Funded by the European Union

UK Research and Innovation

Deliverable number: D3.1

hybrids

Technical report on the state of the art of NLP and AI methods for disinformation detection

A holistic understanding of the evolving landscape of disinformation detection through natural language processing and artificial intelligence.

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:	D3.1
Title:	Technical report on the state of the art of NLP and AI methods for disinformation detection
Work Package	WP3: Democracy Threats
Lead beneficiary:	UNICAEN
Deliverable nature:	R- Document, report
Dissemination level:	PU-Public
Due Date (month):	31/01/2024 (M13)
Submission Date (month):	31/01/2024 (M13)
Keywords:	Misinformation, fact-checking, hate speech detection, hyperpartisanism, disinformation agents and health disinformation

Abstract

This technical report presents a comprehensive overview of the current state of Natural Language Processing and Artificial Intelligence methods in disinformation detection. Focusing on diverse contexts, it begins by examining automated fact-checking, showcasing the latest advancements in NLP and AI to discern truth from falsehood. Shifting to health disinformation, it explores the challenges in detecting false health-related information, drawing connections between health disinformation detection and broader fact-checking frameworks. The discussion then moves to hate speech detection, illustrating how NLP and AI technologies are utilized to combat online hate speech, emphasizing the goal of maintaining a secure and inclusive online environment. Hyperpartisan detection is explored next, revealing the interconnected nature of hyperpartisan content and its impact on the overall disinformation landscape. The report concludes by addressing automated disinformation agents, highlighting the role of NLP and AI in distinguishing between human and bot-generated content to curb the amplification of false narratives. Throughout, the report defines key concepts, reviews current works, discusses approaches and available datasets, and outlines future directions.

Disclaimer & acknowledgment

This project has received funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee - Grant Number: EP/X036758/1. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission or UKRI. The REA or UKRI cannot be held responsible for any use that may be made of the information this document contains.

Deliverable Contributors

	Name	Institution	E-mail
Deliverable leader	Gaël Dias	UNICAEN	gael.dias@unicaen.fr
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	Pablo Gamallo Otero	CiTIUS, USC	pablo.gamallo@usc.es
	Claudia Wagner	GESIS	Claudia.Wagner@gesis.org
	Rubén Miguez	NEWTRAL	ruben.miguez@newtral.es
	Patricia Martín Rodilla	UDC	patricia.martin.rodilla@udc.es
	Javier Parapar López	UDC	javier.parapar@udc.es
	Arkaitz Zubiaga	QMUL	a.zubiaga@qmul.ac.uk

History of Changes

Version	Date	Changes to previous version	Status
0.1	14/11/2023	First Draft	Draft
0.2	23/01/2024	Consortium Internal review	Review
1.0	31/01/2024	Approved version to be submitted	Final

Disclaimer & acknowledgment

This project has received funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee - Grant Number: EP/X036758/1. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission or UKRI. The REA or UKRI cannot be held responsible for any use that may be made of the information this document contains.

Acronym	Meaning
AEDA	An Easier Data Augmentation
ANN	Approximate Nearest Neighbors
BART	Bidirectional Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformers
HNSW	Hierarchical Navigable Small Words
IR	Information Retrieval
LSH	Locality Sensitive Hashing
mBERT	Multilingual BERT
NER	Named Entity Recognition
NLP	Natural Language Processing
OSN	Online Social Network
POS Tag	Part-of-speech Tag
RAG	Retrieval-Augmented Generation
RoBERTa	Robustly Optimized BERT Pre-training Approach
RP	Random Projections
SBERT	Sentence BERT
T5	Text-to-Text Transfer Transformer
TF-IDF	Term Frequency-Inverse Document Frequency
XLM-r	Cross-Lingual Language Model with RoBERTa Architecture

Contents

1 Introduction

In an era marked by the rapid evolution of digital communication, the challenge of identifying and mitigating disinformation has become paramount. This technical report provides a comprehensive overview of the current state of the art in Natural Language Processing (NLP) and Artificial Intelligence (AI) methods tailored for disinformation detection across diverse contexts.

As we delve into various dimensions of disinformation, our exploration begins with the realm of automated fact-checking. This section surveys the latest advancements in NLP and AI, establishing a foundation for understanding how technology is employed to discern truth from falsehood.

Transitioning to health disinformation, we explore the unique challenges posed by the spread of false health-related information. We examine the current landscape of AI-driven solutions, drawing connections between methodologies employed in health disinformation detection and broader fact-checking frameworks.

Moving further, the discussion shifts to hate speech detection. In this segment, we explore how NLP and AI technologies are harnessed to identify and combat the proliferation of hate speech online. We connect the dots between hate speech detection and the overarching goal of maintaining a secure and inclusive online environment.

The report then delves into hyperpartisan detection, elucidating how political polarization contributes to the spread of misinformation. We highlight the interconnected nature of hyperpartisan content and its impact on the broader disinformation landscape.

Finally, we address the pervasive influence of automated disinformation agents, commonly known as bots. This section explores the role of NLP and AI in differentiating between human and bot-generated content, underscoring the integral role of technology in curbing the amplification of false narratives.

Throughout each exploration, we define key concepts, review current works, and discuss the approaches taken. By considering available datasets and outlining future directions, this report aims to provide a holistic understanding of the evolving landscape of disinformation detection, where each section builds upon the interconnected threads of NLP and AI methodologies.

2 Automated fact-checking

2.1 Context

Misinformation poses a significant threat to society, and this threat has escalated with the advent and widespread use of social media platforms. However, manually verifying the content circulating on online platforms is a time-consuming task, and this situation demands the development of automated fact-checking, a process of verifying whether a claim is true or false. This is often carried out as a series of steps involving, the identification of claims to be checked, prioritization of important claims, evidence gathering, and finally verdict prediction. While there are several dedicated organizations such as PoliFact[1] and Fullfact[2] established over the past couple of years, the research in this direction is experiencing a substantial increase due to the presence of emerging challenges.

In this chapter, we outline the recent research and techniques related to automated fact-checking. Specifically, we introduce the fact-checking pipeline, state-of-the-art techniques used at different stages of the pipeline, and the datasets available. Further, the chapter is concluded with some directions for future research.

2.2 Definitions

The fact-checking process is often carried out as a sequence of tasks comprising the detection of claims circulated in online platforms, followed by verification of claims. The process begins with detecting verifiable factual statements, referred to as *claims*. This component is commonly

[1] https://www.politifact.com/

[2] https://fullfact.org/

applied to social platforms and online resources such as news articles to identify statements that require verification. Once the claims are extracted, they go through a prioritization process to estimate the worthiness of the claim to be verified. The criteria used to estimate the worthiness may vary according to the topic or domain of the claim and the user groups who are interested in the veracity of the claim. Some popular criteria used in the literature are the virality of the claim, the interest of the public in the veracity of the claim, the impact that the claim could create, and its timeliness [Das et al., 2023, Micallef et al., 2022].

Following the prioritization step, evidence supporting or refuting the prioritized claims is retrieved, and finally, the verdict of the claim indicating whether the fact discussed in the statement is true or false is predicted with respect to the evidence retrieved [Guo et al., 2022b]. Often, the evidence retrieval and veracity prediction tasks are tackled together in the literature as a *fact verification* process [Guo et al., 2022b]. The latest addition to the automated fact-checking pipeline is the explanation generation [Kotonya and Toni, 2020a], where the researchers aim to automatically generate a reason for the verdict prediction. Apart from these five major components, researchers have focused on retrieving claims similar to an unverified claim from a database of fact-checked claims for effective fact-checking. This task is referred to as *verified claim retrieval* or *claim matching* in the literature. While this can avoid the substantial time involved in processing an unverified claim through the remaining components of the pipeline, the impact and spread of the claim can also be minimized with timely verdicts.

2.3 Approaches

2.3.1 Claim detection

Claim detection task is often treated as a binary classification problem to identify whether a statement is a verifiable claim or not. However, the problem can also be handled as a multi-class classification by either adding an uncertainty label [Kazemi et al., 2021a] or identifying fine-grained verifiable claim types [Konstantinovskiy et al., 2021]. Apart from detecting verifiable claims, prioritizing claims is also often solved as a claim detection problem, where the objective is to classify a factual statement as worthy of verifying or not. Depending on the criteria used to prioritize, various tasks including check-worthy claim detection [Nakov et al., 2022, Shaar et al., 2021a], attention-worthy claim detection [Nakov et al., 2022], and harmful claim detection [Nakov et al., 2022, Shaar et al., 2021a] have been introduced in the literature.

Due to the powerful nature of the transformer architectures in understanding the language and the tasks, several studies utilize models from the transformer family by fine-tuning them in language-specific training data for developing monolingual claim detection solutions [Williams et al., 2021, Zhou et al., 2021] or by fine-tuning multilingual transformers for performing multilingual claim detection. Notable transformer models include mBERT [Hasanain and Elsayed, 2020, Tarannum et al., 2022, Sadouk et al., 2023], XLM-r [Tarannum et al., 2022, Sadouk et al., 2023, Aziz et al., 2023], and GPT-3 [Sadouk et al., 2023]. Several studies reported that fine-tuning multilingual transformers obtain similar or increased performance compared to the monolingual models. Especially, Panda et al. [Panda and Levitan, 2021], who utilized mBERT [Devlin et al., 2018] for the classification task in the NLP4IF 2021 dataset [Shaar et al., 2021a], and the authors reported that mBERT can achieve an impressive score in identifying misinformation labels even without fine-tuning on language-specific training data. Similar studies [Hasanain and Elsayed, 2022, Kartal and Kutlu, 2022] analyzed the zero-shot learning by training the mBERT model only in training data of one language and testing its generalization capability in other languages. Further, a recent study by Agrestia et al. [Agrestia et al., 2022] showed, that fine-tuning GPT-3 model [Radford and Narasimhan, 2018] using English data only gives competitive performance to the BERT models trained on language-specific training data for both verifiable claim detection and claim prioritization tasks.

Limited training data and the imbalanced nature of the data in terms of prediction classes or languages serve as key challenges in claim detection. Various data augmentation techniques including data sampling [Zengin et al., 2021] and machine translation [Suri and Dudeja, 2022, Savchev, 2022, Nakov et al., 2021], and performing multi-tasking [Schlicht et al., 2021] are explored as a solution to overcome these challenges. For example, Savchev et al. [Savchev, 2022] used back translation, translating a text to a target language, and then translating back from the target language to the original language as the data augmentation technique. They observed an increase in overall performance with the incorporation of the data augmentation technique. Similarly, Schlicht et al. [Schlicht et al., 2021] performed multi-task learning by jointly detecting the language and check-worthy claims. The authors used Sentence BERT [Reimers and Gurevych, 2019] trained on a multilingual dataset with dedicated fully connected layers for each task. Du et al. [Du et al., 2022] extended this work by performing a wide range of auxiliary tasks to enhance the performance, and observed an increase in performance for check-worthy detection tasks. Notable auxiliary tasks jointly learned include translation to English, verifiable claim detection, harmful tweet detection, and attention-worthy tweet detection.

Different from these approaches, Konstantinovskiy et al. [Konstantinovskiy et al., 2021] performed a fine-grained analysis of verifiable claims by classifying a sentence into non-claim or six sub-categories of a claim. The authors annotated around 6,300 sentences from subtitles of television shows and trained various traditional machine learning models with a wide range of features. Notable textual features include TF-IDF, Part-of-speech (POS) tags, Named entity recognition (NER), and word embedding. The authors observed, that the logistic regression classifier [LaValley, 2008] obtained the highest F1 score in classifying the sentences as a claim or no claim, and injecting POS and NER information did not improve the performance of the optimal classifier. The proposed solution was tested in a live feed of transcripts from television shows. Similar claim type identification has been carried out in the literature using rule-based approaches [Rony et al., 2020].

The latest attention on claim detection has been given to developing domain-specific solutions. Woloszyn et al. [Woloszyn et al., 2021] focused on identifying *green claim*, a claim discussing an issue related to the environment. The authors compared three pre-trained models, RoBERTa [Liu et al., 2019a], BERTweet, and Flair [Akbik et al., 2018], and observed that generally, RoBERTa outperformed the other two models in the green claim detection task. Smeros et al. [Smeros et al., 2021] extracted scientific claims by introducing three variants of BERT, SciBERT, NewsBERT, and SciNewsBERT fine-tuned using scientific articles and news headlines. Similarly, [Pathak and Srihari, 2021, Pathak et al., 2020] developed solutions specific to news articles based on the assumption that sentences that could well represent the headlines are more check-worthy, and experimented with unsupervised techniques to identify check-worthy sentences. Gollapalli et al. [Gollapalli et al., 2023] attempted to extract medical claims and claim types discussing prevention, diagnoses, cures, treatments, and risks. The authors fine-tuned the Text-to-Text Transfer Transformer (T5) model [Raffel et al., 2020] for identifying claim priority, and the BART [Lewis et al., 2020] model was used to detect the claim types in a zero-shot setting.

2.3.2 Claim matching

Claim matching is the task of identifying a pair of claims that can be addressed with the same fact-check [Kazemi et al., 2021b]. This can be handled as either a classification task to classify whether the two claims match or not, or a regression or semantic similarity task to generate a score indicating the strength of the match. When modeling as a classification problem, the likelihood of the classifier can also be used as a score indicating the probability of the two statements discussing the same claim. The task can be further extended as a search problem in a database of verified claims, by producing a ranked list of verified claims matching the input claim using the scores obtained via classification, or regression, or semantic similarity function. This extended task is referred to as *fact-checked claim retrieval* or *verified claim retrieval*.

Given the growing number of fact-checking organizations, the retrieval of fact-checked claims became an important step in the automated fact-checking process. One of the first papers to address claim-matching was [Shaar et al., 2020a]. The authors set the basic structure of what would be the claim-matching pipeline, namely using a fast lexical search algorithm to select fact-check candidates for a claim, then language models to rank the candidates, and a final layer or re-ranking based on learning to rank. Comparing BERT, RoBERTa and SBERT, the best results were given by SBERT. In the task, the authors were trying to find fact-checked articles for claims, and they discovered that using the body of the fact-checked articles gave better results than just using the title.

A leading component of the automated fact-checking community consists of open competitions. A traditional competition that involved claim-matching was the CLEF-CheckThat, which had claim-matching tasks in the 2020 to 2022 editions [Shaar et al., 2020b, Shaar et al., 2021b, Nakov et al., 2022]. The best-performing team in the 2020 CheckThat competition was Buster.ai which used a mix of data augmentation and training on extra datasets to improve scores [Bouziane et al., 2020]. Besides using the training data provided by the competition, the authors also tried using three datasets known to the fact-checking community: FEVER, SciFact and Liar. Maybe due to having a syntax somewhat different from test data (formed by tweets) or because of not being big datasets, the authors mention in the paper that SciFact and Liar didn't improve the results, which can also suggest that the high scores achieved in the competition might not generalize so well in different contexts. Training on FEVER improved the performance by 1%. The team also tried back-translation and NER as augmentation strategies, but they did not increase performance. Another interesting approach to the task in 2020, done by team UNIPI-NLE [Passaro et al., 2020], was to select the top 2.5k candidates (among 10k) using elastic search, after which they ranked candidates based on SBERT similarity scores. UNIPI-NLE got the second-highest score.

The 2021 edition of the CheckThat [Shaar et al., 2021b] has the same ranking task, but also included tweets in Arabic and fact-checked claims in political debates and speeches. The solutions were usually similar: training a model of the BERT family (like Roberta, or AraBERT) and then using a ranking technique. Team Aschern [Chernyavskiy et al., 2021], the best performing team on the English task, fine-tuned SBERT on the dataset and then used LambdaMART to re-rank the top 20 matches. An approach worth mentioning was the one made by Team DIPS [Mihaylova et al., 2021], which also used SBERT, but did some preprocessing on the data, by splitting hashtags, removing emojis and using the date information, then fed similarity scores to a neural network. Another re-ranking strategy used was rankSVM [Skuczyńska et al., 2021].

Finally, for the 2022 edition [Nakov et al., 2022], the last edition with a claim-matching task, the top performing team was RIET [Shliselberg and Dori-Hacohen, 2022], which used the traditional approach of filtering candidates and then re-ranking. Instead of using BM25, they used sentence-t5 to select candidates, and to re-rank them, they used a large autoregressive language model gpt-neo [Black et al., 2021]. The second place used ElasticSearch for selecting candidates, SBERT and SVM for re-ranking. Other teams used a mix of preprocessing (stemming, removing stop words), cheap algorithms for preselection (BM25) and semantic similarity for re-ranking.

2.3.3 Evidence retrieval

An important research effort to advance evidence retrieval for fact-checking was done at the FEVER/FEVEROUS shared tasks [Aly et al., 2021]. The tasks gave 87K verified claims to participants along with 95M sentences and 12M tables that either support, refute or do not contain enough information about the claim. Most teams applied the same techniques as participants of the claim-matching task of the CLEF-CheckThat challenges, like BM25, TF-IDF, or re-ranking based on models of the BERT family.

As we saw in the strategies of the teams competing in the CLEF-CheckThat competition, a common way of ranking verified claims is to use classic algorithms, like BM25, in a first ranking stage, and then use a better performing model, like rankSVM or a neural network, for re-ranking. Alternatively, one can even generate embeddings for the whole dataset and comparing the query to the documents by brute force. There are, however, other information retrieval (IR) algorithms that have showed great results in general IR tasks and have been adopted in the industry at large, but explored less in automated fact-checking research. These algorithms are especially useful in cases when we have to query large document sets. The techniques presented here can easily be applied to evidence retrieval or to search already fact-checking claims.

One of the reasons for the success of the new large language models was the possibility of encoding contextual text and visual information as dense vectors. This possibility also changed the way document retrieval is performed. Algorithms that perform document search based on vectors became the new paradigm in Approximate Nearest Neighbour search (ANN). We will discuss the main ones in this subsection.

Given the high dimension of current embeddings, a common strategy is to use dimensionality reduction before starting the search. Random projections are the backbone of several ANN algorithms [Bingham and Mannila, 2001]. Unlike PCA, random projections do not need to compute eigenvectors, which make them a convenient dimensionality reduction algorithm. The idea behind it is to project high dimensional data into a lower dimensional space in a way that the distance between the original vectors is preserved into the projected vectors. Instead of projecting the original data orthogonally, a computationally expensive operation, the data is projected using random lower dimensional matrices, which approximate orthogonality [Bingham and Mannila, 2001].

Random projections are a common way of implementing locality sensitive hashing (LSH) search. By projecting the data into a vector space, splitting the space into random hyperplanes and classifying the data vectors into buckets based on their position relative to the hyperplanes, LSH groups the original vectors in a way that similar data fall into similar buckets. Then, instead of searching the whole data, a query can also be put into a bucket and compared to the other buckets, allowing for a sublinear query performance [Jafari et al., 2021]. Another usual way of implementing LSH is using shingling and minhashing [Jafari et al., 2021].

In the past few years, an algorithm that became the state of the art in information retrieval was the Hierarchical Navigable Small Words (HNSW) [Malkov and Yashunin, 2018]. The idea behind it is to build a layered graph structure that balances connectedness and shortest path length in a way that a simple greedy search can efficiently match the nearest neighbor query. Elasticsearch, the well known search engine, uses HNSW as the default vector search algorithm.

2.3.4 Claim verification

Part of the challenges of automated-fact verification is due to the limitations of language models in performing reasoning. Since the development of the transformer architectures, the standard approach for claim verification has been to add an inference layer to a fine-tuned model of the BERT family. In the SCIVER shared task [Wadden and Lo, 2021], which asked participants to find evidence and evaluate scientific claims, the difference between the strategies of the participants resided on the preprocessing steps, the model they chose for inference, or how the evidence ranking was structured, but they all followed this standard approach. Some of the models chosen were RoBERTa [Liu et al., 2019a], SciBERT [Beltagy et al., 2019], BioBERT [Lee et al., 2020] and T5 [Raffel et al., 2020]. Some of the insights from competition were: until then, dense retrieval methods did not perform better than bag-of-words approaches; adding more context when looking for evidence in single sentences tended to perform better than looking for the evidence in the sentences alone; and using larger BERT models like RoBERTa-Large or T5 increased the scores, without the need for any additional changes [Wadden and Lo, 2021].

Another example of this approach was the FEVEROUS shared tasks [Aly et al., 2021] already mentioned in the evidence retrieval section 2.3.3. They involved not only finding textual evidence for the claims, but also classifying the claims based on the found evidence. Most teams used pre-trained BERT models fine-tuned on natural language inference datasets, like [Gi et al., 2021]. Since part of the information was in tables, many teams used the TAPAS model as part of the inference pipeline [Herzig et al., 2020]. The FEVEROUS organizers [Aly et al., 2021] pointed out that a common issue to the solutions presented, and already mentioned in other automated fact-checking reviews [Zeng et al., 2021], is that the claim verification step can be heavily influenced by noisy results from the previous steps, that is, errors are propagated down the pipeline.

2.3.5 Multimodal fact-checking

Since most content in social media tends to explore a mix of text, image, audio and video, in the last three years the automated fact-checking community has increased the efforts to deal with multimodal disinformation. The term multimodal has been applied to automated fact-checking contexts where either the disinformation or the evidence is represented in more than one modality [Mubashara et al., 2023].

An example of this effort was the Factify competitions. The dataset of the last competition (Factify 2) with published results [Suryavardan et al., 2023] contained 50,000 claims with images paired with a document and an image. The task asked participants to establish if the text supported, was insufficient or refuted the claim and the same with respect to the image. For the baseline model [Suryavardan et al., 2023], the organizers used SBERT to compute the similarity between the document and claim, Visual Transformers to do the same with the images and the similarities were fed into a classifier. An alternative baseline used ResNet50 to extract the features, but the results were worse. The best performing team in Factify 2 was team Triple-Check [Du et al., 2023], who used a DeBERTa for text embeddings and Swinv2 for image embeddings. The second on the list was team INO [Zhang et al., 2023], who used CLIP [Radford et al., 2021] and SBERT to extract text features and ResNets [Targ et al., 2016] for image features. Then, the similarities are fed into a Random Forest classifier.

Overall, using CLIP embeddings was quite popular among Factify contestants. Some teams also experimented with multimodal fusion. A limitation of this task and dataset is that the golden labels of the images were created based on the similarity of the document image and the claim's image, and clearly the fact that images are different might not be enough to tell if an image is fake or misleading.

Given that large annotated datasets with fake news are not widely available, some researchers opted to create fake versions of real news by replacing names and locations with tools like spaCy [Honnibal and Montani, 2017] with fake ones. An example of this approach is [Müller-Budack et al., 2020]. The authors develop a method of comparing news articles and their images by identifying the entities mentioned in a text and extracting visual features from the images and querying online databases to establish a measure of entity consistency. The interesting aspect of this approach is that it can work as an automated fact-checking tool in cases where real images are being used to spread misinformation about persons and places not present in the original picture. A drawback of this approach is that it might lack some aspects of real fake news, like the specific appealing language and visual characteristics.

A common strategy to obtain fake news data is to scrape fact-checking websites. A research paper that uses this strategy is [Yao et al., 2023]. The authors scraped from Snopes and Politifact the claims, the claim reviews and the documents and images used as evidence by the journalists to fact-check the claim. The paper then presents a model for ranking evidence, one for claim verification and another one for justification production. The text evidence retrieval is based on a cosine similarity comparison of SBERT embeddings, which are re-ranked with a BERT model pre-trained with MS MARCO Passage Ranking dataset [Nguyen et al., 2016]. The image retrieval is based on the cosine similarities of the CLIP embeddings between the claim and the image set. To carry out the claim verification, the authors generate the CLIP representations of the claim and the textual and image evidence and pass them through an attention head to generate a stance representation and use the representation to predict the entailment based on a cross-entropy objective. The claim, the evidence and the predicted label are then fed into BART [Lewis et al., 2019] to generate the justification.

2.3.6 Generative models

The recent generative large language models (LLMs) have greatly improved the benchmarks on several NLP challenges, like mathematical, commonsense, logical and multimodal reasoning [Chu et al., 2023]. They are naturally a great opportunity for automated fact-checking due to the great amount of world knowledge already built into the models, the lack of necessity of labeled data and the impressive zero-shot reasoning capabilities [Chen and Shu, 2023]. An example of how these models can be used to enhance fact-checking tools can be found in [Wu et al., 2023]. Exploring prompt-engineering, the authors use GPT-3.5 to detect incoherence between images and captions and obtain results more detailed than a usual classifier.

A concerning issue with generative models is hallucination [Rawte et al., 2023]. A way to overcome hallucinations and the lack of up-to-date knowledge of large language models has been to give the models information and context that it has not been trained with. This strategy has been called Retrieval Augmented Generation (RAG). An example of how this methodology has been applied to fact-checking can be found in [Cheung and Lam, 2023]. Given a prompt, the authors implemented a solution that queries GoogleAPI and adds the external knowledge to the instruction given to LLaMA [Touvron et al., 2023], an open-source LLM. An evaluation based on two fact-checking datasets – LIAR [Wang, 2017] and RAWFC [Yang et al., 2022b] — showed that this approach can have state-of-the-art results in fact-checking tasks. Another approach that uses LLaMa can be found at [Leite et al., 2023], but instead of augmenting the model with external knowledge, the authors use the model's reasoning ability to identify credibility signals in claims, like bias, impoliteness, sensationalism, similar to what professional journalists also do when fact-checking content.

A preliminary study [Yang et al., 2023b] of the GPT-4V capacities showed how it can have an enormous potential to advance automated fact-checking research. Just as an anecdotal evidence, when asked to identify the person in a picture (Joe Biden) and what the person was doing, the model not only identified him correctly but was also able to describe that he was at the 2023 G7 Summit delivering a speech. The ability to process and reason over medical images also seemed remarkable. It could identify medical conditions with minimal instruction. It can identify objects and their relative positions in images, explain why certain memes are funny and reason over tables and charts. To the best of our knowledge, there does not seem to be yet a direct application of GPT-4V for fact-checking, but its ability to recognize people and places could represent an unprecedented advance for misinformation detection and verification.

Dataset	Criteria	Label	Language	Topic	Source	Size
NLP4IF 2021 [Shaar et al., 2021a]	Verifiable Interesting Harmful Attention-worthy	Yes No	English Arabic Bulgarian	Covid-19	Twitter	1.3K – 4K
[Kazemi et al., 2021a]	Claim-like Statements	Yes No Probably	English Hindi Bengali Malayalam Tamil	Covid-19 Politics	WhatsApp	5K
ClaimHunter [Beltrán et al., 2021]	Check-worthy	Yes No	Spanish Catalan Galician Basque	Politics	Twitter	30K
[Dutta et al., 2022]	Verifiable Claims	Yes No	English Hindi Bengali Code-mixed	Politics	Twitter	600 – 1.4K
CheckThat 2022 [Nakov et al., 2022]	Verifiable Check-worthy Harmful Attention-worthy	Yes No	English Arabic Bulgarian Dutch Turkish	Covid-19	Twitter	4K – 6K

Table 1: Claim Detection Datasets

2.4 Datasets

2.4.1 Claim detection

Some notable datasets released for verifiable claim detection were the NLP4IF 2021 shared task data [Shaar et al., 2021a] and CheckThat 2022 data [Nakov et al., 2022]. Both datasets contain tweets related to the COVID-19 pandemic. NLP4IF 2021 includes tweets written in English, Arabic, and Bulgarian languages labeled with several misinformation labels, including the label indicating whether the tweet requires fact-checking or not, verifiable or not, and harmful or not. In addition to these three languages, the CheckThat 2022 dataset includes Dutch and Turkish tweets labeled for the verifiable claim detection and claim prioritization tasks. This dataset was expanded with more languages and data via several stages [Alam et al., 2021a, Shaar et al., 2021b, Nakov et al., 2021], and the final version was released in 2022 [Nakov et al., 2022]. Following these, several other claim detection datasets were released, focusing on topics including COVID-19 [Kazemi et al., 2021a] and politics [Kazemi et al., 2021a, Dutta et al., 2022]. Table 1 summarizes the existing claim detection datasets.

2.4.2 Claim matching

As mentioned previously, claim-matching tasks are treated with various objectives in the literature, and a wide range of datasets serving these objectives are available. This includes matching two tweets [Kazemi et al., 2021a], matching tweets with a report [Kazemi et al., 2022], and also matching a verified claim with tweets or social media posts [Shaar et al., 2021b, Nielsen and McConville, 2022, Pikuliak et al., 2023]. Table 2.4.2 summarizes the existing multilingual claim-matching datasets.

Dataset	Label	Language	Topic	Source	Size
[Kazemi et al., 2021a]	Claim Pairs	English Hindi Bengali Malayalam Tamil	Covid-19 Politics	WhatsApp	300 - 650 Pairs
[Shaar et al., 2021b]	Claim-Tweets Pairs	English Arabic	Multitopic	Twitter Snopes AraFact [Ali et al., 2021] ClaimsKG [Tchechmedjiev et al., 2019]	2.5K Pairs
[Kazemi et al., 2022]	Claim-Report Pairs	English Hindi Spanish Portuguese	Multitopic	Twitter Google Fact Check Tools	400 - 4.8K Pairs
MuMiN [Nielsen and McConville, 2022]	Claim-Tweet Pairs	41 Languages	Multitopic	Twitter Google Fact Check Tools	13K Claims 21M Tweets
MultiClaim [Pikuliak et al., 2023]	Claim-Post Pairs	27 Languages	Multitopic	Face book, Twitter Instragam Google Fact Check Tools	31K Pairs
MMTweets [Singh et al., 2023]	Claim-Misinformation Tweet Pairs	English Hindi Spanish Portuguese	Covid-19	Twitter Fact-checking Organizations	1.6K Pairs

Table 2: Claim matching datasets.

2.5 Research directions

Automated fact-checking has drawn considerable attention over the past few decades, and various interesting research directions including multilingual, and multimodal fact-checking-checking, and the adoption of generative pre-trained models and transformer families are explored in the literature. However, the results are still far from the human performance due to the profoundly challenging nature of the issue. We suggest following future research directions:

- **Development of comprehensive datasets**: One of the key aspects hindering the progress of fact-checking research is the unavailability of training data. Especially, comprehensive multi-topic claim detection datasets, verifiable claim type detection datasets, and explainable claim detection are yet to be developed for the research progress even in monolingual settings.

- **Generalized solutions**: 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.

- **Time-aware fact-checking**: Both the true value and the requirement to determine the verifiability, priority, 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.

- **Multi-modal fact-checking**: As mentioned in the literature review, there has been recent efforts to deal with multi-modal misinformation. However, the results are still limited. Current models struggle to identify misinformation disguised as humor, such as in memes, interpret information present in more than one modality or find information necessary to verify a multi-modal claim.

In this chapter, we discussed the main issues about fact checking in the general domain. In the following section, we target disinformation within the context of the health domain, which correlates with fact checking issues.

3 Health disinformation

Fighting the transmission of false information is one of the most urgent problems in the field of information access. Current approaches for detecting misinformation make use of fact-checking techniques, statistical techniques, linguistic features, or neural network models [Kumari et al., 2022]. However, the threat of misleading information appears to be intensifying with the introduction of remarkably creative language models.

3.1 Context

These days, disinformation on the internet and social media is a serious issue that affects society, the economy, and politics greatly, leading to unfavourable outcomes like divisiveness, violence, and election meddling [Scheufele and Krause, 2019]. This was especially evident during the pandemic of 2020 when a lot of information regarding COVID-19 and its therapies was of low-quality or dubious sources [Islam et al., 2020, Pennycook et al., 2020]. On social media, several falsehoods concerning COVID-19 medications and the virus's virality – which in certain situations targets underprivileged populations [Brennen et al., 2020] – have been spreading. For instance, in Iran, many people drank fake alcohol that contained poisonous methanol due to a misconception that claimed alcohol killed the coronavirus. Around 1,000 people needed to be hospitalized, nearly 300 people passed away, and many more suffered irreversible visual loss as a result of this rumour [Kumari et al., 2022]. The myth that hydroxychloroquine (HCQ) and chloroquine (CQ) can treat coronavirus has also been propagated. Therefore, it is essential to identify potentially harmful health-related material as soon as possible [Vigdor, 2020]. Due to low literacy rates, limited technological exposure, and limited comprehension, the problem appears to be particularly severe in developing nations. However, as more people gain access to inexpensive internet, they become more likely to believe and act on false information.

Some websites and individuals intentionally generate harmful content, as exemplified by instances such as pro-eating disorders (pro-ana) sites. These platforms openly share materials that advocate life-threatening behaviours. For instance, in a comprehensive study, Borzekowski et al. [Borzekowski et al., 2010] revealed that 85% of pro-ana sites produce content promoting "thinspiration" and explicit suggestions for engaging in eating-disordered actions. The information disseminated by these detrimental sources poses a risk to over 700 million individuals, predominantly women, who are seeking online health information for eating disorders [Rouleau and von Ranson, 2011].

3.2 Definitions

Navigating the complex realm of online health information requires a foundational understanding of key terms. In this section, we elucidate pivotal definitions surrounding disinformation and health disinformation, laying the groundwork for a comprehensive exploration of challenges and solutions in the context of consumer health search.

Misinformation: Misinformation refers to the unintentional spread of inaccurate information shared in good faith by those unaware that they are passing on falsehoods [Nations, 2023].

Disinformation: Disinformation is information that is not only inaccurate but is also intended to deceive and is spread in order to inflict harm [Nations, 2023].

Fact-checking: Fact-checking is the systematic process of verifying the accuracy and reliability of factual claims, statements, or information presented in various forms of media. This practice involves thorough investigation, cross-referencing with credible sources, and assessing the evidence supporting or refuting a given assertion. Fact-checking aims to provide the public with accurate and unbiased information, helping to counter the spread of misinformation and contribute to informed decision-making [Guo et al., 2022b]. More details about this particular topic can be found in section 2.

Health disinformation: Health disinformation refers to misleading or false information specifically related to health and medical topics. It goes beyond mere inaccuracy, encompassing content deliberately crafted to deceive and disseminated with the intent to cause harm within the domain of public health and individual well-being [Schlicht et al., 2023].

3.3 Approaches

In the ever-expanding landscape of online health information, the quest for reliable guidance prompts the use of Consumer Health Search, demanding robust retrieval algorithms. This exploration delves into the challenges of disinformation detection and credibility estimation, emphasizing the role of language analysis and advanced linguistic models in discerning reliable health-related information from potential misinformation. The subsequent analysis of user responses to false information underscores the multifaceted factors influencing the impact of health-related misinformation, employing diverse approaches, including user studies, expert knowledge, web mining, and personalized models.

3.3.1 Sources

Web search is extensively employed for seeking online guidance, particularly in the realm of medical advice [Pew, 2011]. This category of web search is commonly known as Consumer Health Search [Jimmy et al., 2018]. Effectively accessing health-related information necessitates retrieval algorithms that can prioritize trustworthy documents while excluding unreliable ones. To achieve this goal, various elements, including features for matching queries with documents, estimating passage relevance, evaluating reliability, and employing suitable recommendation models, must be integrated.

The advent of digital media has enhanced the accessibility of information [Reuters, 2021], although the information offered may lack reliability [Abualsaud, 2019], precision [Eysenbach, 2002], or quality [Rieh, 2001]. Pogacar and associates demonstrated that presenting low-quality search results can lead individuals to make erroneous decisions [Weizenbaum, 1966a]. People are susceptible to the influence of search engine outcomes, and exposure to inaccurate information can have detrimental effects. Misinformation disseminated through online channels can be particularly harmful, especially in the initial phases of information dissemination when there is limited awareness of the reliability or accuracy of a specific claim.

Conversely, language serves as a tool for distinguishing trustworthy information from unreliable sources [Matsumoto et al., 2014, **?**]. For instance, the inclusion of technical terminology or formal structures is often linked to higher quality and, frequently, more dependable content. Various machine learning technologies have been employed to leverage the linguistic characteristics of text [Adhikari et al., 2019, Sondhi et al., 2012, Fernández-Pichel et al., 2021b, Fernández-Pichel et al., 2021a].

Consumers exhibit diversity in their information needs, encompassing various patterns when seeking medical information online. Cartright et al. [Cartright et al., 2011] conducted a study involving over 660,000 users, revealing that while a majority focus on symptoms or symptom causes, a significant number are searching for remedies (treatments), symptom-remedy pairs, simple causes, or causes and remedies. Additionally, user sessions can be categorized as evidence-directed, hypothesis-directed with a focus on causes, or hypothesis-directed with a focus on remedies. Furthermore, users vary in their levels of knowledge and expertise. Palotti et al. [Palotti et al., 2015] demonstrated that greater expertise correlates with increased user persistence and more complex information needs. This emphasizes the necessity of designing advanced models to cater to the information requirements of more knowledgeable users, highlighting the importance of tailoring information access to the specific circumstances of individual users.

It is crucial to comprehend the diverse ways in which various users respond to false information and harmful recommendations. The likelihood of experiencing significant consequences following exposure to health-related misinformation is contingent on various user factors [Ellery et al., 2008, Grant et al., 2007].

3.3.2 Disinformation detection and credibility estimation

The extensive volume of interactions and publications accessible on the Internet and social networks allows for comprehensive analyses of content curation focused on consumer health search. People commonly utilize search engines to make health-related queries or share health-related recommendations on blogs, social media posts, or other types of websites. However, tackling the analysis of online health-related misinformation presents challenges in several areas: information filtering and search (to identify relevant online content for consumer health search), linguistic analysis of text and psycho-linguistics, evaluating the quality and reputation of content (e.g., recommending reputable information for individuals with specific disorders), and the efficient processing of large-scale data (requiring scalable distributed computing methods that operate in real-time).

Intelligently integrating various forms of evidence facilitates the differentiation between truthful and untruthful content, as well as assessing the credibility of the sources, where language analysis plays a pivotal role. Language has demonstrated its potential in discerning reliable from unreliable information [Matsumoto et al., 2014, Mukherjee and Weikum, 2015]. For instance, the use of technical terms or formal constructs is often associated with higher quality and more reliable content. Numerous machine learning technologies, such as those leveraging linguistic properties of text [Adhikari et al., 2019, Sondhi et al., 2012, Fernández-Pichel et al., 2021b, Fernández-Pichel et al., 2021a], have been applied, and language-based features might help enhance the detection of health-related misinformation.

Advanced linguistic models, particularly those based on recent deep learning architectures, such as transformers [Yates et al., 2021] have demonstrated good results. But a variety of deep learning models, encompassing feed-forward networks, RNN-based models, CNN-based models, capsule networks, attention-based solutions, memory-augmented networks, graph neural networks, siamese neural networks, and hybrid models [Minaee et al., 2021] were applied to this field. Some of these models have pre-trained versions constructed from extensive corpora, which can be fine-tuned for specific tasks. The integration or fusion of multiple types of evidence is a crucial aspect of credibility estimation. In this regard, advanced score distribution models [Arampatzis and Robertson, 2010] and learning-to-rank techniques [Liu, 2011] have been used.

3.3.3 Analysis of results

To understand how different users react to false information and damaging recommendations we need to evaluate and analyze the results. The risk of suffering severe consequences after being exposed to health-related misinformation depends on a number of user factors [Ellery et al., 2008, Grant et al., 2007]. These factors include social dimensions, personality dimensions and psychological dimensions [Baumgartner and Hartmann, 2011]. To that end, user studies, using expert knowledge, web mining and personalized models and their explainability are some of the used approaches.

3.4 Datasets

Several international challenges have generated openly available datasets focused on consumer health search [Goeuriot et al., 2021], health misinformation [Clarke et al., 2020, Soboroff, 2021], and precision medicine [Roberts et al., 2020]. This provides an opportunity to examine common health-related information needs, analyze search results for various health-related queries, and utilize ground truth data across different dimensions, such as relevance, correctness, and credibility. These shared-task datasets also offer precision medicine data, enabling the identification of treatments tailored to an individual patient's unique characteristics.

Moreover, publicly accessible online sources, including open forums and social networks, serve as platforms where individuals openly discuss health-related issues, sharing medical concerns and worries. This setting allows for a psychological examination, including personality traits, and facilitates the development of models to understand how people react to accurate/inaccurate information and credible/non-credible information. Additionally, established clinical criteria from Diagnostic Manuals for medical disorders, such as mood changes and loss of interest in the case of depression, prove valuable when studying the evolution of language use and expressed concerns over time, including typical health-related queries.

Moreover, there are suggested cost-effective pooling approaches to construct unbiased Information Retrieval (IR) benchmarks that are reusable [Otero et al., 2021, Losada et al., 2016, Losada et al., 2017, Losada et al., 2018, Lipani et al., 2021]. While the previous collections were extensive, they were relatively uniform. In the context of health misinformation, collections are often diverse, and there is a pressing need to develop new collections with limited resources. Consequently, adjusting pooling methods to accommodate the specific characteristics poses a challenge.

Initiatives like TREC [Clarke et al., 2020, Soboroff, 2021] encourage the exploration of retrieval techniques that prioritize accurate and trustworthy information for tasks related to health-related decision-making, discouraging the prevalence of misinformation. One of the tasks is "Participants devise search technologies that promote credible and correct information over incorrect information, with the assumption that correct information can better lead people to make correct decisions". For this task, they have created a collection that contains around 1B English documents.

Moreover, there is a review of publicly available datasets for health misinformation detection [Ni et al., 2023], where the authors gathered works that have emerged from 2020. The authors claim that most of the datasets are based on fact-checkable websites, while only a few are annotated by experts. Table 3 summarises the available datasets gathered in [Ni et al., 2023].

3.5 Future directions

Some future directions are expected from health disinformation to make improvements in the field. This area is novel in itself and, thus, future works should build experimental foundations, like the creation of test collections, definitions of evaluation metrics, etc. We highlight some open research problems:

- It has been shown that it is very difficult to retrieve many helpful documents (relevant + correct + credible) while maintaining at a low level the retrieval of harmful documents (e.g. relevant but incorrect and, in some cases, look credible to the user) [Clarke et al., 2020, Soboroff, 2021]. The inherent complexity of the task lies in the need to intelligently integrate multiple signals, including document retrieval, passage relevance, and reliability estimators. It is crucial to undertake further research to explore the specific type of evidence required and develop suitable methods for combining these signals effectively within the realm of consumer health search. Score distribution models, as demonstrated by successful cases such as [Parapar et al., 2019, Losada et al., 2018, Parapar et al., 2014], have proven effective in tasks involving thresholding and fusion. The future steps involve incorporating Score Distribution models and leveraging learning to rank solutions to enhance the capability of combining multiple sources of evidence, aiming to improve the overall performance and reliability of information retrieval in consumer health search.

- State-of-the-art search methods, represented by language models, have demonstrated success, particularly in recommendation systems. These models yield favourable results when applied to non-textual data, such as product evaluations (ratings) or implicit feedback, as evidenced by studies like [Valcarce et al., 2016b, Valcarce et al., 2016a, Valcarce et al., 2016c, Valcarce et al., 2015a, Valcarce et al., 2015b, Parapar et al., 2013]. It is important to note that available information extends beyond text alone. Some web sources offer additional evidence based on networks and usage, encompassing platform use, interactions, and contacts. Future research endeavours should aim to achieve high search quality by considering multiple types of evidence, including contextual and temporal information. Consequently, language models emerge as promising candidates to address this multifaceted challenge.

- Cost-effective pooling strategies that build unbiased IR benchmarks that are reusable have been proposed in existing literature [Otero et al., 2021, Losada et al., 2016, Losada et al., 2016, Losada et al., 2017]. Those collections were large but homogeneous. In the field of health misinformation, the collections tend to be heterogeneous and, furthermore, it is essential to create new collections with a low budget. Therefore, it is a challenge to adapt pooling to the peculiarities of the health disinformation domain.

- Enhancements in linguistic resources for health consumer search and the detection of health misinformation are imperative. There is a need to develop novel models for discourse analysis, which includes refining existing medical terminologies (e.g., psychological disorders [Losada and Gamallo, 2018]) and recognizing discourse patterns and coherence. The efficacy of these techniques in analyzing specific medical concerns, such as psychological disorders [Hong et al., 2015, Iter et al., 2018, Ash et al., 2006], has already been demonstrated. Additionally, it is critical to automatically identify information needs related to health. Preliminary proposals have been made on automating the identification of queries associated with food and nutrition [Losada et al., 2021].

In conclusion, the future landscape of addressing health disinformation presents both challenges and opportunities. To forge ahead, it is essential to enhance linguistic resources dedicated to health consumer search and misinformation detection, including the development of advanced discourse analysis models. Refining medical terminologies, recognizing discourse patterns, and ensuring coherence are pivotal aspects of this endeavor. The proven potential of these techniques in analyzing specific health concerns underscores their importance. Moreover, the automatic identification of health-related information needs, as exemplified in preliminary proposals for queries associated with food and nutrition, signifies a promising avenue for future research. As we navigate the complex realm of health information, interdisciplinary collaboration and innovative methodologies will be key in our ongoing efforts to foster accurate, accessible, and reliable health information for all. We talked about different websites that intentionally generate harmful content, like pro-eating disorders, emphasising that, although disinformation and hate speech are distinct phenomena, they are related. The next section will dive deeper into hate speech.

Ref.	Topic	Lang.	Data type	Labels	Construction strategy	Size	Evaluation results	Additional information
[Srba et al., 2022]	Medical information	EN	Multimodal	False, Mostly False, Mixture, Mostly True, and True	Based on fact-checking websites	10k+	-	Social engagement, Explanations
[Paka et al., 2021]	COVID-19	EN	Multimodal	Fake, Genuine	Based on fact-checking websites	10k+	F1 = 0.953 (Cross-SEAN)	Social engagement
[Hayawi et al., 2022]	Vaccine; COVID-19	EN	Text	Misinformation, Not misinformation	Expert labeling	5k-10k	F1 = 0.98 (BERT)	-
[Hu et al., 2022]	General health	CN	Text	Supported, Refuted, Not enough information	Based on fact-checking websites	1k-5k	F1, = 79.84 (Graph-Based Model)	Evidence for claims
[Nabożny et al., 2021]	Medical information	EN	Text	Credible, Not credible, Neutral	Expert labeling	5k-10k	-	Reasons for non-credibility (Polish)
[Cui et al., 2020]	General health (Cancer and Diabetes)	EN	Text	Misinformation, fact	Based on fact-checking websites	1k-5k	F1 = 0.8474 on diabetes data; F1 = 0.9309 cancer data	-
[Micallef et al., 2020]	COVID-19	EN	Text	Misinformation, Counter-Misinformation (Professional fact-check tweets/Concerned citizen tweets), Irrelevant	Expert labeling	10k+	F1 = 0.802 on 5G dataset (LR), F1 = 0.709 on fake cures (LR)	-
[Haouari et al., 2021]	COVID-19	Arabic	Multimodal	True, False, Other	Based on fact-checking websites	1k-5k	F1 = 0.762 (MARBERT)	Social engagement
[Cui and Lee, 2020]	COVID-19	EN	Multimodal	True, Fake	Based on fact-checking websites	500-1k	F1 = 0.5814 (dEFEND)	Social engagement
[Kolluri et al., 2022]	Monkeypox	EN	Text	Factual, False	Based on fact-checking websites	100-	ACC = 0.96 (BERT)	-
[Mohr et al., 2022]	COVID-19	EN	Text	Supports, Refutes and Not enough information (NEI)	Expert labeling	100-	F1 = 0.69 (MLP-Evidence)	Evidence for claims; Medical named entity recognition
[Kotonya and Toni, 2020b]	General health	EN	Text	True, False, Mixture, and Unproven	Based on fact-checking websites	1k-5k	F1 = 0.705 (SCIBERT)	Explanations
[Du et al., 2021]	COVID-19	CN	Text	Fake, Real	Based on fact-checking websites	100-	F1 = 0.73 (CrossFake)	-
[Endo et al., 2022]	COVID-19	PT-BR	Text	Rumors, Non-rumors	Based on fact-checking websites	1k-5k	F1 = 0.94 (bi-GRU)	-
[Khan et al., 2022]	COVID-19	EN	Text	Fake, True	-	500-1k	F1 = 0.88 (Random forest)	-
[Mahlous and Al-Laith, 2021]	COVID-19	Arabic	Text	Fake, Genuine	Expert labeling	500-1k	F1 = 0.87 (LR)	-
[Bonet-Jover et al., 2021]	General health	ES	Text	True, False, Unknown	Based on fact-checking websites	100-	F1 = 0.74	5W1H annotation (Who, What, When, Where, Why and How)
[Li et al., 2020]	COVID-19	Multilingual	Multimodal	Fake, Real	1k-5k	-	Social engagement; explanations	
[Dai et al., 2020]	General health	EN	Multimodal	Real, Fake	Based on fact-checking websites	500-1k	F1 = 0.756 on HealthStory, F1 = 0.802 on HealthRelease	Social engagement; explanations
[Zhou et al., 2020]	COVID-19	EN	Multimodal	Reliable, Unreliable	Credibility level of news publishers	500-1k	-	Social engagement
[Alam et al., 2021b]	COVID-19	Multilingual	Multimodal	Whether (binary)/to what extent (5 classes) the tweet appears to contain false information	Expert labeling	1k-5k	F1 = 0.92 (English RoBERTa), F1 = 0.84 (Arabic AraBERT), F1 = 0.95 (Bulgarian XLM-RoBERTa), F1 = 0.87 (Dutch FastText)	-

Table 3: Health misinformation detection datasets gathered by Ni et al.

4 Hate speech detection

4.1 Introduction

There has been a significant growth in abusive content in social media that can have adverse effects on various fringe groups based on ethnicity, color, religious beliefs. Hate Speech as defined in [Davidson et al., 2017] is **speech that targets disadvantaged groups in a manner that is potentially harmful to them**. This chapter will give an overview about the methods and tools that have been developed to combat hate-speech in social media sites and the datasets that have enabled training of models. It will also look at directions that can be taken to build robust models incorporating recent breakthroughs in Natural Language Processing.

4.2 Approaches

There has been a plethora of approaches that have been proposed to tackle hate-speech and different constructs that come under the purview of hate speech. [Schmidt and Wiegand, 2017, Fortuna and Nunes, 2018] do a good job at summarizing the approaches that have been used in hate-speech detection, from collecting data from various social networks using keywords and building classifiers that can aid in the detection of hate-speech. With the introduction of Transformer model [Vaswani et al., 2017] and its variants – BERT [Devlin et al., 2019] and RoBERTa [Liu et al., 2019b] – efforts have been made to build robust classifiers for Hate Speech detection. HateBERT [Caselli et al., 2021] is one such which was trained on RAL-E (Reddit Abusive Language English Dataset). [Antypas and Camacho-Collados, 2023] did a cross dataset evaluation on different English language datasets in an attempt to build robust Hate-Speech classifier. They designed a 2-step method of combining datasets and evaluating it on another held-out dataset. There have been attempts to build models for multilingual hate-speech as well. Attempts have also been made in the detection of multilingual hate-speech. [Bigoulaeva et al., 2021] train a family of Convolutional Neural Network and BiLSTM classifiers on English Language datasets and then use transfer learning on German Hate-Speech datasets (StormFront [de Gibert Bonet et al., 2022] and GermEval [Risch et al., 2018]). [Röttger et al., 2022] proposed a 2-step approach for fine-tuning a multilingual RoBERTa-base model previously fine-tuned on Twitter data (XLM-T) [Barbieri et al., 2022] on English language data; then they used an Active Learning approach to label and fine-tune in the target language. Inspired by this approach, [Goldzycher et al., 2023] proposed an approach which used pre-defined hypothesis to check the label (they used the label ideas from MNLI dataset [Williams et al., 2018]. Despite success in applying new methods to solve the hate-speech detection problem, the problem of bias which ML systems are prone to [Bolukbasi et al., 2016, Garg et al., 2018, Buolamwini and Gebru, 2018] remains and this has led to investigation into the data that these models are trained on, especially when it comes to models trained on hate-speech datasets. To understand this phenomenon [Davidson et al., 2019] has taken a closer look at popular hate-speech datasets, covering different kinds of constructs related to hate-speech, namely, racism, sexism, antisemitism, Islamophobia, sarcasm, and they found racial biases in the models trained on the datasets. Data annotation while curating datasets for task specific applications is an important step, and it has been investigated in [Davani et al., 2022]. In this paper, the authors have investigated the social stereotypes that influence annotation behavior, performance of ML classifiers on annotated hate-speech datasets. The propensity of ML models to capture biases needs to be taken into account when curating and annotating datasets. [Madukwe et al., 2020] has done a detailed critical analysis of all the popular datasets that have been used in detection of hate-speech and has pointed out flaws like class-imbalance biases that exist while labeling the datasets. The following section will give a broad overview about the datasets that are available for hate-speech detection.

4.3 Datasets

One of the challenges faced in hate speech detection is the lack of standardized datasets [ElSherief et al., 2018, Poletto et al., 2020, Toraman et al., 2022], evaluation metrics [Röttger et al., 2021], and benchmark models [Poletto et al., 2020].

Over the past few years, numerous efforts have been made to create datasets for hate speech analysis [Davidson et al., 2017, Golbeck et al., 2017, Founta et al., 2018]. The community has a widely recognized list that aims to collect all available hate speech corpora: hatespeechdata[3]. While this repository is a valuable source, it only provides a list of various dataset publications and their links. Nevertheless, the website also studied 63 datasets, of which 25 are in English, focusing on the best practices for creating datasets for detecting hate speech [Vidgen and Derczynski, 2020]. Studies such as [Poletto et al., 2020] have concentrated on studying all available corpora resources to detect hate speech. This has resulted in a comprehensive survey that highlights the numerous benchmark datasets available for evaluating abusive language.

There is a recent study [Piot et al., 2024] which analyses over 60 English hate speech detection datasets in order to release a meta-collection of more than 1M instances.

We have found datasets in many languages, including Albanian, Arabic, Chinese, Danish, English, French, German, Greek, Hindi, Italian, Latvian, Portuguese, Russian, Spanish, and Turkish, among others. The majority of datasets are textual, but we can find some that are based on image content. The datasets come from a very diverse source of social media networks. Including, but not only: Twitter, Facebook, Reddit, Stormfront, Gab, Whisper, Wikipedia, Civil Comments, YouTube, and BitChute. The strategies employed for the datasets creation, overall, include: (1) utilizing lexicons, keywords, hashtags, and phrase structures, and (2) randomly sampling from sites with a likelihood of containing hate content. In terms of conceptualization, the majority of works adopt a binary strategy. However, a vast of them take a multiclass approach, distinguishing between abusive, hate, offensive, or normal speech, among other terms. A few efforts explore a probabilistic approach, assigning a numerical value between 0 and 1 to gauge the degree of hatefulness in a comment. Additionally, some studies opt for a multiclass and multilabel approach and others go a step further, attempting to extract the specific span that contains hate speech within a larger sentence.

Tables 4, 5, 6, 7, 8 and 9 summarize the available datasets found regarding hate speech. We describe them by showing the language, the task, or purpose the dataset was created, the size (number of entries), the annotation level (post, with context, etc.), the source (from which platform was sourced, e.g., Twitter, YouTube, Facebook, etc.), the medium, specifying if is textual data, images, etc. and the reference of the dataset.

[3] https://hatespeechdata.com/

Lang	Task	Size	Annot	Source	Medium	Reference
SQ	Detecting Abusive Albanian	11,874	Posts	Instagram, Youtube	Text	Nurce et al., 2021
AR	Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language	6,603	Posts	Twitter	Text	Mulki and Ghanem, 2021
AR	Are They our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere	6,136	Posts	Twitter	Text	Albadi et al., 2018
AR	Multilingual and Multi-Aspect Hate Speech Analysis	3,353	Posts	Twitter	Text	Ousidhoum et al., 2019
AR	L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language	5,846	Posts	Twitter	Text	Mulki et al., 2019
AR	Abusive Language Detection on Arabic Social Media	1,100	Posts	Twitter	Text	Mubarak et al., 2017
AR	Abusive Language Detection on Arabic Social Media	32,000	Posts	Al Jazeera	Text	Mubarak et al., 2017
AR	Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic	15,050	Posts	YouTube	Text	Alakrot et al., 2018
BN	Hate Speech Detection in the Bengali language: A Dataset and its Baseline Evaluation	30,000	Posts	Youtube, Facebook	Text	Romim et al., 2021
ZH	SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection	8,969	Posts	Sina Weibo	Text	Jiang et al., 2022
HR	CoRAL: a Context-aware Croatian Abusive Language Dataset	2,240	Posts	Newspaper comments	Text	Shekhar et al., 2022
HR	Datasets of Slovene and Croatian Moderated News Comments	17M	Posts	24sata website	Text	Ljubešić et al., 2018
HR	Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian	21M	Posts	Newspaper comments	Text	Shekhar et al., 2020

Table 4: Hate Speech datasets in Albanian, Arabic, Bengali, Chinese and Croatian.

Lang	Task	Size	Annot	Source	Medium	Reference
DA	Offensive Language and Hate Speech Detection for Danish	3,600	Posts	Twitter, Reddit, newspaper comments	Text	Sigurbergsson and Derczynski, 2019
DA	BAJER: Misogyny in Danish	27,900	Social media post / comment	Twitter, Facebook, Reddit	Text	Zeinert and Derczynski, 2021
NL	The Dutch Abusive Language Corpus v1.0 (DALC v1.0)	8,156	Tweets	Twitter	Text	Caselli et al., 2021
FR	CONAN - COunter NArratives	1,719	Posts	Synthetic, Facebook	Text	Chung et al. (2019)
FR	Multilingual and Multi-Aspect Hate Speech Analysis	4,014	Posts	Twitter	Text	Ousidhoum et al. (2019)
FR	CyberAgressionAdo-v1	19	Messages	Role-playing games	Text	Ollagnier et al. (2022)
DE	DeTox	10,278	Comments	Twitter	Text	Demus et al. (2022)
DE	RP-Crowd	85,000	Comments	German Newspaper	Text	Assenmacher et al. (2021)
DE	Measuring the Reliability of Hate Speech Annotations	469	Posts	Twitter	Text	Ross et al. (2017)
DE	Detecting Offensive Statements Towards Foreigners	5,836	Posts	Facebook	Text	Bretschneider et al. (2017)
DE	GermEval 2018	8,541	Posts	Twitter	Text	Wiegand et al. (2018)
DE	HASOC track at FIRE 2019	4,669	Posts	Twitter, Facebook	Text	Mandl et al. (2019)
EL	Deep Learning for User Comment Moderation	1,45M	Posts	Gazetta	Text	Pavlopoulos et al. (2017)
EL	Offensive Language Identification in Greek	4,779	Posts	Twitter	Text	Pitenis et al. (2020)

Table 5: Hate Speech datasets in Danish, Dutch, Estonian, French, German and Greek.

Lang	Task	Size	Annot	Source	Medium	Reference
HI	Hostility Detection Dataset	8,192	Posts	Twitter, Facebook, WhatsApp	Text	Bhardwaj et al. (2020)
HI	Aggression-annotated Corpus	18,000	Posts	Facebook	Text	Kumar et al. (2018)
HI	Aggression-annotated Corpus	21,000	Posts	Twitter	Text	Kumar et al. (2018)
HI	Offensive Tweets in Hinglish Language	3,189	Posts	Twitter	Text	Mathur et al. (2018)
HI	Hindi-English Code-Mixed Text	4,575	Posts	Twitter	Text	Bohra et al. (2018)
HI	HASOC track at FIRE 2019	5,983	Posts	Twitter, Facebook	Text	Mandl et al. (2019)
ID	Hate Speech Detection	713	Posts	Twitter	Text	Alfina et al. (2017)
ID	Multi-Label Hate Speech Detection	13,169	Posts	Twitter	Text	Ibrohim and Budi (2019)
ID	Abusive Language Detection	2,016	Posts	Twitter	Text	Ibrohim and Budi (2018)
KO	Toxic Speech Detection	9,381	Comments	NAVER	Text	Moon et al. (2020)
LV	User Comment Dataset	12M	Posts	Newspaper comments	Text	Pollak et al. (2021)
IT	Hate Speech against Immigrants	1,827	Posts	Twitter	Text	Sanguinetti et al. (2018)
IT	EVALITA 2018 Hate Speech Task	8,000	Posts	Facebook, Twitter	Text	Bosco et al. (2018)
IT	Automatic Misogyny Identification (AMI)	6,000	Posts	Twitter	Text	Fersini et al. (2020)
IT	CONAN - COunter NArratives	1,071	Posts	Synthetic, Facebook	Text	Chung et al. (2019)
IT	WhatsApp Dataset (Pre-teen Cyberbullying)	14,600	Chats	WhatsApp	Text	Sprugnoli et al. (2018)

Table 6: Hate Speech datasets in Hindi, Indonesian, Korean, Latvian and Italian.

Lang	Task	Size	Annot	Source	Medium	Reference
PL	PolEval 2019 Cyberbullying Detection	10,041	Posts	Twitter	Text	Ogrodniczuk and Kobyliński (2019)
PT	Toxic Language Dataset (ToLD-Br)	21,000	Posts	Twitter	Text	Leite et al. (2020)
PT	Hierarchically-Labeled Hate Speech	3,059	Posts	Twitter	Text	Fortuna et al. (2019)
PT	Offensive Comments (Brazilian Web)	1,250	Posts	g1.globo.com	Text	de Pelle and Moreira (2017)
RU	Toxic Comment Detection	3,000	Comment	VKontakte	Text	Gorbunova (2022)
RU	Hate Speech Detection	100,000	Posts	Youtube	Text	Zueva et al. (2020)
RU	Abusive Speech Detection	2,000	Posts	Youtube	Text	Andrusyak et al. (2018)
RU	South Park Hate Speech	1,400	Sentence	TV Subtitles	Text	Saitov & Derczynski (2021)
SL	Moderated News Comments	7,6M	Posts	MMC RTV website	Text	Ljubešić et al. (2018)
ES	Aggressiveness Analysis	11,000	Posts	Twitter	Text	Alvarez-Carmona et al. (2018)
ES	Misogyny Identification	4,138	Posts	Twitter	Text	Fersini et al. (2018)
ES	Hate Speech Detection (SemEval-2019)	6,600	Posts	Twitter	Text	Basile et al. (2019)
TR	Hate Speech Detection	100,000	Posts	Twitter	Text, Image	Toraman et al. (2022)
TR	Offensive Language Corpus	36,232	Posts	Twitter	Text	Çöltekin (2020)
UK	Abusive Speech Detection	2,000	Posts	Youtube	Text	Andrusyak et al. (2018)
UR	Hate Speech Detection	10,041	Posts	Twitter	Text	Rizwan et al. (2020)

Table 7: Hate Speech datasets in Polish, Portuguese, Russian, Slovene, Spanish, Turkish, Ukrainian and Urdu.

Lang	Task	Size	Annot	Source	Medium	Reference
EN	Pinpointing Fine-Grained Relationships between Hateful Tweets and Replies	5,652	Tweets	Twitter	Text	Albanyan et al. 2022
EN	Large-Scale Hate Speech Detection with Cross-Domain Transfer	100,000	Posts	Twitter	Text and Image	Toraman et al. 2022
EN	ConvAbuse	4,185	Utterance with context	Facebook Messen-ger, Chatbots	Text	Curry et al. 2021
EN	Measuring Hate Speech	39,565	Posts	Twitter, Reddit, YouTube	Text	Kennedy et al. 2020
EN	Learning From the Worst (Dynamically generated hate speech dataset)	41,255	Posts	Synthetic	Text	Vidgen et al. 2021
EN	The 'Call me sexist, but' sexism dataset	6,325	Tweets, survey items	Twitter, Social Psychology scales	Text	Samory et al. 2021
EN	Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection	3,000	Posts	Twitter	Text	Grimminger and Klinger 2021
EN	AbuseEval v1.0	14,100	Tweets	Twitter	Text	Caselli et al. 2020
EN	Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media	1,675	Words	Twitter	Text	Pamungkas et al. 2020
EN	Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text	743	Posts	Kaggle, Reddit, Facebook, Twitter, Instagram	Text, Images, Memes	Suryawanshi et al. 2020
EN	Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate	5,912	Post	Synthetic	Text with emoji	Kirk et al. 2021
EN	HateCheck: Functional Tests for Hate Speech Detection Models	3,728	Post	Synthetic	Text	Röttger et al. 2020
EN	Semeval-2021 Task 5: Toxic Spans Detection	10,629	Posts	Civil Comments	Text	Pavlopoulos et al. 2021

Table 8: Hate Speech datasets in English (I).

Lang	Task	Size	Annot	Source	Medium	Reference
EN	ToxiSpanSE	19,651	Code Reviews	Software	Text	Sarker et al. 2023
EN	Human-in-the-Loop for Data Collection	5,003	Posts	Semi-synthetic	Text	Fanton et al. 2021
EN	HateXplain	20,148	Words, phrases, posts	Twitter, Gab	Text	Mathew et al. 2021
EN	ALONE	688	Post	Twitter	Multimodal	Wijesiriwardene et al. 2020
EN	Online Slur Usage	39,811	Posts	Reddit	Text	Kurrek et al. (2020)
EN	Offensive Content (MultiOFF)	743	Posts	Social Media	Text, Images	Suryawanshi et al. (2020)
EN	Offensive Posts in Social Media	14,100	Posts	Twitter	Text	Zampieri et al. (2019)
EN	Toxicity and Bias	1,8M	Comment, Posts	Civil Comments	Text	Borkan et al. (2019)
EN	Contextual Abuse	25,000	Threads	Reddit	Text	Vidgen et al. (2021)
EN	Hate Speech Detection	24,802	Posts	Twitter	Text	Davidson et al. (2017)
EN	Hate Speech from White Supremacy Forum	9,916	Sentence	Stormfront	Text	de Gibert et al. (2018)
EN	Hateful Symbols on Twitter	16,914	Posts	Twitter	Text	Waseem & Hovy (2016)
EN	Online Hate Speech on Fox News	1,528	Posts	Fox News	Text	Gao & Huang (2018)
EN	Gab Hate Corpus	27,665	Posts	Gab	Text	Kennedy et al. (2018)
EN	Annotator Influence on Hate Speech Detection	4,033	Posts	Twitter	Text	Waseem (2016)
EN	Compliment Become Sexist	712	Posts	Twitter	Text	Jha & Mamidi (2017)
EN	Automatic Misogyny Identification	3,977	Posts	Twitter	Text	Fersini et al. (2018)
EN	Counter Narratives (CONAN)	1,288	Posts	Synthetic, Facebook	Text	Chung et al. (2019)
EN	Detecting Hateful Users on Twitter	4,972	Users	Twitter	Text	Ribeiro et al. (2018)
EN	Gab Benchmark Dataset	33,776	Posts	Gab	Text	Wulczyn et al. (2017)
EN	Offensive Language Identification	24,802	Posts	Twitter	Text	Waseem et al. (2017)

Continued on next page

Continued from previous page

Lang	Task	Size	Annot	Source	Medium	Reference
EN	Twitter Sentiment Analysis	31,961	Posts	Twitter	Text	Ali Toosi, Jan 2019
EN	Toxicity Detection in Software Engineering	19,651	Comment	Software	Text	Sarker et al. (2023)
EN	Toxicity Detection: Does Context Really Matter? CAT-LARGE (No Context)	10,000	Post	Wikipedia Talk Pages	Text	Pavlopoulos et al. (2020)
EN	Toxicity Detection: Does Context Really Matter? CAT-LARGE (With Context)	10,000	Post	Wikipedia Talk Pages	Text	Pavlopoulos et al. (2020)
EN	Anatomy of Online Hate	5,143	Comment	YouTube, Facebook	Text	Salminen et al. (2018)
EN	Automating News Comment Moderation	31.5M	Posts	Newspaper comments	Text	Shekhar et al. (2020)
EN	HateXplain	20,148	Posts	Twitter, Gab	Text	Mathew et al. (2020)

Table 9: Hate Speech datasets in English (II).

4.4 Future direction

Future directions in the field should consider different aspects:

- **Adopting a universal definition of hate speech.** Establishing a universally agreed-upon definition of hate speech is crucial. Currently, there might be variations in how hate speech is defined across different regions, cultures, or platforms. A standardized definition can provide clarity and consistency in identifying and addressing hate speech globally.
- **Creating more datasets and trying to avoid introducing bias during their creation.** The availability of diverse and comprehensive datasets is essential for training effective hate speech detection models. However, the process of creating these datasets should be carefully managed to avoid introducing bias. Bias can arise if data collection strategies disproportionately focus on specific keywords, phrases, or lexicons, leading to a skewed representation of hate speech. Efforts should be made to ensure that the datasets are balanced and reflective of the broader range of hate speech expressions.
- **Considering the context while annotating and training data.** Hate speech often depends on the context in which it is used. Annotating and training data with an understanding of the context surrounding a particular statement can enhance the accuracy of hate speech detection models. This approach recognizes that certain expressions may be considered offensive in one context but not in another.
- **Developing new classification and identification methods.** Continuous improvement of classification and identification methods is necessary to keep pace with the evolving nature of hate speech. This could involve exploring advanced machine learning techniques, NLP

models, or incorporating contextual information to enhance the precision and recall of hate speech detection algorithms

- **Focus on low-resource languages.** Efforts in hate speech detection have predominantly centred around widely spoken languages like English. However, numerous languages are underrepresented in research and development efforts. To address this gap, there should be a concerted focus on developing hate speech detection models for low-resource languages, ensuring that the benefits of such technology are accessible across linguistic diversity

In this chapter, we discussed hate speech detection, especially the approaches that have been taken so far in solving this problem and the datasets that have been curated and collected to make aid in developing solutions. Despite the otherwise remarkable progress that has been made, little has been done to see how classifiers can be made more robust when it comes to data quality, as well as how well we can curate data from multiple data sources. Despite such open problems, hate-speech is still an exciting area of research which needs careful investigation when building solutions. Hate-speech have links with political hyperpartisanship, which will be the topic of the following section 5. For example, [Vasist et al., 2023, Lorenz-Spreen et al., 2023] explore this relationship in great detail regarding political communication in social media. Hate-speech also has links with social media bots

5 Hyperpartisan detection

5.1 Context

This section delves into the concept of hyperpartisan news detection, providing readers with a comprehensive understanding of this emerging phenomenon. The section unfolds as follows: We commence by defining hyperpartisan news detection and establishing a clear understanding of this crucial concept. Next, we shed light on the hazardous consequences of exposure to hyperpartisan news, particularly emphasizing the erosion of democratic cohesion. Successively, we introduce the legal safeguards that governments are implementing to mitigate the spread of disinformation. After that, we explore the detection techniques employed by online publishers to label datasets, providing insights into their practices. Therefore, we present benchmark datasets that have been instrumental in both shared contests and standalone research endeavors, showcasing their significance in this domain. Finally, we deliberate on the most effective approaches to hyperpartisan news detection, offering guidance for future endeavors.

5.2 Definition

Although hyperpartisanship is an attested social phenomenon, it lacks an entry in dictionaries. [Anthonio, 2019] found that the first statement was used during the U.S. 2016 election. Nonetheless, there is a shared definition proposed in diverse papers: one-sided articles that either avoid the dialogue between opposing ideologies or attack the antagonistic party [Kiesel et al., 2019, Jiang et al., 2019]. [Potthast et al., 2018] extend this definition by focusing on the linguistics and stylistic traits, arguing that this is a kind of news manifesting with highly emotional expressions. Moreover, [Potthast et al., 2018] relate hyperpartisan to fake news because of its emphasizing way to present events. Indeed, this news is propaganda-driven. After these considerations, we could stress the biased properties of hyperpartisan news, since it overlaps with the following biases and shows them in exaggerated forms:

1. rhetoric bias: the art of using language effectively to persuade or influence others. Rhetoric can be used to bias information by using persuasive language, emotional appeals, and logical fallacies;
2. ad hominem bias: an attack on the person making an argument, rather than the argument itself. Ad hominem attacks are often used to discredit the speaker and avoid addressing the merits of their argument;
3. opinion statement bias: a statement that expresses a personal belief or belief system. Opinion statements are not always biased, but they can be if they are presented as if they are facts;
4. ideology bias: a set of beliefs about what is right and wrong, good and bad. Ideologies can influence the way people interpret information and make decisions;
5. framing bias: how information is presented can influence how people perceive it. Framing can be used to bias information by highlighting certain aspects of a story while downplaying others;
6. coverage bias: the amount of attention that an issue or event receives from the media can influence how people think about it. Biased coverage can occur when the media disproportionately covers one side of an issue or event;
7. political bias: related to politics or the political system. Political bias can occur when information is presented in a way that favors one political party or ideology over another;
8. slant bias: Slant characterizes a form of media bias where journalists selectively present a portion of a story, emphasizing or spotlighting a specific angle or information fragment. Such slanting limits readers from accessing the complete narrative, thereby constraining the breadth of our comprehension.

As [Tucker et al., 2018] observes, hyperpartisanship can manifest on both the political left and right. We even posit that the center could exhibit hyperpartisanship due to its inherent linguistic tendencies. This form of hyperpartisanship, directly linked to the detection task, coexists with other types of radicalized polarization that transcend the left-right duality. For instance, [McCoy et al., 2018] asserts that polarization patterns remain consistent across socioeconomic, cultural, and historical contexts. Specifically, [McCoy et al., 2018] identifies various hyperpartisan typologies, including “globalist versus nationalist,” “religious versus secular,” and “ traditional versus modern cultural systems.” These typologies often involve the formation of two opposing groups, “Us” and “The Other,” leading to the suppression of the opposing group.

In some instances, hyperpartisan news falls outside the traditional disinformation taxonomy [Kapantai et al., 2021] and is instead considered within the broader realm of online content

[Molina et al., 2021]. However, [Recuero et al., 2020] establishes a strong link between polarization, hyperpartisanship, and disinformation. Moreover, other studies also associate hyperpartisan news with misleading content [Pierri et al., 2020, Huang and Lee, 2019]. Therefore, we position hyperpartisan news within the disinformation domain.

Now that we have a clear definition of hyperpartisan news and its categorization within the disinformation/misinformation spectrum, we can delve into hyperpartisan news detection. This task involves classifying news articles based on their hyperpartisan nature. Given a news article, headline, or fragments of its body, the model should be able to label it as hyperpartisan or mainstream [Kiesel et al., 2019]. Alternatively, the model could assign a political bias score ranging from left to right [Kim and Johnson, 2022, Alzhrani, 2022]. However, [Potthast et al., 2018] argues that binary classification approaches often overlook nuanced differences within various political leanings. [Sridharan, 2022] addressed this by employing a broader range of polarization.

[Aksenov et al., 2021] tackled hyperpartisan detection as a multi-class classification problem, utilizing both 7- and 5-point scales to define political leaning. Similarly, [Baly et al., 2019] employed a polarization scale. [Azizov et al., 2023] sought to manage political orientation by differentiating between right, center, and left.

By approaching these themes, we cannot avoid considering the relationship between governments and the Internet, because political parties rely on media to communicate with their voters. With the advent of cognitive capitalism, data and information is the immaterial reification of the power [Boutang, 2012]. Indeed, digital democracies adopted a one-to-many approach, meaning that each person can access news with multi-behavioral tendencies simultaneously. Through online social platforms, users are not only readers but spreaders and opinion makers [Sgueo, 2023]. Considering that, the more the audience engages with these contents, the greater the risk of polarization, eventually crystallizing into echo chambers that are reluctant to engage with divergent opinions [McCoy et al., 2018]. By entering the realm of post-truth, these tendencies are harmful to social cohesion, because citizens, rather than focusing and being conscious of the political object of the discussion, are being polarized towards a contextualized epistemological position, and have fallen into the illusion that their stance partiality is a comprehensive response, unaware that lacks to discuss the core of the topics. Within this perspective, [Gaughan, 2017] notices that: i) the trust in the institutions and voting systems is falling; ii) opposing parties face adversarial attacks so that voters cannot properly be in the position to adequately understand a topic because they are closed in their political cluster; iii) it is highly improbable that the dominant parties in state legislatures will voluntarily give up the political benefits linked to governing election procedures.

Hyperpartisan articles exhibit a distinct linguistic fingerprint. Studies by [Pérez-Almendros et al., 2019] and [Dumitru and Rebedea, 2019] revealed an abundance of adjectives and adverbs, while [Knauth, 2019b] identified a heavy reliance on pronouns and words conveying disgust. Additionally, these articles typically feature elongated paragraphs [Hanawa et al., 2019] and a sensational writing style characterized by emotionally charged language and uncommon terminology [Sengupta and Pedersen, 2019]. Notably, [Lyu et al., 2023] explored the tendency of right-leaning media to employ hyperpartisan headlines, but research suggests that news headlines in general often exhibit these characteristics [Amason et al., 2019].

5.3 Approaches

AI technologies are applied to detect this kind of news, so we are at the intersection between AI and Linguistics. Far from being a closed task, the applied approaches from distinct researchers show the versatility and complexity of this task. Given the prevalence of hyperpartisan news circulation online, the methodologies discovered target online publishers and social networks specifically. These characteristics infiltrate various online spaces, disseminated by multiple entities.

When considering publishers, a linguistic approach can be applied to detecting hyperpartisanship in news analysis. This involves examining textual information in articles, utilizing style-based or topic-based models [Sánchez-Junquera et al., 2021, Potthast et al., 2018, Lyu et al., 2023, Smădu et al., 2023]. Detection methods may focus on specific sections like titles [Lyu et al., 2023, Amason et al., 2019], sentences [Lim et al., 2018], quotes in the body [Pérez-Almendros et al., 2019], or encompass multiple elements [Naredla and Adedoyin, 2022, Gangula et al., 2019, Papadopoulou et al., 2019, Lyu et al., 2023, Nguyen et al., 2019].

Additionally, the entities involved in the writing and publishing process were considered. Detection of article polarization based on a journalist's leaning was implemented [Alzhrani, 2022]. Considering publishers as interconnected entities forming a polarized network, metadata like external links can be used for analysis [Hrckova et al., 2021, Kulkarni et al., 2018, Alabdulkarim and Alhindi, 2019, Joo and Hwang, 2019]. While bias determination from the source is feasible [Alzhrani, 2022, Alzhrani, 2020], it is crucial to note that bias does not always denote hyperpartisanship [Tran, 2020, Jiang et al., 2019]. In fact, the information source alone might not determine an article's hyperpartisanship [Jiang et al., 2019].

5.3.1 Algorithms

Let us introduce the best methodologies to detect hyperpartisan news distinguishing the nature of each approach in the following paragraphs. Non-Deep Learning algorithms refer to traditional Machine Learning models like Logistic regression; Deep Learning Methods encompass Convolutional Networks and the derived models; instead, Transformers [Devlin et al., 2019] collect the best approaches that use architectures from the Transformer family; finally, in Other Methods, there are mixed methodologies.

Non-Deep learning methods. During the SemEval 2019 task 4, [Srivastava et al., 2019] combined diverse polarity granularities at both sentence and article levels. They also introduced subjectivity, modality, and bias lexicons to analyze the author's point of view. In addition to this, to capture the semantic relationships, they used Universal Sentence Encoder, a document representation agnostic of word order. With a concatenation of handcrafted features and semantic ones, they trained an L2-regularized logistic regression. They achieved the best result by feeding the model with distinct features, finding that the most relevant were bias lexicon and polarity-based ones. In the SemEval 2019 secondary task of detecting the biased news publishers, [Bestgen, 2019] reproduced e̊f's approach. By evaluating the efficacy of a bag-of-words (BoW) methodology, they discovered that publishers differ in the presence of specific tokens. However, they confirmed the effectiveness of [Potthast et al., 2018]'s research. On the other hand, [Bestgen, 2019] ranked first in the detection task using *by-publisher* dataset. Tintin team ranked first in distinguishing documents published by hyperpartisan media outlets from unbiased ones due to its exclusive reliance on the tokens that make up the documents and a standard supervised learning procedure. This approach, although simple, proved to be effective in this specific aspect of the task.

Deep learning methods. [Jiang et al., 2019] successfully participated in and emerged victorious in the SemEval-2019 task 4 Hyperpartisan News Detection challenge performed on the by-article dataset. Their system, employing the ELMo Sentence Representation Convolutional Network, achieved the top spot in accuracy. The central concept of their work revolves around developing and implementing a system capable of anticipating hyperpartisan news, a pressing concern in today's media landscape. The system leverages ELMo embeddings to generate word and sentence representations and utilizes a lightweight Convolutional Neural Network (CNN) with batch normalization for document classification. The capacity of ELMO embeddings to handle polysemy and homonymy reflects its ability to grasp contextualized representations. Their experimental results clearly demonstrate the effectiveness of their approach, encompassing the influence of pre-training and fine-tuning on various datasets. Surprisingly, the by-publisher dataset exhibited a negative impact on the model's performance, potentially due to its abundance of noisy data.

Transformers. To detect the hyperpartisan news headlines, [Lyu et al., 2023] adopted a BERT-base model and developed a new dataset. They discerned between hyperpartisan and non-hyperpartisan words within each title, and also distinguished the political orientation. Based on these findings, they computed the distribution of topics over time, so they were able to analyze the coverage of the news media, reflected in the spread of societal issue topics. In terms of style, Left and Right are very similar to each other, except for the period after each election, when a seasonal pattern is registered.

Other methods. In 2018, [Potthast et al., 2018] built a corpus with news collected from 9 diverse publishers in a period close to the 2016 U.S. election. They demonstrated that political leaning cannot rely on discriminative features like paragraphs, quotes, and hyperlinks. Furthermore, regarding hyperpartisanship, the writing style from Left to Right publishers does not vary so consistently, but hyperpartisan news can be easily differentiated from the mainstream. They found these results when comparing different approaches, including BoW, n-grams and readability scores. With an accuracy of 0.75 for the style-based model, this approach outperforms the topic-based one. In the same year, [Kulkarni et al., 2018] introduced a multi-view document attention model (MVDAM) that can capture the title, structure, and metadata of a news article to effectively infer its political ideology. This Bayesian-based framework employs distinct models to construct a 3D representation: a convolutional neural network (CNN) for processing the title, Node2Vec for analyzing the network, and Hierarchical Attention Network (HAN) for extracting insights from the content.

5.4 Datasets

In the last years, several initiatives like SemEval-2019 task 4 [Kiesel et al., 2019] and CheckThat! [Azizov et al., 2023] encouraged the technological enhancements for hybrid tasks like Natural Language Processing applied in the Social Sciences. The high adhesion from researchers implies the importance of the contributions. Indeed, because hyperpartisan detection became a relevant task, several datasets were built. Nonetheless, due importance has not been recognized in this field of study, which remains relegated to a secondary task. The labelling process of data or its collection is often supported by the usage of specialized platforms like Allsides[4], Factcheck[5], Politifact[6], as ground truth for establishing the bias of an article and as source where to collect data. For instance, Allsides.com gives partisan scores to each article on the same topic but collected from politically opposed publishers. Indeed, [Baly et al., 2020] remarked that voluminous datasets contain noisy data that could reduce the performance. From that moment, researchers started to prefer the quality of the data rather than their volume. Indeed, [Lyu et al., 2023], by analyzing the SemEval 2019 task 4 datasets, discovered issues regarding class imbalance, task-label misalignment, and distribution shift.

We can notice from the data reported that the following issues need to be addressed:

1. The Absence of a distinct dataset for hyperpartisan vs. partisan news impacts classification accuracy.

2. Overemphasis on English news affects the representation of minority languages and their contexts.

3. Lack of computational studies on cross-lingual comparison highlights a knowledge gap.

4. Limited availability of data over time due to paywalls and copyright restrictions poses barriers.

5. Temporal lexicon constraints.

5.5 Research direction

Despite the advancements made in hyperpartisan news detection, there remain significant challenges that hinder its widespread adoption and effectiveness. One critical hurdle is the limited availability of datasets for minority languages, particularly in developing countries, where the prevalence of hyperpartisan news is particularly concerning. This lack of language-specific datasets severely restricts the ability to develop and deploy detection models for these regions. Additionally, the absence of multilingual models hinders the comparison of hyperpartisanship across diverse countries and cultures. This lack of cross-cultural understanding impedes our ability to grasp the nuances of hyperpartisanship and develop effective detection methods that transcend linguistic barriers.

[4] https://allsides.com/

[5] https://mediabiasfactcheck.com/

[6] https://www.politifact.com/

Dataset	Reference	Year	Size	Bias Label	Lang.	Availability
Task 3A	[Azizov et al., 2023]	2023	55,000	AllSides	English	Yes
Task 3B	[Azizov et al., 2023]	2023	8,000	AllSides	English	Yes
Allsides-L	[Ko et al., 2023]	2023	719,256	Allsides	English	Yes
No Name	[Lyu et al., 2023]	2023	1,824,824	AllSides, Media Bias Factcheck	English	No
Framing Triplet Dataset	[Kim and Johnson, 2022]	2022	25,627	Media Bias Factcheck	English	Yes
TVP Info	[Szwoch et al., 2022]	2022	81,694	NONE	Polish	Upon request
TVN 24	[Szwoch et al., 2022]	2022	128,527	NONE	Polish	Upon request
BIGNEWS	[Liu et al., 2022]	2022	3,689,229	Allsides, adfontes-media	English	Upon request
BIGNEWSBLN	[Liu et al., 2022]	2022	2,331,552	Allsides, adfontes-media	English	Upon request
BIGNEWSALIGN	[Liu et al., 2022]	2022	1,060,512	Allsides, adfontes-media	English	Upon request
No Name	[Sánchez-Junquera et al., 2021]	2021	1,555	BuzzFeed	English	No
StereoImmigrants	[Sánchez-Junquera, 2021]	2021	3,704	Manual	Spanish	Yes
No Name	[Pierri et al., 2020]	2020	~37000	None	Italian	No
PoliNews	[Tran, 2020]	2020	~83,000	None	English	No
Presidential	[Alzhrani, 2020]	2020	178,572	Allsides, Media Bias Factcheck	English	No
POLUSA	[Gebhard and Hamborg, 2020]	2020	~0.9M	None	English	Yes
No Name	[Baly et al., 2020]	2020	34,737	Allsides	English	Yes
All-Sides	[Li and Goldwasser, 2019]	2019	10,385	None	English	No
Telugu	[Gangula et al., 2019]	2019	1,327	Manual	Telugu	Yes
SemEval-2019 by-article	[Kiesel et al., 2019]	2019	1273	manually labeled	English	Yes
SemEval-2019 by-publisher	[Kiesel et al., 2019]	2019	754,000	BuzzFeed news, Media Bias Factcheck	English	Yes
The BuzzFeed-Webis Fake News Corpus 2016	[Potthast et al., 2018]	2018	1,627	BuzzFeed	English	Yes

Table 10: This table shows the benchmark datasets for hyperpartisan news detection.

The study by [Aksenov et al., 2021] highlights the importance of employing a more fine-grained label set for hyperpartisan news detection. By classifying news articles into more nuanced categories, such as the level of bias and the type of target audience, these models can provide a more comprehensive understanding of the underlying motivations and strategies employed by hyperpartisan sources. This more refined approach can lead to more effective detection and mitigation of hyperpartisan content.

To address these challenges and advance the state of the art in hyperpartisan news detection, several crucial steps need to be taken. Firstly, there is a need for the creation of comprehensive datasets for minority languages, particularly in developing countries. These datasets should be carefully curated and annotated to reflect the specific linguistic and cultural context of these regions. Secondly, the development of multilingual models is essential to enable cross-cultural comparisons and the identification of universal patterns of hyperpartisanship. These models should be trained on diverse datasets and incorporate linguistic and cultural features to capture the nuances of hyperpartisan content across different languages and geographies. Finally, the exploration of fine-grained labeling schemes should be prioritized to provide a more nuanced understanding of hyperpartisanship and enable the development of more effective detection methods. By addressing these challenges and embracing these advancements, we can move closer to a future where the spread of hyperpartisan news is effectively countered, ensuring a more informed and equitable global information landscape.

6 Automated disinformation agents

6.1 Context

This section of the report will introduce the task of automatically detecting textual content that has been generated by so-called ‘bots’ in online spaces with the malicious intent to deceive its audience. Although there are numerous definitional several issues at stake in the context of this task, this section will restrict itself to reviewing work related to the detection of political social bots. The rest of the subsection will proceed as follows: First, the context of the task of detecting automated deceptive content, i.e., the task of detecting the activities of malicious social bots, will be put into the context of the landscape of online disinformation throughout in the subsequent paragraphs. Next, we will explore how to define the concept of a ‘bot’ and the task of detecting bot-generated content 6.2. Additionally, the third and fourth subsections will deal with highlighting the main approaches that have been employed in the detection of bots in online spaces 6.3, and some of the most commonly used data that is currently available for the task 6.4. The final subsection will highlight some directions that lie open for future research on the task of bot detection 6.5.

In the context of disinformation, examining bots is relevant for the reason that they have been deployed in on social media platforms with the express purpose of disrupting or polluting the informational climate in which political debates and campaigns take place. Given that such malicious bots have mainly been deployed on Online Social Networks (OSNs), research on political bots of this kind (more on the definitions of bots in the subsequent subsection 6.2) has mainly occurred in the previous decade and up to the present day. Despite the relatively short history of bot detection research, the output in this period has been – and continues to be – one of considerable and growing intensity [Cresci, 2020]. Research on social bot detection often focuses on political developments, with elections and referendum votes in the United States [Addawood et al., 2019, Badawy et al., 2018, Luceri et al., 2019], United Kingdom [Howard and Kollanyi, 2016], Germany [Neudert et al., 2017], Spain [Stella et al., 2018, García-Orosa et al., 2021], France [Ferrara, 2017], and Sweden [Fernquist et al., 2018, Martini et al., 2021] constituting some recent case studies in that has been examined in this time period. In this context, research from 2016-onwards has a particularly rapid increase of interest from the research community [Cresci, 2020]. In the specific context of disinformation, at present, relatively little work has been to investigate the direct involvement of social bots (cf. 6.2) exclusively. Some work has examined the partial involvement of bots in the spread of false news [Ruchansky et al., 2017, Vosoughi et al., 2018, Faris et al., 2017]. [Shao et al., 2018] have investigated the role played by disinformation in and immediately after the 2016 U.S. presidential election. In the context of the previous chapters, there is particularly relevant research on social bots and the role they have played in the spread of disinformation in the health domain (specifically in relation to vaccine skepsis) [Broniatowski et al., 2018, Ferrara, 2020, Allem and Ferrara, 2018] and the relationship between bots and hate speech narratives [Uyheng and Carley, 2020, Albadi et al., 2019]. Compared to other types of Social Media Bots (SMBs), more on this later, research in this direction still remains ripe for development.

The increased prevalence and research interest into deceptively malicious bots largely coincides with the advent of social media networks in the web 2.0 era, in a timeline of what [Cresci, 2020] calls "the social bot pandemic". Even benign bots can, for example, play an unwitting part in propagating unverified information or rumors [Gupta et al., 2013, Ferrara et al., 2016, Starbird, 2019].

6.2 Definitions

One initial challenge associated with detecting bots is how to define them. Many researchers lean towards, across disciplines, typically lean towards a provisional definition of bots like the following. **Bot**: Any software or algorithm[7] that by function and design behaves in a goal-oriented manner, typically by mimicking some (set of) human behavior in a virtual or digital environment [Alsmadi and O'Brien, 2020]. It should be noted, that because researchers differ in terms of their focus and methods, there is no precise, universally-agreed upon definition or taxonomy of bots, for example, in the context of social media. This fact complicates the development and accessibility of appropriate labeled data [Orabi et al., 2020] (which will be discussed later 6.4). In this broad sense, the concept of a bot can also be traced to the infancy of Artificial Intelligence research from Turing's imitation game [Turing, 1950, Ferrara et al., 2016].

A first distinction to be made here between **Social bots**: Bots that mimic social human behaviors [Khaund et al., 2018] and which have been deliberately designed to *appear* human[8] [Boshmaf et al., 2013, Ferrara et al., 2016], and 'purely functional' bots, those that mimic other types of behaviors, such as financial trading [Huang et al., 2019, Cresci et al., 2019] or sending (spam) emails. The earliest mentions of the term 'social bot' can be found in [Boshmaf et al., 2013]. Klopfenstein et al. have proposed the term "botplication" for bots of the latter type [Klopfenstein et al., 2017], although this term has seen little adoption. As for the former category, this is where one finds so-called 'chatbots' (or 'chatterbots'). **Chatbots** are bots that are designed to mimic a partner in a conversational situation, be it in speech or text. They include conversational agents, such as the first chatbots like ELIZA [Weizenbaum, 1966b] and its descendants like modern virtual assistants such as Google Assistant, Apple's Siri and Amazon's Alexa, and

[7] Historically speaking, the term 'bot' is derived from software + 'robot' as coined in the English translation of the 1920 play *R.U.R* by Karel Capek (1890-1938). While both terms share connotations of automated, deterministic, artificial, and non-human agency, the difference is that 'bot' typically refers to virtual or software-based informational agents, rather than mechanical, and often anthropomorphic, agents (as is the case with 'robot'). See https://www.britannica.com/topic/RUR, accessed 22-01-2024.

[8] I.e., they do not obviously declare or otherwise give away that they are human.

more task-oriented[9] agents such as customer service bots.

This brings us to the type of bots that is of interest to us here: Social Media Bots (SMBs) that act on Online Social Network (OSN) platforms, meaning essentially, chatbots specifically designed to generate content and interactions in an online space [Klopfenstein et al., 2017, Ferrara et al., 2016, Faris et al., 2017]. Although a number of taxonomies of SMBs have been proposed, with typical distinctions like spambots, social bots, sybils, and cyborgs [Aljabri et al., 2023] the rest of this section will restrict itself to social bots[10]. The reason for this is that social bots have the most relevance in the context of disinformation in this report. Like other kinds of SMBS, social bots can be distinguished based on the intents and goals that guide their design and functions. In short, social bots can be **benign or malign** [Ferrara et al., 2016, Stieglitz et al., 2017]. While benign social bots, can aid in propagating useful information, for example related to a political party's campaigns [García-Orosa et al., 2021], aggregating legitimate news or automatically responding in customer care functions [Ferrara et al., 2016]. Meanwhile, malicious social bots are deployed to 'persuading, smearing and deceiving' [Ferrara et al., 2016], for example by influence public sentiment and opinion through astroturfirng [Ratkiewicz et al., 2011] or spread disinformation in political contexts [Ferrara, 2023], or simply by polluting the online information ecosystem by sending spam [Lee et al., 2011, Cresci et al., 2017] and phishing [Shafahi et al., 2016].

The task of detecting social bots is typically construed as binary classification task with labels such as 'human' and 'bot'. The task is often executed on multilingual datasets, and recently, with labels to further detect the type of large language model that has generated the textual content of the task. To the extent that the approach is text-based, the task of detecting whether a text associated with an account or profile is generated by a machine or a human can be viewed as n applied form of the more generic task of Machine Generated Text detection (MGT).

6.3 Approaches

Various taxonomies of the approaches to social bot detection have been proposed [Orabi et al., 2020, Aljabri et al., 2023]. One can separate bot detection strategies by: Feature and techniques they employ and the scope of the detection strategy.

6.3.1 Feature types

Most approaches to bot detection combine multiple types of features. Some of the most popular categories of features are profile or user features (profile names, pictures), metadata, content features[11] (keywords, hashtags), textual or linguistic features (sentiment, syntax, style, named entities), graph-based features, behavioral (anomalous user action and interaction; likes, follows, (re)posts, # followers/friends) and temporal (time of day, post frequencies) features [Aljabri et al., 2023, Ferrara, 2023].

[9] Conversational agents have been divided into those that participate in conversations in a structured manner and those that do participate in an unstructured manner (typically in order to entertain) [Jurafsky and Martin,]. However, at the time of this writing, this boundary is blurring with the advent of Large Language Models and their imminent integration into virtual agents such as Google's "Assistant with Bard" initiative https://blog.google/products/assistant/google-assistant-bard-generative-ai/ (accessed 22-01-2024)

[10] Some authors refer to social bots as 'socialbots' [Boshmaf et al., 2011]

[11] This including text and other modalities

6.3.2 Detection scope

One second way of dividing social bot detection strategies is in terms of their scope. [Cresci, 2020] have suggested the division between approaches that model and detect social media bots at the individual level (i.e., the level of individual bots and their posts and behaviors) versus the level of groups (botnets and patterns of coordinated behavior). Over the past decade, the trend has been to shift away from individual-detection to group-detection [Cresci, 2020].

6.3.3 Techniques

As with other tasks A.I. and social media network research, social bot detection has also experienced trends. Since the beginning of social bot detection research, there have been three waves. The first was heuristics-based[12] approaches, the second saw the incorporation of natural language processing and machine learning techniques, while the third has seen the adoption of deep learning techniques [Cresci, 2020, Ferrara, 2023], and most recently, adversarial deep learning.

Heuristic approaches utilized shallow patterns of account activity, account metadata, simple content-based features (keywords, hashtags etc.), and simple network-based features (# of followers versus # of accounts followed) [Gorodnichenko et al., 2021, Ferrara, 2023, Cresci, 2020]. See [Howard et al., 2016] for a relatively recent example using only hashtag frequencies an indication of account automation. The examples were picked based on their employing content, text, and optionally, other features to detect social bots since these are closest to the topic of this report.

The most popular ML techniques are by far supervised learning, followed by unsupervised and semi-supervised [Aljabri et al., 2023].

Supervised machine learning approaches Both traditional or 'shallow' learning and deep learning techniques are popular in social bot detection. In approaches that employ traditional machine learning, Random Forest tends to be the best performing and most popular choice of algorithm [Orabi et al., 2020]. Other ensemble approaches like [Sayyadiharikandeh et al., 2020] have also been applied. Support Vector Machines, Decision Trees and (Naïve) Bayes are likewise popular, well-performant choices. Boosting algorithms like AdaBoost, BoostOr and XGBoost are sometimes also used [Aljabri et al., 2023]. Examples of traditional supervised ML approaches that have achieved high performance on detecting social bots while incorporating content features include [Davis et al., 2016], who achieve 0.95 AUC (Area Under ROC Curve) on a dataset of 31K tweets collected with the first version of Botometer (see 6.3.4. Another example is [Fernquist et al., 2018] who achieve F1 of 0.958 using, with their highest performing Random Forest model, and the `cresci-2015` [Cresci et al., 2015] and `varol-2017` [Varol et al., 2017] datasets. If one looks at the most recent PAN shared task on bot detection [Daelemans et al., 2019], the top performing submission used an SVM (average acc. 0.88) on the `pan-2019` dataset [Pizarro, 2020]. Among deep learning approaches to social bot detection, Long short-term memory recurrent neural, along with Convolutional Neural Nets (CNNs), Graph Neural Nets and Multilayer Perceptrons have been popular choices [Ferrara, 2023, Hayawi et al., 2023]. Some examples of performant DL approaches include [Attia et al., 2023], [Wu et al., 2021], [Yang et al., 2022a]

[12] Also referred to as rule-based.

and [Knauth, 2019a]. Using the pan-2019 dataset, [Attia et al., 2023] train a dual-channel multi-dimensional convolutional neural net (DCMDCNN) architecture which achieves a combined F1 score of 0.91. Wu et al. [Wu et al., 2021] implement ResNet [Wang et al., 2017] and BiGRU [Tang et al., 2015] blocks with an attention layer into a model they call a Residual Gated Attention (RGA) architecture, which achieves an F1 score of 0.9886 on data crawled from Weibo in combination with the SWLD-20K dataset. In a similar approach, [Yang et al., 2022a] apply fusion and attention layers to BiGRU blocks on a similarly composed dataset and achieve an F1 score of 0.983. [Knauth, 2019a] combines AdaBoost [Freund and Schapire,] with SMOTE-ENN resampling [Lemaître et al., 2017] and reach an accuracy of 0.988 on the cresci-2017[13] dataset [Cresci et al., 2017]. Kenyeres and Kovacs experiment with combining LSTM, fine-tuned BERT and AdaBoost architectures on the pan-2019 dataset, with the best model being an ensemble model using AdaBoost that achieves 0.9 F1 [Kenyeres and Kovács, 2022].

As mentioned previously, one of the more recent developments is the advent of adversarial training regimes in social bot detection [Cresci et al., 2021, Ferrara, 2023]. Like in other domains, this approach works by training a Generative Adversarial Network (GAN), itself comprised of a generator network which creates adversarial examples that are then passed on to a discriminator network which tries to evaluate whether a given instance is a bot or human [Goodfellow et al., 2014, Yu et al., 2017]. One example of this approach is GANBOT [Najari et al., 2022], where a discriminator and generator are connected using an LSTM layer. After training on the cresci-2017 dataset, the authors achieve better probability scores for bot accounts in the dataset than what can be achieved with a contextual LSTM. Additionally, transfer learning regimes have also been used in [Guo et al., 2022a, Heidari et al., 2022]. Guo et al. [Guo et al., 2022a] propose a model that fuses a two-layer BERT model with a Graph Convolutional Net (GCN), and use the cresci-rtbust-2019 [Mazza et al., 2019], botometer-feedback-2019 [Yang et al., 2019], gilani-2017 [Gilani et al., 2017], cresci-stock-2018 [Cresci et al., 2019, Cresci et al., 2018] and midterm-2018 [Yang et al., 2020] datasets. Heidari et al. [Heidari et al., 2022] base a stacked feed forward neural net with GloVe [Pennington et al., 2014] and ELMO [Peters et al., 2018] embeddings with good results (0.941 F1) on the cresci-2017 dataset.

Unsupervised machine learning approaches to social bot detection are significantly less popular and mainly involve clustering and association algorithms. The most popular among these are K-Nearest Neighbor (KNN), K-Means and Principal Component Analysis [Aljabri et al., 2023]. One interesting recent example is [Mazza et al., 2019] utilize unsupervised clustering with and LSTM and achieve 0.87 F1 by focusing on retweets and temporal features.

Semi-supervised techniques combine labeled with unlabeled training instances. In social bot detection, this kind of approach have seen very little adoption thus far, with a few exceptions [Cao et al., 2014, Shi et al., 2019]. These approaches typically do not focus on (textual) content or linguistic features.

6.3.4 Tools and frameworks

Over the years, researchers have developed off-the-shelf-tools that facilitate the detection and analysis of social bots for non-technical audiences such as social science scholars, (data) jour-

[13] Occasionally referred to as 'MIB'.

nalists and the like. There are 3 prominent tools available as of this writing: Botometer (formerly BotOrNot) [Davis et al., 2016][14], the R package TweetBotOrNot [Kearney, 2023], and Alexandria [Graham and FitzGerald, 2023]. Only the former two are free and open to the public, and only Botometer is aimed at non-technical as well as technical end-users [Sayyadiharikandeh et al., 2020].

6.4 Datasets

As noted by [Aljabri et al., 2023], “the dearth of publicly accessible datasets for OSNs such as Facebook, Instagram, and LinkedIn is one of the greatest obstacles in this research area”. And although X (formerly known as Twitter) has had a history of publicly open APIs that have been free to use for researchers, these too are now severely limited both in terms of the rates with which they can be queried and the (paid) access they have required since 2023 [Yang et al., 2023a]. Despite this fact, most datasets (public as well as private) on bot detection are derived from X platform [Aljabri et al., 2023]. The most relevant datasets for social bot detection in the context of this report are listed in table 11. In addition to social bot datasets, the Botometer repository[15] hosts many other publically available datasets for spam, sybil and fake account detection. The most relevant datasets reviewed in this section in relation to social bot detection are listed in table 11.

Dataset	Reference	Accounts	Size	Source	Language	Public
cresci-2015	[Cresci et al., 2015]	3,900	2,750,057	Twitter	EN	Yes
cresci-2017	[Cresci et al., 2017]	12,736	6,637,615	Twitter	EN	Yes
varol-2017	[Varol et al., 2017]	∼31,000	NA	Twitter	NA	Yes
midterm-2018	[Yang et al., 2020]	50,537	NA	Twitter	NA	Yes
botometer-feedback-2019	[Yang et al., 2019]	528	NA	Twitter	EN	Yes
cresci-rtbust-2019	[Mazza et al., 2019]	467,241	10M	Twitter	IT	Yes
gilani-2017	[Gilani et al., 2017]	3,065	722,109	Twitter	EN	Yes
cresci-stock-2018	[Cresci et al., 2019, Cresci et al., 2018]	467,241	7,855,518	Twitter	EN	Yes
pan-2019	[Rangel and Rosso, 2019]	11,560	1,156,000	Twitter	EN, ES	No

Table 11: Datasets commonly used in malicious social media bot detection.

[14] https://botometer.osome.iu.edu/

[15] https://botometer.osome.iu.edu/bot-repository/datasets.html (accessed 30-01-24)

6.5 Research directions

As mentioned previously, the lack of a precise and widely adopted definition of SMBs 6.2 means that finding ground truth datasets for the task of detecting SMBs in general is difficult [Orabi et al., 2020]. Some of the most commonly identified challenges in social bot detection are 1) the fact that bots develop rapidly and evolve faster than detectors, making it difficult for researchers to keep up in real time [Orabi et al., 2020]. In particular, the rapidly evolving and increasingly sophisticated state of A.I., such as LLMs, provide a serious challenge in this direction, in that it is becoming much more difficult to distinguish human from machine generated textual content [Ferrara, 2023]. 2) data scarcity, and generalizability issues. As mentioned, finding public data for social bot detection is difficult, but it is also a problem that most data available for the task is in English or other large European languages and mainly collected from X (formerly Twitter). Additionally, platforms also differ in terms of their API access, in fact, the same platform might change the data provided via its API over time along with the access to this without warning [Shahid et al., 2022, Yang et al., 2023a]. 3) definitional issues, such as a lack of widely agreed upon definitions and criteria for what constitutes different types of bots have also been raised and pose difficulties for building detection systems and reproducing their results across datasets [Martini et al., 2021].

Based on these challenges, some direction for future research are the following. 1) developing novel and robust detection models, this could extend on transfer learning, adversarial and ensemble approaches which have shown promising results in recent years. It could take the form of hybrid approaches (in the representational sense), such as neuro-symbolic A.I. [Hamilton et al., 2022]. Building systems that can handle real-time data (instead of static platform data) can also be explored in this direction. 2) to mitigate generalizability issues, developing and sharing high quality benchmark datasets of real world examples of social bots in a variety of domains and from a range of platforms, and making these widely available. Developing systems which can learn from feedback via human-in-the loop architectures might also be an option, one which will also aid activists, journalists, and other practitioners in identifying bots in e.g., political campaigns. 3) Not all bots are created equal or for the same purposes. Since bots are purpose-driven software agents, better taxonomies and definitions that can be operationalized in detection could be beneficial both in detection and mitigation efforts. As noted in [Martini et al., 2021, Shahid et al., 2022], differences in conceptualizations and assumptions made about bots impact dataset collection, annotation and ultimately, model training and performance to the point where even the three most popular tools 6.3.4 have a very low degree of agreement on the same datasets. In conclusion, this section has examined the task of detecting automated agents commonly referred to as ‘bots’, which have been shown to interfere in the political domain on online social media platforms. Among other things, the malicious versions of such social bots have been observed to be involved with the proliferation of dis- and misinformation in online social networks. Since the advent of these platforms, detections strategies and the researchers who devise them, have been locked in an increasingly intensifying arms-race, which has scaled along with big-data and modern A.I. compute capabilities and paradigms. New advances, in particular large language models as they can be applied in generative adversarial neural networks, are expected to escalate this race in the near future and further blur the boundaries between human and non-human forms of (linguistic) agency in our modern online informational environments and in our democracies.

In the next chapter, we draw some conclusions about the main topics expressed in this report (Deliverable D3.1) and provide some future work directions for the HYBRIDS project.

7 Conclusions

In this report we analyzed different aspects of disinformation spread and detection. We highlighted how it is important to achieve an agreement on definitions, presented the most successful strategies for identifying disinformation, underscored some of the current limitations and outlined current research directions.

In chapter 2, we discussed the automated fact-checking pipeline. The research has focused on identifying verifiable claims, prioritizing claims, identifying if claims have already been fact-checked, finding evidence for a claim and predicting if a claim is true or false based on the found evidence. Some of the current limitations include the lack of comprehensive datasets, solutions usually restricted to a language or domain, scarcity of time-aware fact-checking research and limited results in multi-modal fact-checking. As seen in the chapter, the recent advances in large language models showed great results in many NLP and multi-modal tasks and opened new possibilities for automated fact-checking.

In the following chapter, we saw how misinformation detection has been applied to the health domain. The spread of misleading and harmful information about health issues became evident during the covid pandemic. Even before that, rumors and false information about therapeutics consolidated in the scientific community, like vaccines, started to be vastly disseminated. Regarding the research trying to prevent misinformation spread, we saw the distinction between misinformation and disinformation, approaches involving language models and linguistic features and also involving consumer research. As future directions, we highlighted the need for refining medical terminologies and the development of advanced discourse analysis models.

Chapter 4 discussed hate speech detection, including models and datasets related to the task. Models like HateBERT were an important step in the advance of hate speech detection. Some of the current research directions include adopting a common definition of hate speech, increasing the availability of datasets, avoiding bias during dataset creation, having the context in mind while annotating the data, developing new models and focusing on low-resource languages.

Then we discussed hyperpartisan news detection. We started by discussing the definitions and categories of hyperpartisan news. Then we presented some of the approaches to detect hyperpartisan news, like logistic regression models, bag-of-words, embeddings, CNNs and transformers. Same as seen in the other chapters, lack of datasets of low-resource languages is a main issue. Other research directions include the development of datasets with more nuanced categories and multilingual models that allow for the identification of hyperpartisan patterns in different languages.

The last chapter examines bot detection. Bot generated content can easily reach large audiences and heavily influence political campaigns and the public opinion in general. After discussing the definitions of bots, we discussed some methods of detecting them, namely the heuristic-based, supervised, semi-supervised and unsupervised machine-learning approaches.

We then present datasets that can be used to detect bots and outline some of the challenges for bot detection, like their rapid evolution, data scarcity and lack of agreement on the definitions. Some research possibilities include using transfer learning, adversarial, ensemble approaches or even neuro-symbolic approaches.

References

[Abualsaud, 2019] Abualsaud, M. (2019). Exposure and order effects of misinformation on health search decisions.

[Addawood et al., 2019] Addawood, A., Badawy, A., Lerman, K., and Ferrara, E. (2019). Linguistic Cues to Deception: Identifying Political Trolls on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 13:15–25.

[Adhikari et al., 2019] Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019). Docbert: Bert for document classification.

[Agrestia et al., 2022] Agrestia, S., Hashemianb, A., and Carmanc, M. (2022). Polimi-flatearthers at checkthat! 2022: Gpt-3 applied to claim detection. *Working Notes of CLEF*.

[Akbik et al., 2018] Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

[Aksenov et al., 2021] 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.

[Alabdulkarim and Alhindi, 2019] Alabdulkarim, A. and Alhindi, T. (2019). Spider-jerusalem at SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 985–989. Association for Computational Linguistics.

[Alam et al., 2021a] Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Da San Martino, G., Abdelali, A., Sajjad, H., Darwish, K., et al. (2021a). Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 913–922.

[Alam et al., 2021b] Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Al-Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., and Nakov, P. (2021b). Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Albadi et al., 2019] Albadi, N., Kurdi, M., and Mishra, S. (2019). Hateful People or Hateful Bots? Detection and Characterization of Bots Spreading Religious Hatred in Arabic Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):61:1–61:25.

[Ali et al., 2021] Ali, Z. S., Mansour, W., Elsayed, T., and Al-Ali, A. (2021). Arafacts: the first large arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236.

[Aljabri et al., 2023] Aljabri, M., Zagrouba, R., Shaahid, A., Alnasser, F., Saleh, A., and Alomari, D. M. (2023). Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining*, 13(1):20. Read_Status: In Progress Read_Status_Date: 2024-01-31T00:39:56.988Z.

[Allem and Ferrara, 2018] Allem, J.-P. and Ferrara, E. (2018). Could Social Bots Pose a Threat to Public Health? *American Journal of Public Health*, 108(8):1005–1006. Publisher: American Public Health Association.

[Alsmadi and O'Brien, 2020] Alsmadi, I. and O'Brien, M. J. (2020). How many bots in russian troll tweets? *Inf. Process. Manag.*, 57(6):102303.

[Aly et al., 2021] Aly, R., Guo, Z., Schlichtkrull, M., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., and Mittal, A. (2021). Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

[Alzhrani, 2020] Alzhrani, K. (2020). Ideology detection of personalized political news coverage: A new dataset. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, pages 10–15. ACM.

[Alzhrani, 2022] Alzhrani, K. (2022). Political ideology detection of news articles using deep neural networks. *Intelligent Automation Soft Computing*, 33:483–500.

[Amason et al., 2019] Amason, E., Palanker, J., Shen, M. C., and Medero, J. (2019). Harvey mudd college at SemEval-2019 task 4: The d.x. beaumont hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 967–970. Association for Computational Linguistics.

[Anthonio, 2019] Anthonio, T. (2019). Robust document representations for hyperpartisan and fake news detection.

[Antypas and Camacho-Collados, 2023] Antypas, D. and Camacho-Collados, J. (2023). Robust hate speech detection in social media: A cross-dataset empirical evaluation. In Chung, Y.-I., R\"ottger, P., Nozza, D., Talat, Z., and Mostafazadeh Davani, A., editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

[Arampatzis and Robertson, 2010] Arampatzis, A. and Robertson, S. (2010). Modeling score distributions in information retrieval. *Information Retrieval*, 14(1):26–46.

[Ash et al., 2006] Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., and Grossman, M. (2006). Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology*, 66(9):1405–1413.

[Attia et al., 2023] Attia, S. M., Mattar, A. M., and Badran, K. M. (2023). Social Bot Detection based on Language-independent Stylometric Analysis using dual-channel multi-dimensional CNN. In *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 241–246, Cairo, Egypt. IEEE.

[Aziz et al., 2023] Aziz, A., Hossain, M., and Chy, A. (2023). Csecu-dsg at checkthat! 2023: transformer-based fusion approach for multimodal and multigenre check-worthiness. *Working Notes of CLEF*.

[Azizov et al., 2023] Azizov, D., Nakov, P., and Liang, S. (2023). Frank at checkthat! 2023: Detecting the political bias of news articles and news media.

[Badawy et al., 2018] Badawy, A., Ferrara, E., and Lerman, K. (2018). Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265, Barcelona. IEEE. 142 citations (Crossref) [2023-12-13].

[Baly et al., 2020] Baly, R., Da San Martino, G., Glass, J., and Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991. Association for Computational Linguistics.

[Baly et al., 2019] Baly, R., Karadzhov, G., Saleh, A., Glass, J., and Nakov, P. (2019). Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In Burstein, J., Doran, C., and Solorio, T., editors, *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 2109–2116. Association for Computational Linguistics.

[Barbieri et al., 2022] Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

[Baumgartner and Hartmann, 2011] Baumgartner, S. E. and Hartmann, T. (2011). The role of health anxiety in online health information search. *Cyberpsychology, Behavior, and Social Networking*, 14(10):613–618.

[Beltagy et al., 2019] Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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

[Bestgen, 2019] Bestgen, Y. (2019). Tintin at SemEval-2019 task 4: Detecting hyperpartisan news article with only simple tokens. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1062–1066. Association for Computational Linguistics.

[Bigoulaeva et al., 2021] Bigoulaeva, I., Hangya, V., and Fraser, A. (2021). Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv. Association for Computational Linguistics.

[Bingham and Mannila, 2001] Bingham, E. and Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250.

[Black et al., 2021] Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with meshtensorflow.

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520.

[Bonet-Jover et al., 2021] Bonet-Jover, A., Piad-Morffis, A., Saquete, E., Martínez-Barco, P., and Ángel García-Cumbreras, M. (2021). Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. *Expert Systems with Applications*, 169:114340.

[Borzekowski et al., 2010] Borzekowski, D. L. G., Schenk, S., Wilson, J. L., and Peebles, R. (2010). e-ana and e-mia: A content analysis of pro–eating disorder web sites. *American Journal of Public Health*, 100(8):1526–1534.

[Boshmaf et al., 2011] Boshmaf, Y., Muslukhov, I., Beznosov, K., and Ripeanu, M. (2011). The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th Annual Computer Security Applications Conference*, ACSAC '11, pages 93–102, New York, NY, USA. Association for Computing Machinery.

[Boshmaf et al., 2013] Boshmaf, Y., Muslukhov, I., Beznosov, K., and Ripeanu, M. (2013). Design and analysis of a social botnet. *Computer Networks*, 57(2):556–578.

[Boutang, 2012] Boutang, Y. (2012). *Cognitive Capitalism*. Polity Press, Cambridge.

[Bouziane et al., 2020] Bouziane, M., Perrin, H., Cluzeau, A., Mardas, J., and Sadeq, A. (2020). Team buster. ai at checkthat! 2020 insights and recommendations to improve fact-checking. In *CLEF (Working Notes)*.

[Brennen et al., 2020] Brennen, J. S., Simon, F. M., Howard, P. N., and Nielsen, R. K. (2020). Types, sources, and claims of covid-19 misinformation.

[Broniatowski et al., 2018] Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., and Dredze, M. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health*, 108(10):1378–1384. Publisher: American Public Health Association.

[Buolamwini and Gebru, 2018] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

[Cao et al., 2014] Cao, Q., Yang, X., Yu, J., and Palow, C. (2014). Uncovering Large Groups of Active Malicious Accounts in Online Social Networks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 477–488, Scottsdale Arizona USA. ACM.

[Cartright et al., 2011] Cartright, M.-A., White, R. W., and Horvitz, E. (2011). Intentions and attention in exploratory health search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 65–74, New York, NY, USA. Association for Computing Machinery.

[Caselli et al., 2021] Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2021). HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

[Chen and Shu, 2023] Chen, C. and Shu, K. (2023). Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

[Chernyavskiy et al., 2021] Chernyavskiy, A., Ilvovsky, D., and Nakov, P. (2021). Aschern at checkthat! 2021: lambda-calculus of fact-checked claims. *Faggioli et al.[12]*.

[Cheung and Lam, 2023] Cheung, T.-H. and Lam, K.-M. (2023). Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853. IEEE.

[Chu et al., 2023] Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., and Liu, T. (2023). A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

[Clarke et al., 2020] Clarke, C. L., Rizvi, S., Smucker, M. D., Maistro, M., and Zuccon, G. (2020). Overview of the trec 2020 health misinformation track. In *TREC*.

[Cresci, 2020] Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, 63(10):72–83. 117 citations (Crossref) [2024-01-21].

[Cresci et al., 2015] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, 80:56–71.

[Cresci et al., 2017] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. (2017). The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, pages 963–972, Perth, Australia. ACM Press.

[Cresci et al., 2018] Cresci, S., Lillo, F., Regoli, D., Tardelli, S., and Tesconi, M. (2018). $FAKE: Evidence of Spam and Bot Activity in Stock Microblogs on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). Number: 1.

[Cresci et al., 2019] Cresci, S., Lillo, F., Regoli, D., Tardelli, S., and Tesconi, M. (2019). Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter. *ACM Transactions on the Web*, 13(2):11:1–11:27.

[Cresci et al., 2021] Cresci, S., Petrocchi, M., Spognardi, A., and Tognazzi, S. (2021). The coming age of adversarial social bot detection. *First Monday*.

[Cui and Lee, 2020] Cui, L. and Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset.

[Cui et al., 2020] Cui, L., Seo, H., Tabar, M., Ma, F., Wang, S., and Lee, D. (2020). Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '20. ACM.

[Daelemans et al., 2019] Daelemans, W., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., and Zangerle, E. (2019). Overview of PAN 2019: Bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection. *Cross-language evaluation forum*, pages 402–416. ECC: 0000015.

[Dai et al., 2020] Dai, E., Sun, Y., and Wang, S. (2020). Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:853–862.

[Das et al., 2023] Das, A., Liu, H., Kovatchev, V., and Lease, M. (2023). The state of human-centered nlp technology for fact-checking. *Information processing & management*, 60(2):103219.

[Davani et al., 2022] Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

[Davidson et al., 2019] Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the ICWSM 2017*, 11(1):512–515.

[Davis et al., 2016] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. (2016). BotOrNot: A System to Evaluate Social Bots. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, pages 273–274, Montr&#233;al, Qu&#233;bec, Canada. ACM Press.

[de Gibert Bonet et al., 2022] de Gibert Bonet, O., Kharitonova, K., Calvo Figueras, B., Armengol-Estapé, J., and Melero, M. (2022). Quality versus quantity: Building Catalan-English MT resources. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 59–69, Marseille, France. European Language Resources Association.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Du et al., 2021] Du, J., Dou, Y., Xia, C., Cui, L., Ma, J., and Yu, P. S. (2021). Cross-lingual covid-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 859–862.

[Du et al., 2022] Du, S., Gollapalli, S. D., and Ng, S.-K. (2022). Nus-ids at checkthat! 2022: identifying check-worthiness of tweets using checkthat5. *Working Notes of CLEF*.

[Du et al., 2023] Du, W.-W., Wu, H.-W., Wang, W.-Y., and Peng, W.-C. (2023). Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. *arXiv preprint arXiv:2302.07740*.

[Dumitru and Rebedea, 2019] Dumitru, V. C. and Rebedea, T. (2019). Fake and hyper-partisan news identification.

[Dutta et al., 2022] Dutta, S., Dhar, R., Guha, P., Murmu, A., and Das, D. (2022). A multilingual dataset for identification of factual claims in indian twitter. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 88–92.

[Ellery et al., 2008] Ellery, P. J., Vaughn, W., Ellery, J., Bott, J., Ritchey, K., and Byers, L. (2008). Understanding internet health search patterns: An early exploration into the usefulness of google trends. *Journal of Communication in Healthcare*, 1(4):441–456.

[ElSherief et al., 2018] ElSherief, M., Kulkarni, V., Nguyen, D., Yang Wang, W., and Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the ICWSM 2018*, 12(1).

[Endo et al., 2022] Endo, P. T., Santos, G. L., de Lima Xavier, M. E., Nascimento Campos, G. R., de Lima, L. C., Silva, I., Egli, A., and Lynn, T. (2022). Illusion of truth: Analysing and classifying covid-19 fake news in brazilian portuguese language. *Big Data and Cognitive Computing*, 6(2):36.

[Eysenbach, 2002] Eysenbach, G. (2002). Infodemiology: the epidemiology of (mis)information. *The American Journal of Medicine*, 113(9):763–765.

[Faris et al., 2017] Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., and Benkler, Y. (2017). Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election.

[Fernquist et al., 2018] Fernquist, J., Kaati, L., and Schroeder, R. (2018). Political Bots and the Swedish General Election. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 124–129, Miami, FL. IEEE.

[Fernández-Pichel et al., 2021a] Fernández-Pichel, M., Losada, D. E., Pichel, J. C., and Elsweiler, D. (2021a). *Comparing Traditional and Neural Approaches for Detecting Health-Related Misinformation*, page 78–90. Springer International Publishing.

[Fernández-Pichel et al., 2021b] Fernández-Pichel, M., Losada, D. E., Pichel, J. C., and Elsweiler, D. (2021b). *Reliability Prediction for Health-Related Content: A Replicability Study*, page 47–61. Springer International Publishing.

[Ferrara, 2017] Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*.

[Ferrara, 2020] Ferrara, E. (2020). What Types of COVID-19 Conspiracies are Populated by Twitter Bots? *First Monday*, pages 1–25. ECC: 0000031 arXiv: 2004.09531.

[Ferrara, 2023] Ferrara, E. (2023). Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*.

[Ferrara et al., 2016] Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7):96–104.

[Fortuna and Nunes, 2018] Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

[Founta et al., 2018] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ..., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the ICWSM 2018*, 12(1).

[Freund and Schapire,] Freund, Y. and Schapire, R. E. A Short Introduction to Boosting.

[Gangula et al., 2019] 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.

[García-Orosa et al., 2021] García-Orosa, B., Gamallo, P., Martín-Rodilla, P., Martínez-Castaño, R., García-Orosa, B., Gamallo, P., Martín-Rodilla, P., and Martínez-Castaño, R. (2021). Hybrid intelligence strategies for identifying, classifying and analyzing political bots. *Social Sciences*, 10(10). 8 citations (Crossref) [2024-01-21] tex.ids= garcia-orosaHybridIntelligenceStrategies2021a number: 10 publisher: Multidisciplinary Digital Publishing Institute (MDPI).

[Garg et al., 2018] Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

[Gaughan, 2017] Gaughan, A. J. (2017). Illiberal democracy: The toxic mix of fake news, hyperpolarization, and partisan election administration.

[Gebhard and Hamborg, 2020] Gebhard, L. and Hamborg, F. (2020). The POLUSA dataset: 0.9m political news articles balanced by time and outlet popularity. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 467–468. Conference Name: JCDL '20: The ACM/IEEE Joint Conference on Digital Libraries in 2020 ISBN: 9781450375856 Place: Virtual Event China Publisher: ACM.

[Gi et al., 2021] Gi, I.-Z., Fang, T.-Y., and Tsai, R. T.-H. (2021). Verdict inference with claim and retrieved elements using roberta. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65.

[Gilani et al., 2017] Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., and Crowcroft, J. (2017). Of Bots and Humans (on Twitter). In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 349–354, Sydney Australia. ACM.

[Goeuriot et al., 2021] Goeuriot, L., Suominen, H., Kelly, L., Alemany, L. A., Brew-Sam, N., Cotik, V., Filippo, D., Gonzalez Saez, G., Luque, F., Mulhem, P., Pasi, G., Roller, R., Seneviratne, S., Vivaldi, J., Viviani, M., and Xu, C. (2021). *CLEF eHealth Evaluation Lab 2021*, page 593–600. Springer International Publishing.

[Golbeck et al., 2017] Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., ..., and Wu, D. M. (2017). A large labeled corpus for online harassment research. In *Proceedings of the ACM WebSci 2017*. ACM.

[Goldzycher et al., 2023] Goldzycher, J., Preisig, M., Amrhein, C., and Schneider, G. (2023). Evaluating the effectiveness of natural language inference for hate speech detection in languages with limited labeled data. In Chung, Y.-I., R\"ottger, P., Nozza, D., Talat, Z., and Mostafazadeh Davani, A., editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 187–201, Toronto, Canada. Association for Computational Linguistics.

[Gollapalli et al., 2023] Gollapalli, S. D., Du, M., and Ng, S.-K. (2023). Identifying checkworthy cure claims on twitter. In *Proceedings of the ACM Web Conference 2023*, pages 4015–4019.

[Goodfellow et al., 2014] 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.

[Gorodnichenko et al., 2021] Gorodnichenko, Y., Pham, T., and Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from #Brexit and #USElection. *European Economic Review*, 136:103772.

[Graham and FitzGerald, 2023] Graham, T. and FitzGerald, K. M. (2023). *Bots, Fake News and Election Conspiracies: Disinformation During the Republican Primary Debate and the Trump Interview*. Digital Media Research Centre, Queensland University of Technology, Brisbane, Qld.

[Grant et al., 2007] Grant, R., Clarke, R. J., and Kyriazis, E. (2007). A review of factors affecting online consumer search behaviour from an information value perspective. *Journal of Marketing Management*, 23(5–6):519–533.

[Guo et al., 2022a] Guo, Q., Xie, H., Li, Y., Ma, W., and Zhang, C. (2022a). Social Bots Detection via Fusing BERT and Graph Convolutional Networks. *Symmetry*, 14(1):30. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute Read_Status: To Read Read_Status_Date: 2024-01-31T00:01:23.074Z.

[Guo et al., 2022b] Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022b). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

[Gupta et al., 2013] Gupta, A., Lamba, H., and Kumaraguru, P. (2013). $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. In *2013 APWG eCrime Researchers Summit*, pages 1–12.

[Hale et al., 2024] 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*.

[Hamilton et al., 2022] Hamilton, K., Nayak, A., Božić, B., and Longo, L. (2022). Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web*, pages 1–42. 2 citations (Crossref) [2024-01-08].

[Hanawa et al., 2019] Hanawa, K., Sasaki, S., Ouchi, H., Suzuki, J., and Inui, K. (2019). The sally smedley hyperpartisan news detector at SemEval-2019 task 4. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1057–1061. Association for Computational Linguistics.

[Haouari et al., 2021] Haouari, F., Hasanain, M., Suwaileh, R., and Elsayed, T. (2021). ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 72–81, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

[Hasanain and Elsayed, 2020] Hasanain, M. and Elsayed, T. (2020). bigir at checkthat! 2020: Multilingual bert for ranking arabic tweets by check-worthiness. In *CLEF (Working Notes)*.

[Hasanain and Elsayed, 2022] Hasanain, M. and Elsayed, T. (2022). Cross-lingual transfer learning for check-worthy claim identification over twitter. *arXiv preprint arXiv:2211.05087*.

[Hayawi et al., 2023] Hayawi, K., Saha, S., Masud, M. M., Mathew, S. S., and Kaosar, M. (2023). Social media bot detection with deep learning methods: a systematic review. *Neural Computing and Applications*. Read_Status: In Progress Read_Status_Date: 2024-01-31T00:40:32.999Z.

[Hayawi et al., 2022] Hayawi, K., Shahriar, S., Serhani, M., Taleb, I., and Mathew, S. (2022). Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health*, 203:23–30.

[Heidari et al., 2022] Heidari, M., Jones, J. H. J., and Uzuner, O. (2022). Online User Profiling to Detect Social Bots on Twitter. Publisher: arXiv Version Number: 1.

[Herzig et al., 2020] Herzig, J., Nowak, P. K., Müller, T., Piccinno, F., and Eisenschlos, J. M. (2020). Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*.

[Hong et al., 2015] Hong, K., Nenkova, A., March, M. E., Parker, A. P., Verma, R., and Kohler, C. G. (2015). Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls. *Psychiatry Research*, 225(1–2):40–49.

[Honnibal and Montani, 2017] Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

[Howard and Kollanyi, 2016] Howard, P. N. and Kollanyi, B. (2016). Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. Publisher: arXiv Version Number: 1.

[Howard et al., 2016] Howard, P. N., Kollanyi, B., and Woolley, S. (2016). Bots and automation over twitter during the US election. *Computational propaganda project: Working paper series*, 21(8).

[Hrckova et al., 2021] Hrckova, A., Moro, R., Srba, I., and Bielikova, M. (2021). Quantitative and qualitative analysis of linking patterns of mainstream and partisan online news media in central europe. 46(5):954–973. Publisher: Emerald Publishing Limited.

[Hu et al., 2022] Hu, X., Guo, Z., Wu, G., Liu, A., Wen, L., and Yu, P. (2022). CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

[Huang et al., 2019] Huang, B., Huan, Y., Xu, L. D., Zheng, L., and Zou, Z. (2019). Automated trading systems statistical and machine learning methods and hardware implementation: a survey. *Enterprise Information Systems*, 13(1):132–144. ECC: 0000021 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17517575.2018.1493145.

[Huang and Lee, 2019] Huang, G. K. W. and Lee, J. C. (2019). Hyperpartisan news and articles detection using bert and elmo. In *2019 International Conference on Computer and Drone Applications (IConDA)*, pages 29–32.

[Islam et al., 2020] Islam, M. S., Sarkar, T., Khan, S. H., Mostofa Kamal, A.-H., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Amin Chowdhury, K. I., Anwar, K. S., Chughtai, A. A., and

Seale, H. (2020). Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621–1629.

[Iter et al., 2018] Iter, D., Yoon, J., and Jurafsky, D. (2018). Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146, New Orleans, LA. Association for Computational Linguistics.

[Jafari et al., 2021] Jafari, O., Maurya, P., Nagarkar, P., Islam, K. M., and Crushev, C. (2021). A survey on locality sensitive hashing algorithms and their applications. *arXiv preprint arXiv:2102.08942*.

[Jiang et al., 2019] Jiang, Y., Petrak, J., Song, X., Bontcheva, K., and Maynard, D. (2019). Team bertha von suttner at SemEval-2019 task 4: Hyperpartisan news detection using ELMo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Jimmy et al., 2018] Jimmy, Zuccon, G., Palotti, J., Goeuriot, L., and Kelly, L. (2018). Overview of the clef 2018 consumer health search task. In *Conference and Labs of the Evaluation Forum*.

[Joo and Hwang, 2019] Joo, Y. and Hwang, I. (2019). Steve martin at SemEval-2019 task 4: Ensemble learning model for detecting hyperpartisan news. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 990–994. Association for Computational Linguistics.

[Jurafsky and Martin,] Jurafsky, D. and Martin, J. H. *Speech and Language Processing*. 3 (draft) edition. ECC: 0000004.

[Kapantai et al., 2021] Kapantai, E., Christopoulou, A., Berberidis, C., and Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society*, 23(5):1301–1326.

[Kartal and Kutlu, 2022] Kartal, Y. S. and Kutlu, M. (2022). Re-think before you share: A comprehensive study on prioritizing check-worthy claims. *IEEE transactions on computational social systems*, 10(1):362–375.

[Kazemi et al., 2021a] Kazemi, A., Garimella, K., Gaffney, D., and Hale, S. (2021a). Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517, Online. Association for Computational Linguistics.

[Kazemi et al., 2021b] Kazemi, A., Garimella, K., Gaffney, D., and Hale, S. (2021b). Claim matching beyond english to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4504–4517.

[Kazemi et al., 2022] Kazemi, A., Li, Z., Peréz-Rosas, V., Hale, S. A., and Mihalcea, R. (2022). Matching tweets with applicable fact-checks across languages.

[Kearney, 2023] Kearney, M. W. (2023). tweetbotornot2. original-date: 2020-01-10T02:16:36Z.

[Kenyeres and Kovács, 2022] Kenyeres, A. and Kovács, G. (2022). Twitter bot detection using deep learning.

[Khan et al., 2022] Khan, S., Hakak, S., Deepa, N., Prabadevi, B., Dev, K., and Trelova, S. (2022). Detecting covid-19-related fake news using feature extraction. *Frontiers in Public Health*, 9.

[Khaund et al., 2018] Khaund, T., Al-Khateeb, S., Tokdemir, S., and Agarwal, N. (2018). Analyzing Social Bots and Their Coordination During Natural Disasters. In Thomson, R., Dancy, C., Hyder, A., and Bisgin, H., editors, *Social, Cultural, and Behavioral Modeling*, volume 10899, pages 207–212. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

[Kiesel et al., 2019] 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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Kim and Johnson, 2022] Kim, M. Y. and Johnson, K. M. (2022). CLoSE: Contrastive learning of subframe embeddings for political bias classification of news media. 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 2780–2793. International Committee on Computational Linguistics.

[Klopfenstein et al., 2017] Klopfenstein, L. C., Delpriori, S., Malatini, S., and Bogliolo, A. (2017). The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 555–565, Edinburgh United Kingdom. ACM. ECC: 0000199.

[Knauth, 2019a] Knauth, J. (2019a). Language-Agnostic Twitter Bot Detection. In *Proceedings - Natural Language Processing in a Deep Learning World*, pages 550–558. Incoma Ltd., Shoumen, Bulgaria.

[Knauth, 2019b] Knauth, J. (2019b). Orwellian-times at SemEval-2019 task 4: A stylistic and content-based classifier. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 976–980. Association for Computational Linguistics.

[Ko et al., 2023] 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.

[Kolluri et al., 2022] Kolluri, A., Vinton, K., and Murthy, D. (2022). Poxverifi: An information verification system to combat monkeypox misinformation.

[Konstantinovskiy et al., 2021] Konstantinovskiy, L., Price, O., Babakar, M., and Zubiaga, A. (2021). Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital threats: research and practice*, 2(2):1–16.

[Kotonya and Toni, 2020a] Kotonya, N. and Toni, F. (2020a). Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443.

[Kotonya and Toni, 2020b] Kotonya, N. and Toni, F. (2020b). Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

[Kulkarni et al., 2018] Kulkarni, V., Ye, J., Skiena, S., and Wang, W. Y. (2018). Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527. Association for Computational Linguistics.

[Kumari et al., 2022] Kumari, R., Ashok, N., Ghosal, T., and Ekbal, A. (2022). What the fake? probing misinformation detection standing on the shoulder of novelty and emotion. *Inf. Process. Manage.*, 59(1).

[LaValley, 2008] LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18):2395–2399.

[Lee et al., 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

[Lee et al., 2011] Lee, K., Eoff, B., and Caverlee, J. (2011). Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):185–192.

[Leite et al., 2023] Leite, J. A., Razuvayevskaya, O., Bontcheva, K., and Scarton, C. (2023). Detecting misinformation with llm-predicted credibility signals and weak supervision. *arXiv preprint arXiv:2309.07601*.

[Lemaître et al., 2017] Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17):1–5.

[Lewis et al., 2019] 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. *arXiv preprint arXiv:1910.13461*.

[Lewis et al., 2020] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

[Li and Goldwasser, 2019] Li, C. and Goldwasser, D. (2019). Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy. Association for Computational Linguistics.

[Li et al., 2020] Li, Y., Jiang, B., Shu, K., and Liu, H. (2020). Toward a multilingual and multimodal data repository for covid-19 disinformation. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE.

[Lim et al., 2018] Lim, S., Jatowt, A., and Yoshikawa, M. (2018). Understanding characteristics of biased sentences in news articles. In *CIKM Workshops*.

[Lipani et al., 2021] Lipani, A., Losada, D. E., Zuccon, G., and Lupu, M. (2021). Fixed-cost pooling strategies. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1503–1522.

[Liu, 2011] Liu, T.-Y. (2011). *Learning to Rank for Information Retrieval*. Springer Berlin Heidelberg.

[Liu et al., 2019a] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019a). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Liu et al., 2019b] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

[Liu et al., 2022] 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 Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374. Association for Computational Linguistics.

[Lorenz-Spreen et al., 2023] Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., and Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7(1):74–101.

[Losada and Gamallo, 2018] Losada, D. E. and Gamallo, P. (2018). Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, 54:1–24.

[Losada et al., 2021] Losada, D. E., Herrmann, M., and Elsweiler, D. (2021). Cost-effective identification of on-topic search queries using multi-armed bandits. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 645–654, New York, NY, USA. Association for Computing Machinery.

[Losada et al., 2016] Losada, D. E., Parapar, J., and Barreiro, A. (2016). Feeling lucky? multi-armed bandits for ordering judgements in pooling-based evaluation. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, page 1027–1034, New York, NY, USA. Association for Computing Machinery.

[Losada et al., 2017] Losada, D. E., Parapar, J., and Barreiro, A. (2017). Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing amp; Management*, 53(5):1005–1025.

[Losada et al., 2018] Losada, D. E., Parapar, J., and Barreiro, A. (2018). A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation. *Information Fusion*, 39:56–71.

[Luceri et al., 2019] Luceri, L., Deb, A., Badawy, A., and Ferrara, E. (2019). Red Bots Do It Better:Comparative Analysis of Social Bot Partisan Behavior. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1007–1012, San Francisco USA. ACM. 49 citations (Crossref) [2023-12-13].

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

[Madukwe et al., 2020] Madukwe, K., Gao, X., and Xue, B. (2020). In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

[Mahlous and Al-Laith, 2021] Mahlous, A. R. and Al-Laith, A. (2021). Fake news detection in arabic tweets during the covid-19 pandemic. *International Journal of Advanced Computer Science and Applications*, 12(6).

[Malkov and Yashunin, 2018] Malkov, Y. A. and Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

[Martini et al., 2021] Martini, F., Samula, P., Keller, T. R., and Klinger, U. (2021). Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data & Society*, 8(2):205395172110335.

[Matsumoto et al., 2014] Matsumoto, D., Hwang, H. C., and Sandoval, V. A. (2014). Cross-language applicability of linguistic features associated with veracity and deception. *Journal of Police and Criminal Psychology*, 30(4):229–241.

[Mazza et al., 2019] Mazza, M., Cresci, S., Avvenuti, M., Quattrociocchi, W., and Tesconi, M. (2019). RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter. In *Proceedings of the 10th ACM Conference on Web Science*, pages 183–192, Boston Massachusetts USA. ACM.

[McCoy et al., 2018] McCoy, J., Rahman, T., and Somer, M. (2018). Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62(1):16–42.

[Micallef et al., 2022] Micallef, N., Armacost, V., Memon, N., and Patil, S. (2022). True or false: Studying the work practices of professional fact-checkers. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–44.

[Micallef et al., 2020] Micallef, N., He, B., Kumar, S., Ahamad, M., and Memon, N. (2020). The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE.

[Mihaylova et al., 2021] Mihaylova, S., Borisova, I., Chemishanov, D., Hadzhitsanev, P., Hardalov, M., and Nakov, P. (2021). Dips at checkthat! 2021: Verified claim retrieval. In *CLEF (Working Notes)*, pages 558–571.

[Minaee et al., 2021] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

[Mohr et al., 2022] Mohr, I., Wührl, A., and Klinger, R. (2022). CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

[Molina et al., 2021] 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.

[Mubashara et al., 2023] Mubashara, A., Michael, S., Zhijiang, G., Oana, C., Elena, S., and Andreas, V. (2023). Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*.

[Mukherjee and Weikum, 2015] Mukherjee, S. and Weikum, G. (2015). Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM'15. ACM.

[Müller-Budack et al., 2020] Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., and Ewerth, R. (2020). Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 16–25.

[Nabożny et al., 2021] Nabożny, A., Balcerzak, B., Wierzbicki, A., Morzy, M., and Chlabicz, M. (2021). Active annotation in evaluating the credibility of web-based medical information: Guidelines for creating training data sets for machine learning. *JMIR Medical Informatics*, 9(11):e26065.

[Najari et al., 2022] Najari, S., Salehi, M., and Farahbakhsh, R. (2022). GANBOT: a GAN-based framework for social bot detection. *Social Network Analysis and Mining*, 12(1):4. 16 citations (Crossref) [2023-12-04] Read_Status: To Read Read_Status_Date: 2024-01-30T23:46:45.823Z.

[Nakov et al., 2021] Nakov, P., Alam, F., Shaar, S., Da San Martino, G., and Zhang, Y. (2021). A second pandemic? analysis of fake news about covid-19 vaccines in qatar. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1010–1021.

[Nakov et al., 2022] Nakov, P., Da San Martino, G., Alam, F., Shaar, S., Mubarak, H., and Babulkov, N. (2022). Overview of the clef-2022 checkthat! lab task 2 on detecting previously fact-checked claims.

[Naredla and Adedoyin, 2022] Naredla, N. R. and Adedoyin, F. F. (2022). Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights*, 2(1):100064.

[Nations, 2023] Nations, U. (2023). Hate speech, mis- and disinformation.

[Neudert et al., 2017] Neudert, L., Kollanyi, B., and Howard, P. (2017). Junk news and bots during the German parliamentary election: What are German voters sharing over twitter? Publisher: Computational Propaganda Project.

[Nguyen et al., 2019] Nguyen, D.-V., Dang, T., and Nguyen, N. (2019). NLP@UIT at SemEval-2019 task 4: The paparazzo hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 971–975. Association for Computational Linguistics.

[Nguyen et al., 2016] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

[Ni et al., 2023] Ni, Z., Bousquet, C., Vaillant, P., and Jaulent, M.-C. (2023). *Rapid Review on Publicly Available Datasets for Health Misinformation Detection*. IOS Press.

[Nielsen and McConville, 2022] Nielsen, D. S. and McConville, R. (2022). Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

[Orabi et al., 2020] Orabi, M., Mouheb, D., Al Aghbari, Z., and Kamel, I. (2020). Detection of Bots in Social Media: A Systematic Review. *Information Processing & Management*, 57(4):102250. Read_Status: Read Read_Status_Date: 2024-01-28T21:48:38.769Z.

[Otero et al., 2021] Otero, D., Parapar, J., and Barreiro, A. (2021). The wisdom of the rankers: A cost-effective method for building pooled test collections without participant systems. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC ’21, page 672–680, New York, NY, USA. Association for Computing Machinery.

[Paka et al., 2021] Paka, W. S., Bansal, R., Kaushik, A., Sengupta, S., and Chakraborty, T. (2021). Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, 107:107393.

[Palotti et al., 2015] Palotti, J., Hanbury, A., Müller, H., and Kahn, C. E. (2015). How users search and what they search for in the medical domain: Understanding laypeople and experts through query logs. *Information Retrieval Journal*, 19(1–2):189–224.

[Panda and Levitan, 2021] Panda, S. and Levitan, S. I. (2021). Detecting multilingual covid-19 misinformation on social media via contextualized embeddings. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–129.

[Papadopoulou et al., 2019] Papadopoulou, O., Kordopatis-Zilos, G., Zampoglou, M., Papadopoulos, S., and Kompatsiaris, Y. (2019). Brenda starr at SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 924–928. Association for Computational Linguistics.

[Parapar et al., 2013] Parapar, J., Bellogín, A., Castells, P., and Barreiro, (2013). Relevance-based language modelling for recommender systems. *Information Processing amp; Management*, 49(4):966–980.

[Parapar et al., 2019] Parapar, J., Losada, D. E., Presedo-Quindimil, M. A., and Barreiro, A. (2019). Using score distributions to compare statistical significance tests for information retrieval evaluation. *Journal of the Association for Information Science and Technology*, 71(1):98–113.

[Parapar et al., 2014] Parapar, J., Presedo-Quindimil, M. A., and Barreiro, (2014). Score distributions for pseudo relevance feedback. *Information Sciences*, 273:171–181.

[Passaro et al., 2020] Passaro, L., Bondielli, A., Lenci, A., Marcelloni, F., et al. (2020). Unipi-nle at checkthat! 2020: approaching fact checking from a sentence similarity perspective through the lens of transformers. In *CEUR WORKSHOP PROCEEDINGS*, volume 2696. CEUR.

[Pathak et al., 2020] Pathak, A., Shaikh, M. A., and Srihari, R. K. (2020). Self-supervised claim identification for automated fact checking. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 213–227.

[Pathak and Srihari, 2021] Pathak, A. and Srihari, R. K. (2021). Assessing effectiveness of using internal signals for check-worthy claim identification in unlabeled data for automated fact-checking. *arXiv preprint arXiv:2111.01706*.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

[Pennycook et al., 2020] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7):770–780.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[Pew, 2011] Pew, R. C. (2011). Pew research center: Internet, science & tech. Accessed: 05/01/2024.

[Pierri et al., 2020] Pierri, F., Artoni, A., and Ceri, S. (2020). Hoaxitaly: a collection of italian disinformation and fact-checking stories shared on twitter in 2019. *arXiv preprint arXiv:2001.10926*.

[Pikuliak et al., 2023] Pikuliak, M., Srba, I., Moro, R., Hromadka, T., Smolen, T., Melisek, M., Vykopal, I., Simko, J., Podrouzek, J., and Bielikova, M. (2023). Multilingual previously fact-checked claim retrieval. *arXiv preprint arXiv:2305.07991*.

[Piot et al., 2024] Piot, P., Martín-Rodilla, P., and Parapar, J. (2024). Metahate: A dataset for unifying efforts on hate speech detection.

[Pizarro, 2020] Pizarro, J. (2020). Profiling Bots and Fake News Spreaders at PAN'19 and PAN'20 : Bots and Gender Profiling 2019, Profiling Fake News Spreaders on Twitter 2020. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 626–630.

[Poletto et al., 2020] Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

[Potthast et al., 2018] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

[Pérez-Almendros et al., 2019] Pérez-Almendros, C., Espinosa-Anke, L., and Schockaert, S. (2019). Cardiff university at SemEval-2019 task 4: Linguistic features for hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 929–933. Association for Computational Linguistics.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[Radford and Narasimhan, 2018] Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training.

[Raffel et al., 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

[Rangel and Rosso, 2019] Rangel, F. and Rosso, P. (2019). PAN19 Author Profiling: Bots and Gender Profiling.

[Ratkiewicz et al., 2011] Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Flammini, A., and Menczer, F. (2011). Detecting and Tracking Political Abuse in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):297–304.

[Rawte et al., 2023] Rawte, V., Sheth, A., and Das, A. (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

[Recuero et al., 2020] Recuero, R., Soares, F. B., and Gruzd, A. (2020). Hyperpartisanship, disinformation and political conversations on twitter: The brazilian presidential election of 2018. In *Proceedings of the international AAAI conference on Web and social media*, volume 14, pages 569–578.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

[Reuters, 2021] Reuters, I. (2021). Reuters institute digital news report 2021. Accessed: 05/01/2024.

[Rieh, 2001] Rieh, S. Y. (2001). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161.

[Risch et al., 2018] Risch, J., Krebs, E., Löser, A., Riese, A., and Krestel, R. (2018). Fine-grained classification of offensive language. In *Proceedings of GermEval 2018 (co-located with KONVENS)*, pages 38–44.

[Roberts et al., 2020] Roberts, K., Demner-Fushman, D., Voorhees, E., Bedrick, S., and Hersh, W. (2020). Overview of the trec 2020 precision medicine track.

[Rony et al., 2020] Rony, M. M. U., Hoque, E., and Hassan, N. (2020). Claimviz: Visual analytics for identifying and verifying factual claims. In *2020 IEEE Visualization Conference (VIS)*, pages 246–250. IEEE.

[Röttger et al., 2022] Röttger, P., Nozza, D., Bianchi, F., and Hovy, D. (2022). Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Röttger et al., 2021] Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., and Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

[Rouleau and von Ranson, 2011] Rouleau, C. R. and von Ranson, K. M. (2011). Potential risks of pro-eating disorder websites. *Clinical Psychology Review*, 31(4):525–531.

[Ruchansky et al., 2017] Ruchansky, N., Seo, S., and Liu, Y. (2017). CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, Singapore Singapore. ACM.

[Sadouk et al., 2023] Sadouk, H. T., Sebbak, F., and Zekiri, H. E. (2023). Es-vrai at checkthat! 2023: Analyzing checkworthiness in multimodal and multigenre.

[Sánchez-Junquera, 2021] Sánchez-Junquera, J. (2021). On the detection of political and social bias. *Doctoral Symposium on Natural Language Processing from the PLN.net network*.

[Sánchez-Junquera et al., 2021] Sánchez-Junquera, J., Rosso, P., Montes-y Gómez, M., and P. Ponzetto, S. (2021). Masking and transformer-based models for hyperpartisanship detection in news. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*, pages 1244–1251. INCOMA Ltd. Shoumen, BULGARIA.

[Savchev, 2022] Savchev, A. (2022). Ai rational at checkthat! 2022: using transformer models for tweet classification. *Working Notes of CLEF*.

[Sayyadiharikandeh et al., 2020] Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2020). Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 2725–2732, New York, NY, USA. Association for Computing Machinery. Read_Status: To Read Read_Status_Date: 2024-01-30T00:18:17.074Z.

[Scheufele and Krause, 2019] Scheufele, D. A. and Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

[Schlicht et al., 2021] Schlicht, I. B., de Paula, A. F. M., and Rosso, P. (2021). Upv at checkthat! 2021: mitigating cultural differences for identifying multilingual check-worthy claims. *arXiv preprint arXiv:2109.09232*.

[Schlicht et al., 2023] Schlicht, I. B., Fernandez, E., Chulvi, B., and Rosso, P. (2023). Automatic detection of health misinformation: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*.

[Schmidt and Wiegand, 2017] Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

[Sengupta and Pedersen, 2019] Sengupta, S. and Pedersen, T. (2019). Duluth at SemEval-2019 task 4: The pioquinto manterola hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 949–953. Association for Computational Linguistics.

[Sgueo, 2023] Sgueo, G. (2023). *The Design of Digital Democracy*. Springer Textbooks in Law. Springer Nature Switzerland.

[Shaar et al., 2021a] Shaar, S., Alam, F., Da San Martino, G., Nikolov, A., Zaghouani, W., Nakov, P., and Feldman, A. (2021a). Findings of the NLP4IF-2021 shared task on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF@NAACL' 21, Online. Association for Computational Linguistics.

[Shaar et al., 2021b] Shaar, S., Haouari, F., Mansour, W., Hasanain, M., Babulkov, N., Alam, F., Da San Martino, G., Elsayed, T., and Nakov, P. (2021b). Overview of the clef-2021 checkthat! lab task 2 on detecting previously fact-checked claims in tweets and political debates. In *CLEF (Working Notes)*, pages 393–405.

[Shaar et al., 2020a] Shaar, S., Martino, G. D. S., Babulkov, N., and Nakov, P. (2020a). That is a known lie: Detecting previously fact-checked claims. *arXiv preprint arXiv:2005.06058*.

[Shaar et al., 2020b] Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barrón-Cedeno, A., Elsayed, T., Hasanain, M., Suwaileh, R., Haouari, F., Da San Martino, G., et al. (2020b). Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media. *CLEF (Working Notes)*, 2696.

[Shafahi et al., 2016] Shafahi, M., Kempers, L., and Afsarmanesh, H. (2016). Phishing through social bots on Twitter. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3703–3712.

[Shahid et al., 2022] Shahid, W., Li, Y., Staples, D., Amin, G., Hakak, S., and Ghorbani, A. (2022). Are You a Cyborg, Bot or Human?—A Survey on Detecting Fake News Spreaders. *IEEE Access*, 10:27069–27083. Conference Name: IEEE Access Read_Status: Read Read_Status_Date: 2024-01-29T11:42:19.187Z.

[Shao et al., 2018] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787. ECC: 0000462 arXiv: 1707.07592.

[Shi et al., 2019] Shi, P., Zhang, Z., and Choo, K.-K. R. (2019). Detecting Malicious Social Bots Based on Clickstream Sequences. *IEEE Access*, 7:28855–28862. Conference Name: IEEE Access.

[Shliselberg and Dori-Hacohen, 2022] Shliselberg, S.-H. M. and Dori-Hacohen, S. (2022). Riet lab at checkthat! 2022: improving decoder based re-ranking for claim matching. *Working Notes of CLEF*, pages 05–08.

[Singh et al., 2023] Singh, I., Scarton, C., Song, X., and Bontcheva, K. (2023). Finding already debunked narratives via multistage retrieval: Enabling cross-lingual, cross-dataset and zero-shot learning. *arXiv preprint arXiv:2308.05680*.

[Skuczyńska et al., 2021] Skuczyńska, B., Shaar, S., Spenader, J., Nakov, P., et al. (2021). Beasku at checkthat! 2021: fine-tuning sentence bert with triplet loss and limited data. *Faggioli et al.[33]*.

[Smeros et al., 2021] Smeros, P., Castillo, C., and Aberer, K. (2021). Sciclops: Detecting and contextualizing scientific claims for assisting manual fact-checking. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1692–1702.

[Smădu et al., 2023] Smădu, R.-A., Echim, S.-V., Cercel, D.-C., Marin, I., and Pop, F. (2023). From fake to hyperpartisan news detection using domain adaptation.

[Soboroff, 2021] Soboroff, I. (2021). Overview of trec 2021. In *30th Text REtrieval Conference. Gaithersburg, Maryland*.

[Sondhi et al., 2012] Sondhi, P., Vydiswaran, V. G. V., and Zhai, C. (2012). *Reliability Prediction of Webpages in the Medical Domain*, page 219–231. Springer Berlin Heidelberg.

[Srba et al., 2022] Srba, I., Pecher, B., Tomlein, M., Moro, R., Stefancova, E., Simko, J., and Bielikova, M. (2022). Monant medical misinformation dataset: Mapping articles to fact-checked claims. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. ACM.

[Sridharan, 2022] Sridharan, A. (2022). An automated news bias classifier using caenorhabditis elegans inspired recursive feedback network architecture.

[Srivastava et al., 2019] Srivastava, V., Gupta, A., Prakash, D., Sahoo, S. K., R.R, R., and Kim, Y. H. (2019). Vernon-fenwick at SemEval-2019 task 4: Hyperpartisan news detection using lexical and semantic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1078–1082. Association for Computational Linguistics.

[Starbird, 2019] Starbird, K. (2019). Disinformation's spread: bots, trolls and all of us. *Nature*, 571(7766):449–450. Publisher: Nature Publishing Group.

[Stella et al., 2018] Stella, M., Ferrara, E., and De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440. 240 citations (Crossref) [2023-12-13].

[Stieglitz et al., 2017] Stieglitz, S., Brachten, F., Ross, B., and Jung, A.-K. (2017). Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts.

[Suri and Dudeja, 2022] Suri, P. M. and Dudeja, S. (2022). Asatya at checkthat! 2022: multimodal bert for identifying claims in tweets. *Working Notes of CLEF*.

[Suryavardan et al., 2023] Suryavardan, S., Mishra, S., Chakraborty, M., Patwa, P., Rani, A., Chadha, A., Reganti, A., Das, A., Sheth, A., Chinnakotla, M., et al. (2023). Findings of factify 2: multimodal fake news detection. *arXiv preprint arXiv:2307.10475*.

[Szwoch et al., 2022] Szwoch, J., Staszkow, M., Rzepka, R., and Araki, K. (2022). Creation of polish online news corpus for political polarization studies. In Afli, H., Alam, M., Bouamor, H., Casagran, C. B., Boland, C., and Ghannay, S., editors, *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 86–90. European Language Resources Association.

[Tang et al., 2015] Tang, D., Qin, B., and Liu, T. (2015). Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal. Association for Computational Linguistics.

[Tarannum et al., 2022] Tarannum, P., Hasan, M. A., Alam, F., and Noori, S. R. H. (2022). Z-index at checkthat! lab 2022: Check-worthiness identification on tweet text.

[Targ et al., 2016] Targ, S., Almeida, D., and Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

[Tchechmedjiev et al., 2019] Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., and Todorov, K. (2019). Claimskg: A knowledge graph of fact-checked claims. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*, pages 309–324. Springer.

[Toraman et al., 2022] 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 et al., 2023] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[Tran, 2020] Tran, M. (2020). How biased are american media outlets? a framework for presentation bias regression. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4359–4364. IEEE.

[Tucker et al., 2018] Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

[Turing, 1950] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236):433–460.

[Uyheng and Carley, 2020] Uyheng, J. and Carley, K. M. (2020). Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. *Journal of Computational Social Science*, 3(2):445–468. Read_Status: To Read Read_Status_Date: 2024-01-29T21:38:55.796Z.

[Valcarce et al., 2015a] Valcarce, D., Parapar, J., and Barreiro, A. (2015a). A study of priors for relevance-based language modelling of recommender systems. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15. ACM.

[Valcarce et al., 2015b] Valcarce, D., Parapar, J., and Barreiro, (2015b). *A Study of Smoothing Methods for Relevance-Based Language Modelling of Recommender Systems*, page 346–351. Springer International Publishing.

[Valcarce et al., 2016a] Valcarce, D., Parapar, J., and Barreiro, (2016a). *Efficient Pseudo-Relevance Feedback Methods for Collaborative Filtering Recommendation*, page 602–613. Springer International Publishing.

[Valcarce et al., 2016b] Valcarce, D., Parapar, J., and Barreiro, (2016b). Item-based relevance modelling of recommendations for getting rid of long tail products. *Knowledge-Based Systems*, 103:41–51.

[Valcarce et al., 2016c] Valcarce, D., Parapar, J., and Barreiro, (2016c). *Language Models for Collaborative Filtering Neighbourhoods*, page 614–625. Springer International Publishing.

[Varol et al., 2017] Varol, O., Ferrara, E., Davis, C., Menczer, F., and Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):280–289.

[Vasist et al., 2023] Vasist, P. N., Chatterjee, D., and Krishnan, S. (2023). The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Inf Syst Front*, pages 1–26.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Vidgen and Derczynski, 2020] Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

[Vigdor, 2020] Vigdor, N. (2020). Man fatally poisons himself while self-medicating for coronavirus, doctor says.

[Vosoughi et al., 2018] Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151. 3314 citations (Crossref) [2023-12-14].

[Wadden and Lo, 2021] Wadden, D. and Lo, K. (2021). Overview and insights from the sciver shared task on scientific claim verification. *arXiv preprint arXiv:2107.08188*.

[Wang, 2017] Wang, W. Y. (2017). " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

[Wang et al., 2017] Wang, Z., Yan, W., and Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1578–1585. ISSN: 2161-4407.

[Weizenbaum, 1966a] Weizenbaum, J. (1966a). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

[Weizenbaum, 1966b] Weizenbaum, J. (1966b). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45. ECC: 0004196.

[Williams et al., 2018] Williams, A., Nangia, N., and Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference.

[Williams et al., 2021] Williams, E., Rodrigues, P., and Tran, S. (2021). Accenture at checkthat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation.

[Woloszyn et al., 2021] Woloszyn, V., Kobti, J., and Schmitt, V. (2021). Towards automatic green claim detection. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 28–34.

[Wu et al., 2023] Wu, G., Wu, W., Liu, X., Xu, K., Wan, T., and Wang, W. (2023). Cheap-fake detection with llm using prompt engineering. *arXiv preprint arXiv:2306.02776*.

[Wu et al., 2021] Wu, Y., Fang, Y., Shang, S., Jin, J., Wei, L., and Wang, H. (2021). A novel framework for detecting social bots with deep neural networks and active learning. *Knowledge-Based Systems*, 211:106525.

[Yang et al., 2019] Yang, K., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., and Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61. 218 citations (Crossref) [2024-01-21].

[Yang et al., 2020] Yang, K.-C., Varol, O., Hui, P.-M., and Menczer, F. (2020). Scalable and Generalizable Social Bot Detection through Data Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1096–1103. 145 citations (Crossref) [2024-01-21] Read_Status: To Read Read_Status_Date: 2024-01-31T00:39:35.009Z.

[Yang et al., 2023a] Yang, K.-C., Varol, O., Nwala, A. C., Sayyadiharikandeh, M., Ferrara, E., Flammini, A., and Menczer, F. (2023a). Social Bots: Detection and Challenges. arXiv:2312.17423 [cs] Read_Status: Read Read_Status_Date: 2024-01-29T09:23:34.191Z.

[Yang et al., 2022a] Yang, Z., Chen, X., Wang, H., Wang, W., Miao, Z., and Jiang, T. (2022a). A New Joint Approach with Temporal and Profile Information for Social Bot Detection. *Security and Communication Networks*, 2022:1–14. 1 citations (Crossref) [2023-12-04].

[Yang et al., 2023b] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., and Wang, L. (2023b). The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1).

[Yang et al., 2022b] Yang, Z., Ma, J., Chen, H., Lin, H., Luo, Z., and Chang, Y. (2022b). A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. *arXiv preprint arXiv:2209.14642*.

[Yao et al., 2023] Yao, B. M., Shah, A., Sun, L., Cho, J.-H., and Huang, L. (2023). End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

[Yates et al., 2021] Yates, A., Nogueira, R., and Lin, J. (2021). Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

[Yu et al., 2017] Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). Number: 1.

[Zeng et al., 2021] Zeng, X., Abumansour, A. S., and Zubiaga, A. (2021). Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

[Zengin et al., 2021] Zengin, M. S., Kartal, Y. S., and Kutlu, M. (2021). Tobb etu at checkthat! 2021: Data engineering for detecting check-worthy claims. In *CLEF (Working Notes)*, pages 670–680.

[Zhang et al., 2023] Zhang, Y., Tao, Z., Wang, X., and Wang, T. (2023). Ino at factify 2: Structure coherence based multi-modal fact verification. *arXiv preprint arXiv:2303.01510*.

[Zhou et al., 2020] Zhou, X., Mulay, A., Ferrara, E., and Zafarani, R. (2020). Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, CIKM '20. ACM.

[Zhou et al., 2021] Zhou, X., Wu, B., and Fung, P. (2021). Fight for 4230 at checkthat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness. In *CLEF (Working Notes)*, pages 681–692.

Author's contributions

Michele Joshua Maggini was responsible for section 5 and overall revision of the manuscript.
Paloma Piot was responsible for section 1, section 3, subsection 4.3 and subsection 4.4 and overall revision of the manuscript.
Rabiraj Bandyopadhyay was responsible for section 4 and overall revision of the manuscript.
Rafael Martins Frade was responsible for section 2 and overall revision of the manuscript.
Rrubaa Panchendrarajan was responsible for section 2 and overall revision of the manuscript.
Søren Fomsgaard was responsible for section 6 and overall revision of the manuscript.