezvan et al., 2018]. They annotated a Twitter corpus of $25\,000$ entries, categorizing content into types of harassment: sexual, racial, appearance-related, intellectual, and political. Political harassment, as defined by them, involves discussions on political views regarding issues influenced by the government, such as global warming, the opioid epidemic, immigration, or gun control. Politicians and politically active individuals are typical targets. They created this dataset using a lexicon that included terms associated with different types of hate speech, including words like cockmuncher, towel head, dickwad, propaganda, and demon.

Other authors decided to focus on the impact of politically biased data on hate speech classification in German language [Wich et al., 2020]. Therefore, they constructed three politically biased datasets (left-wing, right-wing, politically neutral) and compared the performance of classifiers trained on them. They used an existing Twitter hate speech corpus with binary labels (offensive, non-offensive), extracted the offensive records, and combined them with three data sets each (politically left-wing, politically right-wing, politically neutral) implicitly labeled as non-offensive. These three new created datasets were built by using keywords from different topics that were generated from the existing dataset.

In general, the methods used to create datasets in this area heavily rely on using keywords. However, this approach can lead to biases in the data. For instance, if these keywords focus too much on certain individuals or politicians, or if they include offensive language (which might not always be indicative of hate speech), it could skew the dataset. This approach might also overlook other ways hate speech can be expressed in this context. Some suggestions to gather diverse data from social media without bias, are adopting a random sampling approach across platforms like Twitter, Facebook, and Reddit. Moreover, filtering sampled data using automated methods for relevance, and then engage human annotators from diverse backgrounds to manually label potential hate speech. In addition, iterating on sampling and annotation processes to enhance the dataset quality and representativeness, can foster inclusivity in hate speech detection model development.

There are several notable issues with existing datasets for hate speech detection in the political domain. Firstly, many datasets lack sufficient context, which can lead to misinterpretations of the data. Additionally, there is a scarcity of multilingual data, although some efforts have been made for German language datasets. Another significant gap is the absence of a universally agreed-upon definition of hate speech, which hinders efforts to effectively address this problem.

Different authors have also run experiments and baselines on datasets of hate speech in political discourse. Grimminger and Klinger [2021] experimented with their curated dataset and achieved a F1-score of 0.53 for the detection of hate speech posts. Other efforts trained different models on the multiclass task of hate speech detection (hate, offensive, normal), and achieved a F1-score of 0.83 for English data using Megraton model (Megatron introduces an efficient parallel training approach for BERT-like models to increase parameter size) [Toraman et al., 2022].

Advancements in NLP and ML have revolutionized our ability to detect and analyze hate speech within the realm of political discourse. In recent years, researchers and technologists have developed sophisticated models trained on large datasets to automatically identify and categorize hateful language in political contexts. Now, we delve into the landscape of existing NLP and ML models tailored specifically for detecting hate speech, exploring their methodologies, strengths, and limitations.

4.2 Implicit Hate Speech

We introduced the concept of implicit hate speech in Subsection 2.3, which is defined as the use of coded language to incite hate against people belonging to particular religion, ethnicity or gender. We also mentioned that such types of coded hate speech is difficult to detect by classifiers trained on more overt forms of hate speech because of it's dependence on tokens found in more overt forms of hate speech [Waseem et al., 2017; Wiegand et al., 2019]. In this section we will describe briefly the datasets that have been created, and the SOTA models that have been developed to detect such coded language. As mentioned in 3.3, [ElSherief et al., 2021] released a dataset to encourage efforts in detection of implicit hate speech. They collected data from Twitter by focusing on eight different ideological clusters in the US defined by SPLC [2019] namely Black Separatist (27.1%), White Nationalist (16.4%), NeoNazi (6.2%), Anti-Muslim (8.9%), Racist Skinhead (5.1%), Ku Klux Klan (5.0%), Anti-LGBT (7.4%), and Anti-Immigrant (2.12%). Before this Sap et al. [2020] released another dataset named SBIC dataset where they used a method known as Social Bias Frames to annotate data. Hartvigsen et al. [2022] also released a dataset called ToxiGen where they used Chat-GPT to generate text data containing implicit hate speech. Progress have also been made in development of novel methodologies, Ocampo et al. [2023a] performed an in-depth study to facilitate a taxonomy of implicit hate speech messages, and also discovered misannotations in SBIC dataset using their taxonomy and qualitative analysis. Ocampo et al. [2023b] also proposed a methodology of increasing the performance of classifiers to detect implicit hate speech by using adversarial training. Huang et al. [2023] used a method which they named chain-of-explanation method inspired by the chain-of-thought prompting method developed by Wei et al. [2023] using Large Language models like GPT-2 [Radford et al., 2019], GPT-Neo [Gao et al., 2020], BART [Zhang et al., 2022], OPT [Lewis et al., 2019] and T5 [Raffel et al., 2019] models. Yang et al. [2023] proposed another framework which leverages the reasoning capabilities of LLMs to detect implicit hate as well as provide the reasons for detection.

Although these methods have shown great promise in detection of hate speech, the question remains that how we can build better detection systems because the Large Language Models are prone to biases [Gallegos et al., 2024]. Additionally some of the prompting methods used augment data (as in the case of ToxiGen dataset [Hartvigsen et al., 2022]) will not work at present because of guardrails that are present when interacting with closed source LLMs [Dong et al., 2024]. So one should avoid being over dependent on closed models like ChatGPT and GPT-4 and develop new methodologies that can leverage off the shelf pre-trained LLMs [Gururangan et al., 2020] for tuning the models based on data selected using different quality and diversity metrics. The detection of hate speech is still an open problem and with rising usage of social media it has become imperative to not only put efforts into detecting overt forms of hate but also the insidious forms of hate hidden in implicit hate speech. In the next section,we will focus on the datasets and methodologies that have been built to tackle the problem of the spread of Deepfakes and their limitations.

4.3 Deepfakes

4.3.1 Audio

The definition of deepfake audio is commonly accepted as an audio in which “important attributes have been manipulated via AI technologies while still retaining its perceived naturalness” [Yi et al., 2023]. The authors [Yi et al., 2023] classify the types of deepfake audio as text-to-speech, voice conversion, emotion fake, scene fake and partially fake. The most relevant for deepfake generation are text-to-speech and voice conversion.

Text-to-speech refers to the creation of natural speech from text based on deep learning models [Yi et al., 2023]. As pointed out by the authors, two recent results achieved the state of the art in text-to-speech using generative techniques. The first, by Huang et al. [2022], use a diffusion model as a generating method and the second, by Kim et al. [2021], use conditional variational autoencoders with adversarial learning.

Voice conversion is the process of producing audio content that simulates the voice of a given person [Walczyna and Piotrowski, 2023]. It typically consists of components that extract characteristics of a persons’ voice and another component that uses them to generate the audio content. A seminal work in this task was done by Jia et al. [2018]. The authors create a text-to-speech model capable of synthesizing voices of multiple speakers. In order to avoid recording large amounts of data for many speakers, the authors decouple speaker feature extraction from speech generation. Doing these tasks separately allows each of them to be trained independently. Their speech encoder follows the same architecture of Wan et al. [2018] and is trained to generate embeddings that maximize the cosine similarity for utterances of the same speaker and minimize for different speakers. As a vocoder, they use wavenet [Van Den Oord et al., 2016]. More recent works tended to use Variational Autoencoders (VAE), like Long et al. [2022].

The speaker features and the text are passed to a generative architecture, typically consisting of an encoder/decoder, to generate a speech spectrogram, which then will be fed into a vocoder to produce the audio record. As pointed out by Walczyna and Piotrowski [2023], generators and vocoders tend to use generative adversarial networks (GANs).

As pointed out by Zhang et al. [2023], deepfake audio detections may suffer from catastrophic forgetting, that is, a model trained for one dataset, when fine tuned for another one, may lose part of its predictive capabilities on the first dataset. To mitigate this issue, Zhang et al. [2023] propose Regularized Adaptive Weight Modification, which projects new learned knowledge close to the old knowledge feature space, to allow the model to learn how to classify the new data, while preserving the already learned knowledge.

Deepfake audio detection has also been applied in anti spoofing scenarios. As noted by Khan and Malik [2023], Siamese Networks have been a popular choice in detecting generated audios. The authors themselves use Siamese Networks with shared weights to extract voice embeddings and metric learning to differentiate between true and spoofed audios.

4.3.2 Image and video

GANs and autoencoders are the most prevalent deepfake generation techniques [Naitali et al., 2023]. In the training process, the autoencoders learn the capacity to compress learned facial features into a lower-dimensional space and this space can be used to generate new faces. Akhtar [2023] presents a detailed review of methods for generation and detection of the following categories of deepfakes:

- Identity swap: Identity swap is the substitution of a person's face in an image with the face of another person. CNNs and GANs have been used in face swap models.
- Attribute manipulation: it's the modification of facial attributes. GANs and VAEs have been the traditional models used for attribute manipulation.
- Face synthesis: this is the generation of non-real faces. Similar to other deepfake methods, GANs have been used in face synthesis.

Nevertheless, according to Akhtar [2023], there are some issues with deepfake detection:

- Generalizability: models tend to perform well on training data, but have a hard time reproducing this performance on unseen data.
- Explainability: like any deep learning model, it's not totally clear how deepfake detection models work and which characteristics in the data influence them the most.

- Technological evolution: it's hard for current models to keep up with the constant emergence of new deep fake models.
- Vulnerability to adversarial attacks: Adversarial models can be trained to fool detection models, making them vulnerable to this type of attacks.

Table 1 shows common deepfake datasets used for deepfake detection.

Name	Reference	Media	Size
ASVspoof	Nautsch et al. [2021]	Audio	107 speakers
Celeb-DF	Li et al. [2020]	Video	590 real + 5639 fake
UADFV	Yang et al. [2019]	Video	49 real + 49 fake
DF-TIMIT	Korshunov and Marcel [2018]	Video	640 fake
FF-DF	Rossler et al. [2019]	Video	1000 real + 1000 fake
DFD	Trinh and Liu [2021]	Video	363 original + 3068 fake
DFDC	Dolhansky et al. [2019]	Video	1132 real + 4113 fake

Table 1: Deepfake datasets.

4.4 Hyperpartisan News

Understanding the limitations of datasets is crucial in the context of hyperpartisan detection. While datasets serve as the backbone of machine learning algorithms in identifying hyperpartisan content, they are not immune to certain constraints. These limitations encompass various factors such as biases, imbalances, and incompleteness, which can significantly impact the effectiveness and reliability of detection models. By acknowledging these constraints, researchers and practitioners can adopt more nuanced approaches and develop strategies to mitigate potential pitfalls in hyperpartisan content detection.

In this section, we will explore the inherent limitations present in datasets utilized within the domain of hyperpartisan news detection. For a comprehensive overview of these datasets, we encourage you to consult [Maggini et al., 2024]. Firstly, existing datasets often fall short of meeting the criteria necessary for accurately defining the problem at hand. Specifically, due to the nuanced rhetorical structures inherent in hyperpartisan articles, there is a notable absence of a finely-grained, all-encompassing dataset capturing biases and rhetorical fallacies at the same time. Essentially, what is lacking is a dataset where foundational biases are meticulously annotated at the word or sentence level. Such a resource would not only deepen our understanding of the rhetorical strategies employed by partisan journalists but also shed light on the minimal elements responsible for bias, which could potentially be neutralized.

Moreover, the concept of hyperpartisan news is broadly defined [Schedler, 2023], and within the realm of Computer Science, there is no consensus on its precise definition. This suggests that the datasets constructed thus far may not accurately represent any particular political theory but rather rely primarily on general definition [Kiesel et al., 2019] or political leaning features coming from the news source [Liu et al., 2022]. Introducing a methodological framework could potentially facilitate the application of political scoring metrics such as scale-points. This approach could enable the computational assessment of partisanship levels in the media to be objectively quantified. It is imperative not only to accurately delineate the problem but also to consider the socio-political dynamics of the relevant countries.

Additionally, the predominance of English datasets underscores the dominance of Anglophone countries in this field, resulting in a skewed portrayal of polarization levels in other democracies. This issue is compounded by the current limitations in data availability. For instance, the legal dispute between the New York Times and OpenAI highlighted tensions surrounding the use of copyrighted materials. Consequently, to safeguard intellectual property, news outlets have restricted access to and usage of their content, further constraining data accessibility. This factor contributes to the prevailing scarcity of data. Furthermore, the rapid obsolescence of data on polarizing topics is another significant factor to consider. Trending issues that fuel societal polarization can quickly become outdated. Addressing contemporary societal challenges necessitates the availability of up-to-date datasets.

In conclusion, numerous factors impede progress in the field, primarily contingent upon the availability of current data. Future research efforts are likely to focus on addressing the gaps identified herein, ultimately contributing to advancements in the detection of hyperpartisan political content.

4.5 Automated Fact-checking

Our previous deliverable D3.1 provides a comprehensive view of existing datasets and state-of-the-art models in fact-checking research. This section discusses the research gap and open challenges associated with the existing datasets and models in the fact-checking research era.

Limited Datasets: one of the key aspects hindering the progress of fact-checking research is the unavailability of training data for specific needs. Especially, comprehensive multitopic claim detection datasets, verifiable claim type detection datasets, claim clustering datasets, and explainable claim detection are yet to be developed for the research progress even in monolingual settings. Automated generative approaches may be used as an alternative to generating suitable datasets [Bussotti et al., 2023; Veltri et al., 2023].

Validity of Data Sources: annotating a massive amount of factual statements for their verifiability, priority, similarity, and veracity is a tedious and expensive task. This resulted in relying on existing tools such as the Google Fact Checking tool to partially automate the creation of training datasets. Further, the transparency in the data gathering and annotation process often does not persist, and these factors question the credibility of the existing dataset as well as the solutions developed on it.

Consolidated Definition of the Tasks: defining verifiability, priority, and similarity of claims may depend on various factors such as source, topic, target audience, etc. Therefore, a wide range of definitions are used in the literature to tackle all three aspects of the claim detection problem. This highly hinders the research progress with unified agreement on the definition of the tasks.

Change of Claim Status with Time: both the true value and the requirement to determine the verifiability, priority, similarity, and veracity of claims may change over time. Further, incorporating this temporal nature of the problem is scarcely explored in the literature, mainly due to the unavailability of datasets meeting these objectives, and the challenges associated with simulating the real-time environment for accurate experiments.

Demand of Generalizable Solutions: as previously discussed, the source of factual statements can be from various platforms and can be articulated in various formats, languages, and modalities. Recent studies [Hale et al., 2024] have shown evidence of the existence of the same claims across multiple platforms, written in multiple formats, lengths, and details. While this demands more generalizable solutions to identify claims regardless of these factors, most of the existing research focuses on developing solutions specific to a source, data format, language, and modality.

Language Imbalance in Datasets: most of the existing multilingual datasets are composed of a higher number of annotated samples for high-resource languages such as English, compared to lesser-resourced languages. While existing research tried to tackle this problem via sampling, and augmenting data in the underrepresented languages through machine translation techniques, this could lead to biases in the training when the model is provided with more data on certain languages.

In summary, various factors affect the progress of fact-checking research. Notably, the unavailability of datasets meeting the task requirement serves as a key challenge. Promising future direction includes the development of credible and comprehensive datasets, generalized solutions, explainable fact-checking, and time-aware fact-checking.

5 Conclusions

In conclusion, our exploration into the landscape of political disinformation in social networks and the press reveals a complex and multifaceted phenomenon with far-reaching implications for democratic discourse and societal cohesion.

Firstly, the intertwining of political disinformation with hate speech underscores the urgent need for concerted efforts to combat online toxicity and promote respectful dialogue. Addressing hate speech requires a multifaceted approach that encompasses not only regulatory measures but also education and community-driven initiatives aimed at fostering empathy and understanding. Also this survey focused on the concept of implicit hate speech which can be used by dissenting parties to create problems while holding no accountability and the efforts that have been made so far into detecting them.

Secondly, the emergence of deep fake news poses a formidable challenge to the integrity of visual media and the authenticity of online content. As deep fake technology continues to evolve, it is imperative to invest in advanced detection methods and robust authentication mechanisms to mitigate the risks posed by synthetic media.

Moreover, the proliferation of hyperpartisan news highlights the corrosive effects of echo chambers and ideological polarization on public discourse. Efforts to counter hyperpartisanship must focus on promoting media literacy and critical thinking skills, empowering individuals to discern fact from opinion and navigate diverse viewpoints with discernment.

Furthermore, while fact-checking represents a crucial line of defense against political disinformation, its effectiveness hinges on the availability of reliable data and sophisticated algorithms capable of discerning truth from falsehood amidst the deluge of online content. Investing in the development of advanced fact-checking tools and fostering collaborations between researchers, journalists, and technology companies can enhance the efficacy of fact-checking initiatives and bolster public trust in information sources.

In essence, addressing the scourge of political disinformation demands a comprehensive and collaborative approach that engages stakeholders across government, civil society, and the private sector. By fostering greater transparency, accountability, and media literacy, we can fortify democratic resilience in the face of emerging threats and uphold the principles of truth, integrity, and informed citizenship in the digital age.

References

Agarwal, P., Hawkins, O., Amaxopoulou, M., Dempsey, N., Sastry, N., and Wood, E. (2021). Hate speech in political discourse: A case study of uk mps on twitter. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 5–16, New York, NY, USA. Association for Computing Machinery.

Akhtar, Z. (2023). Deepfakes generation and detection: A short survey. *Journal of Imaging*, 9(1):18.

Aksenov, D., Bourgonje, P., Zaczynska, K., Ostendorff, M., Moreno-Schneider, J., and Rehm, G. (2021). Fine-grained classification of political bias in german news: A data set and initial experiments. In Mostafazadeh Davani, A., Kiela, D., Lambert, M., Vidgen, B., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131. Association for Computational Linguistics.

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

Azizov, D., Nakov, P., and Liang, S. (2023). Frank at checkthat! 2023: Detecting the political bias of news articles and news media. In *Conference and Labs of the Evaluation Forum*.

Baines, D. and Elliott, R. (2020). Defining misinformation, disinformation and malinformation: An urgent need for clarity during the COVID-19 infodemic.

Barnidge, M. and Peacock, C. (2019). A third wave of selective exposure research? the challenges posed by hyperpartisan news on social media. *Media and Communication*, 7(3):4–7.

Beltrán, J., Míguez, R., and Larraz, I. (2021). Claimhunter: An unattended tool for automated claim detection on twitter. In *KnOD@ WWW*.

Blokker, N., Dayanık, E., Lapesa, G., and Padó, S. (2020). Swimming with the tide? positional claim detection across political text types. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 24–34.

Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Krack, N., Teyssou, D., France-Presse, A., Cuccovillo, L., and Verdoliva, L. (2024). Generative ai and disinformation: Recent advances, challenges, and opportunities. *European Digital Media Observatory*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Buchanan, B., Lohn, A., Musser, M., and Sedova, K. (2021). Truth, Lies, and Automation. Technical report, Center for Security and Emerging Technology.

Bussotti, J.-F., Veltri, E., Santoro, D., and Papotti, P. (2023). Generation of training examples for tabular natural language inference. *Proceedings of the ACM on Management of Data*, 1(4):1–27.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., and Granitzer, M. (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Chen, C. and Shu, K. (2024). Can LLM-Generated Misinformation Be Detected?

Crothers, E., Japkowicz, N., and Viktor, H. (2023). Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods.

Danry, V., Leong, J., Pataranutaporn, P., Tandon, P., Liu, Y., Shilkrot, R., Punpongsanon, P., Weissman, T., Maes, P., and Sra, M. (2022). AI-Generated Characters: Putting Deepfakes to Good Use. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–5.

Das, M., Raj, R., Saha, P., Mathew, B., Gupta, M., and Mukherjee, A. (2023). Hatemm: A multimodal dataset for hate video classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:1014–1023.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Dénigot, Q. and Burnett, H. (2020). Dogwhistles as Identity-based interpretative variation. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 17–25, Gothenburg. Association for Computational Linguistics.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.

Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., Meng, J., Ruan, W., and Huang, X. (2024). Building guardrails for large language models.

Dumitru, V.-C. and Rebedea, T. (2021). Topic-based models with fact checking for fake news identification. In *RoCHI - International Conference on Human-Computer Interaction*, pages 182–190. MATRIX ROM.

ElSherief, M., Kulkarni, V., Nguyen, D., Yang Wang, W., and Belding, E. (2018a). Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018b). Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., and Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ernesto de León, M. M. and Adam, S. (2024). Hyperpartisan, alternative, and conspiracy media users: An anti-establishment portrait. *Political Communication*, 0(0):1–26.

Fallis, D. (2014). The Varieties of Disinformation. In Floridi, L. and Illari, P., editors, *The Philosophy of Information Quality*, volume 358, pages 135–161, Cham. Springer International Publishing.

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, 34(4):623–643.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. (2024). Bias and fairness in large language models: A survey.

Gangula, R. R. R., Duggenpudi, S. R., and Mamidi, R. (2019). Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84. Association for Computational Linguistics.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. (2020). The pile: An 800gb dataset of diverse text for language modeling.

Garg, S. and Sharma, D. K. (2022). Role of ELMo embedding in detecting fake news on social media. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 57–60. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Google (2024). Gemma: Introducing new state-of-the-art open models.

Graham, T. and Bogle, A. (2022). Digital disinformation. In Glasser, R., Johnstone, C., and Kapetas, A., editors, *The Geopolitics of Climate and Security in the Indo-Pacific*, pages 84–90. Australian Strategic Policy Institute, Barton, ACT.

Grapă, T.-E. and Mogoș, A.-A. (2023). The spectacle of “patriotic violence” in romania: Populist leader george simion’s mediated performance. *Media and Communication*, 11(2).

Grimminger, L. and Klinger, R. (2021). Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings* *of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *CoRR*, abs/2004.10964.

Hale, S. A., Belisario, A., Mostafa, A., and Camargo, C. (2024). Analyzing misinformation claims during the 2022 brazilian general election on whatsapp, twitter, and kwai. *arXiv preprint arXiv:2401.02395*.

Hancock, J. T. and Bailenson, J. N. (2021). The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3):149–152.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Hassan, F. M. and Lee, M. (2020). Political fake statement detection via multistage feature-assisted neural modeling. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.

Hassan, N., Arslan, F., Li, C., and Tremayne, M. (2017). Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.

Hobolt, S. B., Lawall, K., and Tilley, J. (2023). The polarizing effect of partisan echo chambers. *American Political Science Review*, pages 1–16.

Huang, F., Kwak, H., and An, J. (2023). Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

Huang, R., Lam, M. W., Wang, J., Su, D., Yu, D., Ren, Y., and Zhao, Z. (2022). Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*.

Hussaini, S. A., Ali, M. T., Ali, M., and Vishvapathi, P. (2022). Rumour detection in the political domain from twitter using machine learning techniques. In *International Conference on Information and Management Engineering*, pages 669–677. Springer.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I., Wu, Y., et al. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31.

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Karimi, H., Roy, P., Saba-Sadiya, S., and Tang, J. (2018). Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1546–1557.

Khan, A. and Malik, K. M. (2023). Securing voice biometrics: One-shot learning approach for audio deepfake detection. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., and Potthast, M. (2019). SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839. Association for Computational Linguistics.

Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Kim, Y., Park, S., and Han, Y.-S. (2022). Generalizable implicit hate speech detection using contrastive learning. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ko, Y., Ryu, S., Han, S., Jeon, Y., Kim, J., Park, S., Han, K., Tong, H., and Kim, S.-W. (2023). KHAN: Knowledge-aware hierarchical attention networks for accurate political stance prediction. *Proceedings of the ACM Web Conference 2023*, pages 1572–1583. Conference Name: WWW '23: The ACM Web Conference 2023 ISBN: 9781450394161 Place: Austin TX USA Publisher: ACM.

Korshunov, P. and Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Kristoffer Holt, T. U. F. and Frischlich, L. (2019). Key dimensions of alternative news media. *Digital Journalism*, 7(7):860–869.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.

Liu, Y., Zhang, X. F., Wegsman, D., Beauchamp, N., and Wang, L. (2022). POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Long, Y., Lu, Q., Xiang, R., Li, M., and Huang, C.-R. (2017). Fake news detection through multi-perspective speaker profiles. In *Proceedings of the eighth international joint conference on natural language processing (volume 2: Short papers)*, pages 252–256.

Long, Z., Zheng, Y., Yu, M., and Xin, J. (2022). Enhancing zero-shot many to many voice conversion with self-attention vae. *arXiv preprint arXiv:2203.16037*.

Lyu, H., Pan, J., Wang, Z., and Luo, J. (2023). Computational assessment of hyperpartisanship in news titles.

Maggini, M., Bassi, D., Gamallo, P., Piot, P., and Dias, G. (2024). A systematic review of hyperpartisan news detection: A comprehensive framework for definition, detection, and evaluation.

Masud, S. and Charaborty, T. (2023). Political mud slandering and power dynamics during indian assembly elections. *Social Network Analysis and Mining*, 13(1).

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

McCoy, J. and Somer, M. (2019). Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *The ANNALS of the American Academy of Political and Social Science*, 681(1):234–271.

Molina, M. D., Sundar, S. S., Le, T., and 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.

Naitali, A., Ridouani, M., Salahdine, F., and Kaabouch, N. (2023). Deepfake attacks: Generation, detection, datasets, challenges, and research directions. *Computers*, 12(10):216.

Nations, U. (2023). What is hate speech?

Nautsch, A., Wang, X., Evans, N., Kinnunen, T. H., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J., and Lee, K. A. (2021). Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2):252–265.

Nguyen, T. T., Huynh, T. T., Yin, H., Weidlich, M., Nguyen, T. T., Mai, T. S., and Nguyen, Q. V. H. (2023). Detecting rumours with latency guarantees using massive streaming data. *The VLDB Journal*, 32(2):369–387.

Nguyen, T. T., Ren, Z., Nguyen, T. T., Jo, J., Nguyen, Q. V. H., and Yin, H. (2024). Portable graph-based rumour detection against multi-modal heterophily. *Knowledge-Based Systems*, 284:111310.

Novoa, G., Echelbarger, M., Gelman, A., and Gelman, S. A. (2023). Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, 120(47):e2309361120.

Ocampo, N., Sviridova, E., Cabrio, E., and Villata, S. (2023a). An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Ocampo, N. B., Cabrio, E., and Villata, S. (2023b). Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

Patel, S., Bansal, P., and Kaur, P. (2022). Rumour detection using graph neural network and oversampling in benchmark twitter dataset. *arXiv preprint arXiv:2212.10080*.

Patwari, A., Goldwasser, D., and Bagchi, S. (2017). Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2259–2262.

Pawelec, M. (2022). Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *Digital Society*, 1(2):19.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. L., and Sheth, A. (2018). A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 33–36, New York, NY, USA. Association for Computing Machinery.

Ross Arguedas, A., Robertson, C. T., Fletcher, R., and Nielsen, R. K. (2022). Echo chambers, filter bubbles, and polarisation: a literature review. Technical report, [object Object].

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). Whose opinions do language models reflect? In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Schedler, A. (2023). Rethinking Political Polarization. *Political Science Quarterly*, 138(3):335–359.

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., and Zhao, J. (2023). Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. *2023 International Conference on Computer and Applications (ICCA)*, pages 1–7.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2021). Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):687–690.

Søe, S. O. (2014). Information, misinformation og disinformation : En sprogfilosofisk analyse. In *Nordisk Tidsskrift for Informationsvidenskab Og Kulturformidling*, volume 3, pages 21–30.

Solovev, K. and Pröllochs, N. (2022). Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. In *Proceedings of the ACM Web Conference 2022*, WWW '22. ACM.

SPLC (2019). Hate map.

Tam, N. T., Weidlich, M., Zheng, B., Yin, H., Hung, N. Q. V., and Stantic, B. (2019). From anomaly detection to rumour detection using data streams of social platforms. *Proceedings of the VLDB Endowment*, 12(9):1016–1029.

Toraman, C., Şahinuç, F., and Yilmaz, E. (2022). Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Trinh, L. and Liu, Y. (2021). An examination of fairness of ai models for deepfake detection. *arXiv preprint arXiv:2105.00558*.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Veltri, E., Badaro, G., Saeed, M., and Papotti, P. (2023). Data ambiguity profiling for the generation of training examples. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 450–463. IEEE.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., and Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, 595(7866):197–204.

Walczyna, T. and Piotrowski, Z. (2023). Overview of voice conversion methods based on deep learning. *Applied Sciences*, 13(5):3100.

Wan, L., Wang, Q., Papir, A., and Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Weiai Wayne Xu, Y. S. and Kim, C. (2020). What drives hyper-partisan news sharing: Exploring the role of source, style, and content. *Digital Journalism*, 8(4):486–505.

Wich, M., Bauer, J., and Groh, G. (2020). Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of Abusive Language: The Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Wintersieck, A., Fridkin, K., and Kenney, P. (2021). The message matters: The influence of fact-checking on evaluations of political messages. *Journal of Political Marketing*, 20(2):93–120.

Yang, J., Rojas, H., Wojcieszak, M., Aalberg, T., Coen, S., Curran, J., Hayashi, K., Iyengar, S., Jones, P. K., Mazzoleni, G., Papathanassopoulos, S., Rhee, J. W., Rowe, D., Soroka, S., and Tiffen, R. (2016). Why Are “Others” So Polarized? Perceived Political Polarization and Media Use in 10 Countries. *Journal of Computer-Mediated Communication*, 21(5):349–367.

Yang, X., Li, Y., and Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE.

Yang, Y., Kim, J., Kim, Y., Ho, N., Thorne, J., and young Yun, S. (2023). Hare: Explainable hate speech detection with step-by-step reasoning.

Yasser, K., Kutlu, M., and Elsayed, T. (2018). bigir at clef 2018: Detection and verification of check-worthy political claims. In *CLEF (Working Notes)*.

Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., and Zhao, Y. (2023). Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*.

York, C., Ponder, J. D., Humphries, Z., Goodall, C., Beam, M., and Winters, C. (2020). Effects of fact-checking political misinformation on perceptual accuracy and epistemic political efficacy. *Journalism & Mass Communication Quarterly*, 97(4):958–980.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models.

Zhang, X., Yi, J., Tao, J., Wang, C., and Zhang, C. Y. (2023). Do you remember? overcoming catastrophic forgetting for fake audio detection. In *International Conference on Machine Learning*, pages 41819–41831. PMLR.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *Acm Computing Surveys (Csur)*, 51(2):1–36.

Author's contributions

Michele Joshua Maggini was responsible for section 2.5, 3.5, 4.4 and overall revision of the manuscript.

Paloma Piot was responsible for sections 2.2, 3.2 and 4.1 and overall revision of the manuscript.

Rabiraj Bandyopadhyay was responsible for section 2.3, 3.3, 4.2 and overall revision of the manuscript.

Rafael Martins Frade was responsible for section 4.3 and overall revision of the manuscript.

Rrubaa Panchendrarajan was responsible for section 3.6, section 4.5 and overall revision of the manuscript.

Søren Fomsgaard was responsible for sections 2.1, 2.4, 3.1, 4.3 and overall revision of the manuscript.