A Survey on Automatic Credibility Assessment of Textual Credibility Signals in the Era of Large Language Models

IVAN SRBA, Kempelen Institute of Intelligent Technologies, Slovakia
OLESYA RAZUVAYEVSKAYA, The University of Sheffield, UK
JOÃO A. LEITE, The University of Sheffield, UK
ROBERT MORO, Kempelen Institute of Intelligent Technologies, Slovakia
IPEK BARIS SCHLICHT, Deutsche Welle, Germany and Universitat Politècnica de València, Spain
SARA TONELLI, Fondazione Bruno Kessler, Italy
FRANCISCO MORENO GARCÍA, Universidad Politécnica de Madrid, Spain
SANTIAGO BARRIO LOTTMANN, Universidad Politécnica de Madrid, Spain
DENIS TEYSSOU, Agence France-Presse, France
VALENTIN PORCELLINI, Agence France-Presse, France
CAROLINA SCARTON, The University of Sheffield, UK
KALINA BONTCHEVA, The University of Sheffield, UK
MARIA BIELIKOVA, Kempelen Institute of Intelligent Technologies, Slovakia

In the current era of social media and generative AI, an ability to automatically assess the credibility of online social media content is of tremendous importance. Credibility assessment is fundamentally based on aggregating credibility signals, which refer to small units of information, such as content factuality, bias, or a presence of persuasion techniques, into an overall credibility score. Credibility signals provide a more granular, more easily explainable and widely utilizable information in contrast to currently predominant fake news detection, which utilizes various (mostly latent) features. A growing body of research on automatic credibility assessment and detection of credibility signals can be characterized as highly fragmented and lacking mutual interconnections. This issue is even more prominent due to a lack of an up-to-date overview of research works on automatic credibility assessment. In this survey, we provide such systematic and comprehensive literature review of 175 research papers while focusing on textual credibility signals and Natural Language Processing (NLP), which undergoes a significant advancement due to Large Language Models (LLMs). While positioning the NLP research into the context of other multidisciplinary research works, we tackle with approaches for credibility assessment as well as with 9 categories of credibility signals (we provide a thorough analysis for 3 of them, namely: 1) factuality, subjectivity and bias, 2) persuasion techniques and logical fallacies, and 3) claims and veracity). Following the description of the existing methods, datasets and tools, we identify future challenges and opportunities, while paying a specific attention to recent rapid development of generative AI.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → *Social media.*

Additional Key Words and Phrases: Credibility Assessment, Credibility Signals, Natural Language processing, NLP, Literature Survey

Authors' addresses: Ivan Srba, ivan.srba@kinit.sk, Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia; Olesya Razuvayevskaya, o.razuvayevskaya@sheffield.ac.uk, The University of Sheffield, Sheffield, UK; João A. Leite, jaleite1@sheffield.ac.uk, The University of Sheffield, Sheffield, UK; Robert Moro, robert.moro@kinit.sk, Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia; Ipek Baris Schlicht, ipek.baris-schlicht@dw.com, Deutsche Welle, Bonn/Berlin, Germany and Universitat Politècnica de València, Valencia, Spain; Sara Tonelli, satonelli@fbk.eu, Fondazione Bruno Kessler, Trento, Italy; Francisco Moreno García, fran.moreno@upm.es, Universidad Politécnica de Madrid, Madrid, Spain; Santiago Barrio Lottmann, s.barrio@upm.es, Universidad Politécnica de Madrid, Madrid, Spain; Denis Teyssou, Denis.TEYSSOU@afp.com, Agence France-Presse, Paris, France; Valentin Porcellini, Valentin.PORCELLINI@afp.com, Agence France-Presse, Paris, France; Carolina Scarton, c.scarton@sheffield.ac.uk, The University of Sheffield, Sheffield, UK; Kalina Bontcheva, k.bontcheva@sheffield.ac.uk, The University of Sheffield, Sheffield, UK; Maria Bielikova, maria.bielikova@kinit.sk, Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia.

1 INTRODUCTION

Tackling false information (i.e., disinformation and misinformation) has attracted significant attention in recent years from researchers, media professionals, AI/social media practitioners as well as a general public. From the research perspective, a considerable focus emerged notably after the US presidential elections in 2016. Since then, researchers have been addressing false information by means of a wide range of classification task, leveraging AI technologies such as machine learning, natural language processing, computer vision, or social network analysis [269]. The primary stream of research works, commonly denoted as a *fake news detection*, focuses on predicting whether a piece of content (considering different modalities and context) contains false information[1] [274]. Proposed solutions to fake news detection can be broadly categorized into style-based, propagation-based, source-based, and knowledge-based [341]. Style-based methods analyze linguistic and stylistic features of content, propagation-based methods focus on the spread patterns within networks, and source-based approaches assess the credibility of the information's origin. Knowledge-based methods, on the other hand, involve content verification (fact-checking) against reliable sources, such as research studies, encyclopedias, knowledge graphs, or even large language models.

Despite commendable efforts with a clear contribution to protecting societal democratic values, proposed solutions for fake news detection have been recognized to lack (i) sufficient accuracy on out-of-distribution data (mostly due to over- or under-fitting of the detection models) [111], (ii) generalizability (to new topics and emerging techniques for creating and spreading false information) [291] and (iii) explainability (especially in the case of state-of-the-art deep learning models) [1]. All these factors play a crucial role in effectively combating false information.

Since these drawbacks of fake news detection are inherently related to the high-level nature of the task itself, a sibling task of *credibility assessment* has a potential to bring more accurate, better generalizable and explainable solutions as it is rather based on a two step approach. Firstly, a more granular *credibility signals* (sometimes denoted also as *credibility indicators*) are detected. Secondly, such credibility signals are aggregated by a *credibility assessment algorithm* resulting into a single *credibility score*. Typical examples of credibility signals include factuality, subjectivity and bias present in the content, persuasion techniques, logical fallacies or detection of recently emerging machine-generated text. While credibility signals can be analyzed in a piece of online content by a user manually, they provide a real value especially when detected automatically. Such *automatic detection of credibility signals* can range from simple (e.g., rule- or heuristic- based) approaches to complex solutions requiring AI that is capable of in-depth content understanding.

As a result, credibility assessment provides an interesting potential to extend and supplement the existing stream of works on fake news detection. Credibility signals are more flexible and more widely-applicable as a binary/multiclass classification of fake news. Besides their primary usage, to be further aggregated into a credibility score, they can also be utilized and interpreted by a user directly (e.g., as an "information nutrition label" [94]), provide valuable inputs to information retrieval engines or recommender systems in order to prefer content/sources associated with a high level of credibility, or even serve as features in subsequent classification tasks (including fake news detection itself [278]). Study by Lu et al. [181] showed that even if people are influenced by others when judging the veracity of online content, providing accurate AI-based credibility indicators can effectively improve people's ability in detecting misinformation. Credibility assessment and credibility signals are, furthermore, especially useful in cases in which we cannot easily ascertain that something is true or false (on a single dimension) – concept of credibility provides a necessary level of granularity to represent a potentially complex information from multiple perspectives.

[1] Some works go beyond a binary classification and predict multiple classes, such as veracity labels with varying degrees of trustworthiness commonly used by fact-checking organizations: true, mostly true, mixture, mostly false, false.

Credibility signals are also more easily interpretable and naturally support semi-automatic human-centered AI approaches that attempt supporting instead of replacing a human expert or a lay person (individual credibility signals can be automatically detected and their interpretation can be subsequently done by a human). Last but not least, they are more aligned with journalistic/fact-checking workflows for identifying false information.

In contrast to other related research areas (including fake news detection), research works on automatic credibility assessment are mostly carried out in *isolation* (as we show in this survey, there is a lack of works addressing multiple categories of credibility signals at once and, moreover, many research works do not explicitly state that the outcome of their solution can actually serve as a credibility signal). Getting familiar with the existing works on automatic credibility assessment and detection of credibility signals is challenging due to an *ambiguous and inconsistent terminology* used by researchers, a lack of *clear definition of credibility and credibility signals* as well as due to *a missing standardized taxonomy* of various credibility signals' categories. At the same time, automatic credibility assessment can significantly benefit from a *holistic approach*. For example, due to common underlying similarities, there is a big potential of multi-task learning.

While there is a wide range of possible categories of credibility signals, we specifically focus on such signals that relate to textual content and can be automatically detected by *Natural Language Processing (NLP)* methods. This area is currently at the spotlight due to the recent advancements of *Large Language Models (LLMs)*. Their rapid development has a revolutionizing effect on many text classification tasks, automatic credibility assessment and detection of credibility signals not being an exception.

Despite multiple surveys addressing fake news detection (e.g., [1, 127, 274, 296, 341]), to the best of our knowledge, there is *no survey on automatic credibility assessment and detection of credibility signals from the NLP perspective*. In this paper, we therefore provide the first comprehensive and systematic literature overview of credibility signal detection with the focus on NLP and LLMs. The main contributions are as follows:

(1) We provide a necessary background overview of definitions and dimensions of credibility assessment (including various taxonomies of credibility signals used in the existing literature) to setup the common understanding that is currently missing in the existing works.
(2) We analyze and describe a total of 175 NLP works tackling with automatic credibility assessment and detection of credibility signals. We provide an in-depth description of 3 major categories of credibility signals, which have been selected due to a considerable NLP research interest and their recognition as important by end users, namely: (i) factuality, subjectivity and bias; (ii) persuasion techniques and logical fallacies; and (iii) claims and veracity. This analysis is complemented with an overview of additional 6 categories of credibility signals.
(3) We thoroughly position the NLP research on credibility assessment into the context of other modalities as well as multidisciplinary works. Gaps in the existing literature, future challenges and opportunities, considering also the rapid development of generative AI (especially LLMs), are discussed.

Besides the review of published research works, this survey builds on extensive past research activities and acquired knowledge of its authors in the target area. A unique composition of authors consisting of experts not only on the primary NLP area, but also on other related research domains (such as a computer vision, a social network analysis) as well as media professionals, provides a novel holistic perspective on the addressed topic. Such diversity contributes, besides others, to identification of open problems and future challenges, in which an application of research in practice plays a prominent role.

This survey is structured into 10 sections. Section 2 provides definitions of core concepts related to credibility, describes various dimensions of credibility assessment as well as defines the scope of this survey while situating it into the context of other existing surveys. In Section 3, the methodology employed in the survey process is thoroughly described. Section 4 focuses on research papers that address automatic credibility assessment (i.e., approaches addressing both steps — detection of multiple categories of credibility signals and their aggregation into a credibility score). Subsequently, Sections 5-7 address an automatic detection of 3 selected categories of credibility signals. Section 8 supplements this analysis with a brief overview of additional 6 categories of credibility signals, including the complementary perspective of credibility signals addressed in the non-NLP research. In Section 9, we provide an orthogonal analysis of challenges and open problems that characterise the state of the art in this area. Finally, conclusions are drawn in Section 10.

2 BACKGROUND

2.1 Definitions

Credibility is a complex and multidimensional concept which lacks a single, unified and widely-accepted definition. Existing definitions vary especially across various disciplines and fields, which study credibility from different perspectives, such as information science, psychology, marketing or human-computer interaction (HCI) [247]. It is commonly defined as a high-level construct with a help of related concepts, such as believability (which is considered to be roughly a synonym with credibility), trust/trusthworthiness, reliability or their various combinations [15, 241, 247, 264]. Credibility is always tied to a *target entity* (an object of assessment), that can be either a piece of content or a source spreading such content [155]. Due to ambiguity of terminology, in this survey we opt for community-contributed definitions [123] gathered by the *Credible Web Community Group*[2].

Credibility assessment is a process of ascertaining some level/degree of credibility to a target entity. Practically, credibility assessment can be formulated as [241]: (i) a classification problem (i.e., assigning a label from a pre-defined categories, such as a low/high level of credibility); or as (ii) ranking/scoring problem (i.e., assigning a numerical score). In both cases, there are no standard scales adopted yet. Credible Web Community Group distinguishes two kinds of credibility assessment acts [123]:

(1) *Making credibility decisions*, the act of deciding for oneself what to believe, which is often informal, immediate, and unconscious. Doing this incorrectly often leads to being misled, although it may not be practical to do it correctly at all times.
(2) *Credibility analysis*, manually by people or automatically by machines, is gathering, organizing, and analyzing evidence to help people make credibility decisions about a particular information item. A credibility analysis process might produce some kind of report which might itself be called a *credibility assessment* and might include a *credibility score*. In the following text, we will focus especially on this act.

Consequently, a *credibility signal* [123] is a small unit of information used in making a credibility assessment as an evidence. This can be a measurable feature of the information being assessed for credibility, or information about it (metadata), or information about entities which relate to it in various ways, such as the entity who provided it.

Credibility indicator [123] is commonly used interchangeably with credibility signal. In some communities, *credibility signal* is used for inputs to *credibility assessment algorithms* and *credibility indicator* is used for the display features added to the output, to communicate results to a human consumer. In this survey, we prefer to use *credibility signal*,

[2] https://credweb.org/

while we would like to emphasize that in practice many signals can be used directly also as indicators communicated to end users.

Finally, *credibility assessment tools* [123] are software features or applications which perform credibility assessment or help people do so. Such tools can either derive credibility signals, or implement also a *credibility assessment algorithm* to aggregate such signals into a credibility score/credibility indicators.

Credibility and credibility assessment is inherently very close to *false information* (including fake news that intentionally create and spread disinformation) and the task of *fake news detection*. Nevertheless, credibility (as also defined above) is a broader concept and non-credible content does not necessary needs to be false one only. We would like to emphasize that credibility signals have been already recognized in fake news research as a possible solution towards more accurate and explainable methods. The recent work by Grieve and Woodfield [101] delve into the complexities of identifying and analysing disinformation deceptive news. The authors study the way language adapts to different contexts and purposes, arguing that the linguistic choices in deceptive news differ systematically from those in genuine news. The difference in intent – to deceive versus to inform – leads to detectable variations in language use. The authors illustrate this with the famous case of fake news involving Jayson Blair of The New York Times, where false news were less informationally dense and less confident than real ones. The pressure to invent news and the lack of factual grounding influenced writing style, making it less precise and authoritative. The authors propose that this type of register variation could be a broader indicator of fake news. They call for continued research to understand the linguistic patterns of fake news and to develop strategies to combat its spread. Although the above framework is only based on the analysis of one reporter's articles, this approach may be extended to corpora of non-credible/fake news and credible/real news on various topics to try to identify credibility signals/register variation between the two.

To complement this recent case, we can also recall that the idea of analysing the language of disinformation and propaganda has its source in the pioneering work of the German linguist and philologist Viktor Klemperer [156]. Klemperer showed how hate speech and conspiracy theories were used by the Nazis to justify the persecution of Jews. He also showed how these discourses were repeated and amplified by the media and institutions, until they became the norm.

2.2 Dimensions of credibility assessment

Approaches to study credibility. First, the research works on credibility and credibility assessment can be broadly separated into multiple groups following their primary objective [15].

(1) There is a considerable number of (typically highly multidisciplinary) works focusing on definition of *credibility, credibility components or theoretical frameworks*.
(2) The second group corresponds to *human-based approaches (studies)*, in which various aspects of credibility are studied from the perspective of end-users.
(3) Finally, a large group of approaches focuses on *automation of credibility assessment*, either detecting the credibility signals or aggregating them into the credibility indicators. The complexity of such automation can be very diverse, ranging from simple rule- and heuristic-based methods to supervised/unsupervised machine learning. This group also covers research works focusing on building datasets appropriate to create such methods/models; as well as systems (web applications, browser extensions, etc.) built on the top of such automated methods.

Levels of credibility assessment. Following a type of a target entity, which is a subject of credibility assessment, we can distinguish five levels of credibility assessment as follows [15, 241]:

(1) *Content level*. At the content level, the task is to analyze the content attributes typically of a social media post or a news article. It is the most fundamental and important type of assessment, which is commonly used to determine the credibility at other levels (e.g., if a post is assessed to be credible, also a corresponding user/source or an associated topic/event is considered to be credible).
(2) *User/source level*. This level of credibility assessment depends on features extracted from user accounts (e.g., age, education, profile image) and a history of user-generated content. Besides a specific user, it can relate to a broader source (e.g., a news portal, an organization).
(3) *Topic/event level*. At this level, a credibility is assessed by proceeding from a cluster of posts falling under a specific trending or potentially high-impact topic or event (e.g., elections, societal crisis situations).
(4) *Media level*. At the media level, the medium used to communicate and spread information is a target of credibility analysis (e.g., an online social network). This level typically encompasses credibility analysis of authors, spreaders as well as posts themselves.
(5) *Hybrid level*. To optimally utilize the advantages of individual levels and take advantage of high correlation between them, researchers commonly apply a hybrid approach in which the credibility is assessed at multiple levels at the same time.

Taxonomies of credibility signals. Credibility signals, as a core unit in the credibility assessment process, can be categorized by various taxonomies, following different views.

First, many works broadly distinguish between content- and context-based credibility signals [40, 56, 63, 84, 89, 95, 133, 139, 146, 230, 240, 325, 337]. Content-based signals are derived from the content itself, such as partisanship, emotional appeal, or persuasion techniques. On the other hand, *context-based* signals, also referred to as provenance- [91] or meta-information, take into account contextual (e.g., user, source) clues, such as user statistics, location, domain reputation, or the number of shares.

Second, despite many commendable efforts, there is no unified list or categorization of credibility signals. As a result, research works commonly utilize various and very diverse lists of credibility signals. For example, Molina et al. [198] proposed separate lists of credibility signals (or features as they have been denoted by the authors) to distinguish between eight types of online content, such as real news, fabricated news, or satire. Each list grouped signals into 4 main categories roughly corresponding to levels of credibility assessment: (i) message and linguistic, (ii) sources and intentions, (ii) structural, and (iv) network.

An alternative approach was used by the *Credibility Coalition*[3], which organized weekly remote sessions during which participants drafted about 100 credibility signals specifically aimed at the credibility of web pages [337]. They have been later coalesced into 12 major categories, including reader behavior, revenue models, publication metadata, and inbound and outbound references. In 2017, the *Credibility Coalition* formed the *World Wide Web Consortium (W3C) Credible Web Community Group*, which continued in this initiative. It crowdsourced from human experts (researchers as well practitioners) an extensive list containing more than 200 credibility signals, commonly referred to as *W3C signals* [204]. This list introduces a number of signals categorized by target entity (subject) type: claim, text, image, audio, video, article, title, web page, website, aggregation (e.g., RSS), venue, provider, creator, person, or organization. While such a list is currently just an informal draft and has not been standardized yet, it has been adopted and served as a foundation to several research works (e.g., [230, 234]).

[3] https://credibilitycoalition.org/

Finally, the credibility signals differ in their polarity. In the existing works, some signals are formulated as a predictor of credible content (e.g., high factuality), while other signals are formulated as a predictor of non-credible content (e.g., presence of logical flaws). When interpreting or aggregating individual credibility signals (either by an end user or a credibility assessment algorithm), such polarity must be explicitly taken into consideration (e.g., in [163, 234]).

Importance and effects of credibility signals on end users. Due to a broad definition of credibility, as well as rich taxonomies of credibility signals/indicators, a number of human-based studies have analysed the importance of individual credibility signals and their effects on the perceived credibility, in general as well as for specific groups of end users (according to their expertise, age, education etc.).

First, these studies address various content types and platforms, such as social media posts [10, 275], news articles [124, 128, 181], videos and video sharing platforms [91, 130], images [91], search engines [143, 184, 305], or online encyclopedias [85].

At the same time, they address various types of end users. Piccolo et al. [230] perform a human study with two groups of participants, academics/students and random social media users. Rijo and Waldzus [248] looked into how voting patterns and political beliefs influence the way people evaluate information credibility. Jia et al. [141] focused on liberal and conservative users. Other works (e.g., [130]) focus on social media users in general.

Regarding the scope of credibility signals, some works focus on signals provided directly by (social media) platforms [184, 305], some investigate a single credibility signal (e.g., hyper-partisanship in [141]) or a whole spectrum of signals (e.g., 28 signals analysed in [230]). Jia et al. [141] compared credibility signals produced by three various sources – an algorithm, a community, and a third-party fact-checker. Besides content- and context-based signals, Chang et al. [56] considered also an additional group of design signals (consisting of four signals: interaction design, interface design, navigation design, and security settings).

The number of studies comparing the importance of credibility signals on non-English data is rather limited. The work by Gao et al. [95] performs the human study on the importance of credibility signals for the Weibo Chinese misinformation dataset regarding health and safety information.

All these humans studies resulted into valuable findings about which credibility signals are the most important and have the greatest effect on the end-users. Out of 28 credibility signals, Piccolo et al. [230] found that context-based (presence of links, publisher, author) contribute most towards human judgement. Interestingly, different groups of participants exhibit different patterns towards content-based signals. For example, fallacy and partisanship, as well as clickbait title, were ranked high by academics and students only, while social media users did not find this information important. In the human study on e-Health information, Chang et al. [56] found that content signals were most important, accounting for 57.3% of cases, followed by source-related or context-based signals (26.0%). Finally, they found that design-related signals, such as website layouts and overall design are least important and account for 16.7% of decisions. Lin et al. [175] found that the authority and retweet patterns are most important for human participants when deciding on information credibility. Gao et al. [95] showed that the impact of source credentials, such as the authenticity of the authors, on users' credibility assessment was limited. On the other hand, content-based signals, such as objective claims, are considered more insightful for credibility assessment.

2.3 Existing surveys

The majority of the surveys on information credibility focus on specialised domains. Alrubaian et al. [15] conducted a literature review on the credibility of social media information while considering 4 levels of credibility assessment (i.e., post, member, topic/event, hybrid level). The analysed credibility analysis methods were categorised into automated-, human-, and hybrid-based approaches. Automated-based approaches are further subcategorized into machine learning approaches, graph-based semisupervised approaches, and weighted and IR algorithms. Their analysis spans until 2018, and therefore, the automated methods are limited to traditional machine learning (e.g., classifications algorithms like decision trees, logistic regression) and do not include more recent transformer or LLM-based approaches.

Similarly, Qureshi et al. [241] perform an analysis of the credibility detection methods in the social media and microblog domains. The authors adopt the level-based classification of the task proposed by Alrubaian et al. [15]. In addition, they propose a fine-grained framework based on the existing approaches that targets various aspects behind credibility assessment. Their theoretical framework consists of content-, media-, website- and interaction-based assessment of social media information. Within this framework, they identify two main directions of research, namely, *post-level* and *user-level* constructs. The post-level constructs are further divided into the detection of: (i) deceptive and trustful information (rumor and fake news, hoax, spam and phishing and meme detection tasks), (ii) bias and objectivity detection (hyperpartisan, bias and polarisation detection tasks), and (iii) hate speech and offensive language detection tasks. The user-based constructs, on the other hand, are based on: (i) user deception level (bot and suspicious behaviour detection tasks), (ii) user competence (topic-specific expert detection; and opinion credibility detection tasks), and (iii) user authority (post ranking, personality detection, influence and trust propagation).

Another group of the surveys focuses on particular categories of credibility signals, such as, a survey by Nakov et al. [208] that surveys the literature on factuality and biases in social media, or a survey by Panchendrarajan and Zubiaga [216] on claim detection for automated fact checking.

In contrast to these existing surveys, we focus on in-depth analysis of *Natural Language Processing (NLP)* research, with a specific focus on *automatic credibility assessment and detection of individual textual credibility signals* in the context of *online social media content*. This focus is motivated and determined by the recent emergence of large language models, that have significantly influenced the NLP field. The research works on credibility assessment are already adopting this new generation of models, nevertheless, their potential is still far from being explored completely. This survey thus stands on the intersection of credibility assessment (which is an application domain of utmost importance for protecting our society and democratic values) and artificial intelligence (which is a research domain rapidly advancing thanks to many recent discoveries).

2.4 Scope of the survey

Following the existing surveys and recent development in the relevant research field, we define the scope of this survey as follows:

(1) Out of *approaches to study credibility*, this survey specifically tackles *automation of credibility assessment*.
(2) Regarding the *levels of credibility assessment*, the research works falling into the scope of this survey typically address the credibility assessment at the post level, but indirectly also at user/source, topic/event, media and hybrid levels.
(3) Out of various *taxonomies of credibility signals*, the survey covers content- as well as context-based signals that relate to text (while relation to other modalities is briefly documented in Section 8.7). We specifically focus on a selection of such credibility signals that: (i) have attracted a considerable amount of interest in the NLP community, and (ii) have been recognized by the human-based studies as important and impactful on social media users and human experts (e.g., media professionals). Firstly, the survey covers papers that address

automatic credibility assessment with multiple categories of credibility signals at once. Subsequently, it provides an in-depth analysis for three individual categories of credibility signals:

(a) Factuality, subjectivity and bias;
(b) Persuasion techniques and logical fallacies;
(c) Claims and veracity.

For survey completeness, we provide a brief analysis for six additional categories of credibility signals, which are either covered by recent surveys (e.g., machine-generate text, offensive language) or have been addressed by a limited number of works yet:

(a) Machine-generated text;
(b) Text quality;
(c) References and citations;
(d) Clickbaits and title representativeness;
(e) Originality and content reuse;
(f) Offensive language.

Regarding the *time span*, the survey covers the research papers published before May 2024, when the systematic identification of research works was done. The first papers covered by the survey originate in 2004.

3 METHODOLOGY

3.1 Research paper collection process

In order to collect the relevant papers, we adopted the conventional and standardized methodology for literature review called *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)* [292] and specifically its updated version PRISMA 2020 [214]. In addition to being commonly used in literature surveys, PRISMA was also previously used to perform the analysis of the topics related to credibility signals, such as media bias [250]. Figure 1 depicts the standardized PRISMA 2020 flow diagram [107] illustrating individual steps and the number of research papers processed.

Identification – Search queries. In order to conduct a systematic literature search that covers the papers that have a high potential of being relevant, we generated custom search queries for papers tackling with: (i) credibility assessment utilizing multiple categories of credibility signals, as well as (ii) automatic detection of individual categories of credibility signals. Each query combined keywords and logical search operations (*AND*, *OR*) to achieve the highest possible coverage and reduce false positives at the same time. The keywords have been carefully selected to capture inconsistent and ambiguous terminology used in the existing works. The first version of queries was proposed following our past experience and research in the respective area, and iteratively refined to cover additional alternative phrases with the same semantics identified in the research papers. Table 1 represents the search queries employed to acquire the desired list of research publications.

Identification – Databases and registers. The proposed queries were submitted to four large repositories of peer-reviewed scientific publications, namely:

- ACL Anthology[4]
- ScienceDirect[5]

[4] https://aclanthology.org/
[5] https://www.sciencedirect.com/

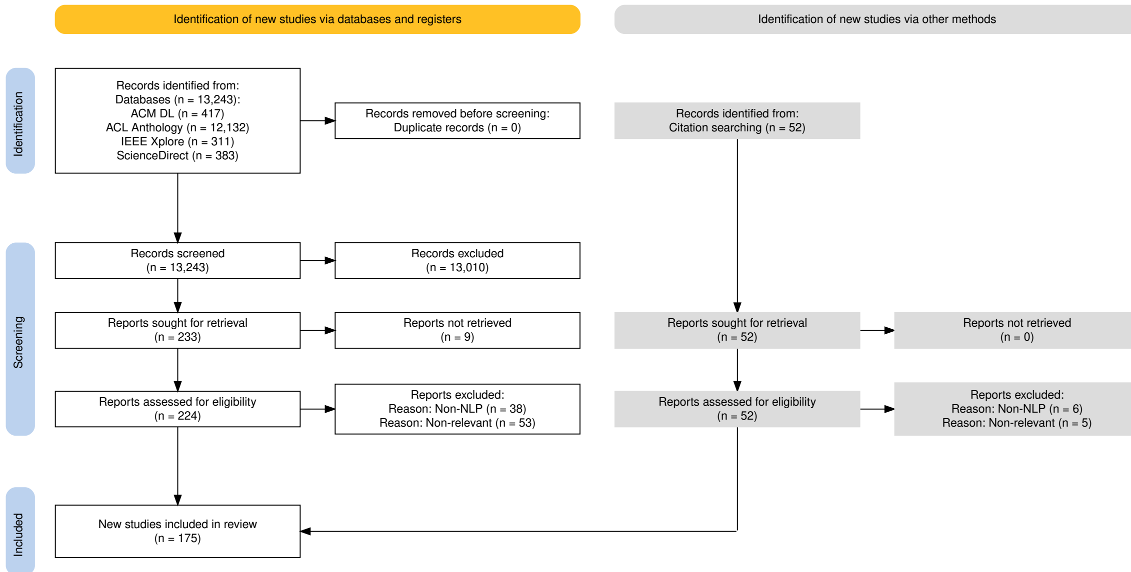

Fig. 1. PRISMA 2020 flow diagram [107] depicting the standardized methodology applied to collect the relevant research papers, together with a number of research papers processed in each step.

- ACM DL[6]
- IEEE Xplore[7]

Identification – Search scope. Papers addressing credibility assessment with multiple categories of credibility signals are quite scarce, therefore, for this type of papers we performed the search for the query terms in the whole body of the document. For individual categories of credibility signals, on the other hand, we searched for the given query term in the title, abstract and user-specified keywords[8] only.

For ACM Digital Library (n = 417), IEEE Xplore (n = 311) and ScienceDirect (n = 383) we considered all returned papers as identified. Since ACL Anthology uses Google Search as a search engine, the search returned an extensive number of results, many of them being irrelevant. By a manual check, we found out that the relevant results appeared in first 10 pages or less, therefore for each search query we considered as identified papers that appear on first 10 pages of results (still resulting in a large number of 12,132 identified papers).

Identification – Extended search. Besides identification of the relevant research papers by the direct search, we applied also citation searching (i.e., search for additional potentially-relevant papers in references from/to the papers identified by the direct search). To some extent, we also checked additional papers published by the same author and at the same venue (e.g., in case of data challenges that are directly focused on the respective category of credibility signals). In this way, we identified additional papers (n = 52), that partially comes from the pre-print servers, especially arXiv. Being a pre-print (a paper that have not underwent the peer-review yet), we carefully checked the quality of such specific cases to make sure that only high-quality papers are included in our survey.

[6]https://dl.acm.org/
[7]https://ieeexplore.ieee.org/
[8]The search in author-specified keywords is available only at the ScienceDirect digital library.

Table 1. Search queries used during the systematic literature search

Credibility signal category	Description	Search query	Search scope
Credibility assessment with multiple categories of credibility signals	Generic credibility assessment approaches utilizing multiple categories of credibility signals are analysed, with no specific focus	"credibility signals" OR "credibility indicators" OR "confidence indicators" OR "non-credibility index" OR "noncredibility index" OR "credibility features"	Title, abstract, keywords, the whole article body
Factuality, subjectivity and bias	Factuality, bias objectivity-reporting, subjectivity, partisanship, tone, emotional valence, framing	("factuality" OR "text subjectivity" OR "text objectivity" OR "subjective bias" OR "sentiment bias" OR "media bias" OR "news framing" OR "partisan") AND (detection OR classification OR identification OR extraction)	Title, abstract, keywords
Persuasion techniques and logical fallacies	Propaganda and persuasion techniques, logical fallacies	("propaganda" OR "persuasion" OR "logical fallacies") AND (detection OR classification OR identification OR extraction)	Title, abstract, keywords
Claims and veracity	Check-worthy claims, previously fact-checked claims	("check-worthy" OR "fact-checked") AND (claim OR statement OR misinformation) AND (detection OR classification OR identification OR extraction)	Title, abstract, keywords
Additional credibility signals	Machine-generated text, Text Quality, References and citations, Clickbaits and title representativeness, Originality and content reuse, Offensive language	Search is done on ad-hoc basis, no systematic keywords were used	Title, abstract, keywords

Screening. Upon identification of the papers that satisfy the search queries (n = 13,243), title and abstract of the papers have been screened directly in the user interface of corresponding digital library. Following this screening, the *non-relevant* papers (n = 13,010) have been excluded. As excluded papers in this step, we considered those that: (i) used the search keywords in a different context that is out of scope this survey as defined in Section 2.4 (i.e., not related to online content credibility and NLP), or (ii) are not full/short/journal/workshop papers (e.g., abstracts of invited talks or workshops).

Out of the remaining relevant papers (n = 233), the full body (PDF file) cannot be retrieved by the authors of this survey only in a negligible number of cases (n = 9).

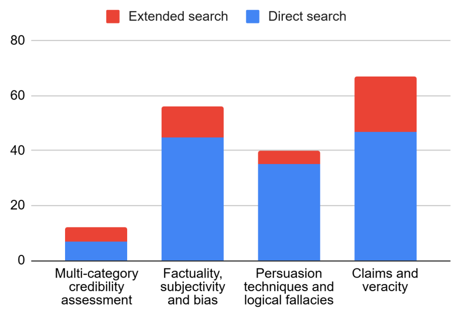

Fig. 2. Number of included papers according the direct (by keywords) and extended (by citations) search grouped by each credibility signal category.

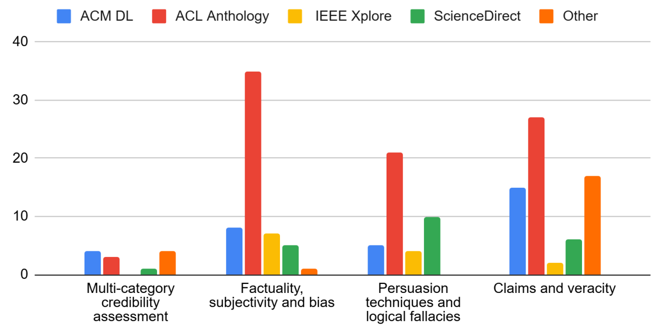

Fig. 3. Number of included papers according to the digital library grouped by each credibility signal category. Other refers to additional repositories, such as arXiv.

For all papers with a full body (n = 224), a final eligibility check was done (to further exclude such papers that have not been already excluded in the first screening step). By in-depth analysis of the full version of the obtained papers, we excluded papers due to two specific reasons:

- Non-NLP (n = 38 for direct search, n = 6 for extended search) – the paper is exploring the general or specific notion of credibility and credibility signals, but from either theoretical, behavioural, human annotation, or other area of computer science (computer vision, social network analysis, etc.).
- Non-relevant (n = 53 for direct search, n = 5 for extended search) – the search terms (co)-occurr in the paper in the context different from the notion of credibility signals. For example, the bi-gram "confidence indicator" can often occur in research on construction risk assessment. Such papers were removed from the analysis.

Included. Finally, we obtained relevant NLP papers (n = 175) that are exploring a specific category or multiple categories of credibility signals while the primary contribution is related to NLP domain (i.e., a paper performs an automatic detection and/or aggregation of credibility signals, introduces a novel dataset or a system for automatic credibility assessment).

Out of included papers, the smallest number of papers tackles with credibility assessment with multiple categories of credibility signals, while the most populated is claims and veracity (see Figure 2). Due to NLP focus, the ACL Anthology is clearly the most common digital library for all three specific categories of credibility signals, while number of papers from individual digital libraries is rather evenly distributed for multi-category credibility assessment papers (see Figure 3). From the time evolution, we can see an initial increase of papers in 2016, corresponding to the overall research focus on false information after the US presidential elections (see Figure 4). A significant increase in recent years, partially slow downed during the COVID-19 pandemic, demonstrates a high research interest in the area addressed in this survey.

3.2 Paper categorization and in-depth analysis

Included papers were subsequently thoroughly analysed by the authors of this survey. Firstly, for each paper the basic metadata have been collected:

- paper type (one of the following options: a full paper, a short paper, a journal paper, a workshop paper, other),
- venue (an abbreviation of a conference, a journal or a pre-print),

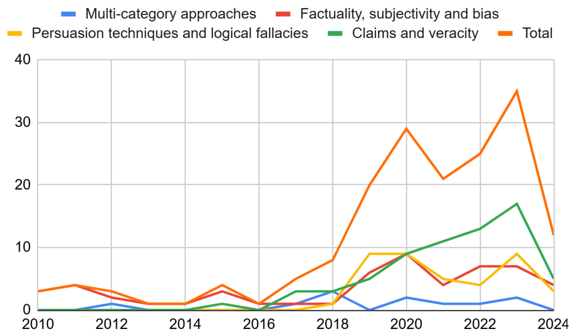

Fig. 4. Evolution of included papers according to year they have been published in. Year 2024 includes papers published online before May 2024.

- year, when the paper was published in,
- main outcome (one of the following options: a method/model, a study, a dataset, a tool, a survey).

Moreover, we annotated several additional types of information that have been identified as the most useful for the analysis and description of the current state of the art, as well as for recognition of remaining open problems and potential future work, namely:

- dataset used (if authors used an existing dataset, a name of such dataset, otherwise information about how own dataset was obtained),
- content/context signals (whether the paper tackles with content, context or both types of credibility signals, see Section 2.2 for more information),
- list of credibility signals (a specific list of signals addressed in the paper),
- automated detection (whether a paper attempts to perform also an automatic credibility assessment and/or automatic detection of credibility signals),
- models (if paper attempts to perform automated detection, what kind of models have been used),
- metrics (if paper attempts to perform automated detection, what kind of metrics have been used),
- human agreement (if paper introduces an annotated dataset, what kind of human agreement was achieved).

The full list of papers included in this survey, including all annotations is available as a digital appendix of this survey[9].

Following this analysis, we describe multi-category credibility assessment papers as well as each individual category of credibility signals from three perspectives (that are also utilized to structure the consequent categories): (i) datasets, (ii) methods and models, and (iii) tools and services. For each category we summarized the current state of the art and identified problems and challenges that are specifically present in such credibility signal category (such analysis is further elaborated in Section 9, with an orthogonal discussion across all included papers).

4 CREDIBILITY ASSESSMENT WITH MULTIPLE CATEGORIES OF CREDIBILITY SIGNALS

This section provides an overview of research works that tackle with credibility assessment by considering multiple categories of credibility signals at once. This group of approaches typically explicitly uses the terminology relevant

[9]Available at: https://docs.google.com/spreadsheets/d/1-6ueO3a_VxgBuiv0QluUY3Un9YEps4YPhtbtESdTaTM/edit?usp=sharing

to credibility (such as credibility assessment, credibility signal, credibility score), in contrast to approaches described in the following sections that tackle with detection of a particular category of credibility signals, but typically do not denote them as such.

4.1 Datasets

The most of datasets used in the credibility assessment works are annotated at the content level (see Section 2.2 for the overview of levels of credibility assessment). Table 2 provides an overview of the selected datasets created or adapted for the purpose of credibility assessment task.

Zhang et al. [337] created a thoroughly annotated, however, only very small credibility-related dataset. Out of news articles, being the most shared on social media, 40 articles were selected (covering multiple topics: public health, climate science, diseases, vaccines). In total, 6 annotators were recruited to annotate the credibility signals. These indicators were adopted from the initial list of signals created by the *Credibility Coalition* (see Section 2.2). Out of them, 16 content and context indicators were selected to be annotated. Namely, content indicators included: title representativeness, clickbait title, quotes from outside experts, citation of organizations and studies, calibration of confidence, logical fallacies, tone, inference. Context indicators included: originality, fact-checked, representative citations, reputation of citations, number of ads, spammy ads, number of social calls, and finally placement of ads and social calls. Besides that, all articles were evaluated for an overall credibility by domain experts on a 5-point scale. In the following dataset analysis, two backward stepwise multiple regression models were utilized to measure the predictive value provided by both sets of indicators. For content-based signals, after model convergence, two variables remained: clickbait title and logical fallacies (slippery slope). This model was found to significantly predict credibility ($F = 13.972$, $p < 0.001$). For context-based signals, 6 variables remained: fact-checked–reported false, fact-checked–reported mixed results, number of social calls, number of mailing list calls, and placement of ads and social calls. Together, they were also found to significantly predict credibility ($F = 12.986$, $p < 0.001$). Despite a wide range of annotated credibility signals as well as overall credibility score, the practical value of the resulting dataset, which was created only as a proof of concept, is somehow limited by its size ($N = 40$) which is insufficient even in few-shot learning scenarios.

In the follow-up study, Bhuiyan et al. [40] gathered a dataset of over 4,000 credibility assessments taken from 2 crowd groups (journalism students and Upwork workers) and 2 expert groups (journalists and scientists). Similarly as in the previous case, a varied set of 50 news articles related to climate science were selected. News articles were annotated by both crowd and expert groups on a 5-point Likert scale. The in-depth analysis of the obtained annotations revealed differences in annotation not only between crowd and expert groups, but also within expert groups between journalists and scientists due to differing expert criteria that journalism versus science experts use. Following the observations, authors proposed directions how to better design crowdsourcing of content credibility.

El Ballouli et al. [84] collected 17 million tweets in a period of two weeks that were written in Arabic and contained any hashtag. After preprocessing and grouping tweets by their hashtags, a topic-independent subset of 9,000 tweets was selected. Each of these tweets was annotated by three annotators on a binary scale (credible, non-credible), while the majority vote was used to determine the final label.

Due to a lack of sufficiently large credibility-annotated datasets, some authors (e.g., [163, 234, 325]) decided to use available fake news datasets, such as LIAR [307], Weibo[10], FakeNewsNet [273], FakeNewsAMT [228] or Celebrity dataset [228]. In such a case, various fake-news labels are mapped to credibility labels by taking an assumption that a fake news content is considered to be non-credible. Qureshi and Malick [240] also proceeded from the FakeNewsNet dataset, however, only to identify a potential set of tweets that were further manually annotated by 12 experts for credibility on a 5-point scale. The resulting dataset consists of 4,958 tweets, out of them some were discarded due to low annotators' agreement.

Another group of datasets used in the existing works relates to credibility annotated at source (webpage) level. Microsoft Credibility [263] dataset consist of top 40 search results on 25 pre-defined queries on the topics of Health, Politics, Finance, Environmental Science, and Celebrity News, which were manually annotated for credibility on a 5-point scale. Another dataset, Content Credibility Corpus (C3) [145], consists of 15,750 evaluations of 5,543 pages by more than 2,000 annotators recruited through Amazon Mechanical Turk.

Finally, at the topic/event level, Mitra and Gilbert [194] introduced a large scale dataset containing 60M tweets collected during a period of more than three months and grouped into 1,049 real-world events, each annotated by 30 human annotators on a 5-point scale.

Table 2. Selected datasets used in the works on credibility assessment with multi-category credibility signals

Dataset	Language	# Instances	Content type	Classes	Used by
Zhang et al. [337]	English	40	News articles	8 content-based signals, 8 context-based signals, overall credibility on a 5-point scale	[234, 337]
Bhuiyan et al. [40]	English	50	News articles	overall credibility on a 5-point scale	[40]
El Ballouli et al. [84]	Arabic	9,000	Tweets	credible, non-credible	[84]
LIAR [307]	English	12,800	Short statements from politifact.com	pants on fire, false barely-true, half-true mostly-true, true	[325]
Weibo[10]	Chinese	18,000	Microblogs	same 6 as in LIAR	[325]
FakeNewsNet [273]	English	23,196	News articles, microblogs	true news, false news	[163, 234]
FakeNewsAMT [228]	English	480	Political news articles	fake, legitimate	[163]
Celebrity [228]	English	500	Celebrity news articles	fake, legitimate	[163]
Qureshi and Malick [240]	English	4,958	Microblogs	overall credibility on a 5-point scale	[240]
Microsoft Credibility [263]	English	1,000	Websites (search results)	overall credibility on a 5-point scale	[89, 234]
Content Credibility Corpus (C3) [145]	English	5,691	Websites	25 signals grouped into 6 categories	[89, 234]
CREDBANK [194]	English	60M	Microblogs	credibility of 1,049 real-world events annotated on a 5-point scale	

4.2 Methods and models

The methods and models used in the research works falling into this group typically employ a methodology consisting of three steps. Firstly, a set of credibility signals is selected. Secondly, various techniques are employed for their automatic detection. Finally, detected credibility signals are then either used to predict an overall credibility which can be subsequently utilized for example as an input to information retrieval engine. In the following description, we provide an in-depth description of each of these three steps.

Credibility signals selection. In the earliest works, credibility signals were rather approximated by various (mostly shallow) linguistic features that can be easily detected by the machine, such as a number of words or exclamation marks. While being potentially helpful content representations for traditional machine learning algorithms, people would not usually use such features for credibility assessment due to the absence of their direct interpretability. Weerkamp and De Rijke [311] adopted a subset of 11 such features for blog posts proposed in [252] while focusing on those that are text-based and can be easily detected automatically, such as *capitalization*, *shouting*, *spelling* or *post length*. Similarly, Kang et al. [146] opted for 12 numeric and 7 binary content-based credibility indicators for Twitter, such as *a positive sentiment factor* or *a number of mentions*. Approximately half of these indicators were adopted from the previous work by Castillo et al. [53].

[10] https://service.account.weibo.com

El Ballouli et al. [84] selected 48 credibility signals, out of them 26 signals were content-based. Besides shallow linguistic features, as already included in the previous works (e.g., *count of hashtags*, *count of unique words*, *count of exclamation/question marks*) the authors utilized also more advanced sentiment extraction to assess *positive sentiment*, *negative sentiment*, and *objectivity*. A shift towards more complex credibility signals is also present in the work by Qureshi and Malick [240]. Authors initially complied a list of 33 potential credibility signals. By proceeding from the initial experiments, a final set of 11 content- and context-based signals were used. In the case of content-based signals, shallow features (such as *a number of hashtags*, *length of text*) were discarded in favor of more complex signals like, *post hate*, *informativeness*, or *deception*.

For the purpose of website credibility evaluation, Esteves et al. [89] selected 15 content-based indicators by proceeding from the taxonomy created by Olteanu et al. [211], such as *authority* (authoritative keywords within the page HTML content), *readability metrics*, *text category* or *sentiment*.

In contrast to works mentioned so far, the following works selected signals from the list created by the Credible Web Community Group (see Section 2.2). At first, Podgurski et al. [234] proposed a modular credibility assessment system of web pages using the 23 most frequent content- and context-based signals. The employed signals range from naive numerical and linguistic features, such as the *number of external links*, *linguistic statistics* (average text and word count, overall word count, etc.), *spelling and grammar*, *domain* endings and *author information*, to more complex features, such as sentiment, readability and presence of a clickbait title. Similarly, Leite et al. [163] employed 18 credibility signals, including advanced ones that have not been considered in the previous works, such as *call to action*, *impoliteness*, *sensationalism* or *explicitly unverified claims*.

Credibility signal detection. Early shallow linguistic features were detected primarily by simple rule- and heuristic-based techniques [146, 311], such as a predefined lexicon of positive and negative words, counting/searching for specific characters/patterns (e.g., capital letters, emoticons).

By proceeding from such superficial linguistic features to more advanced credibility signals, also their detection methods became more complex. For sentiment analysis various lexicon and rule-based sentiment analysis were employed, such as Vader for English [89, 234], or ArSenL for Arabic [84]. For text category classification, well-known approaches employed in other NLP tasks were adopted, such binary multinomial Naive Bayes (NB) classifiers or Latent Semantic Analysis (LSA) [89]. Various NLP libraries were used for automatic detection of additional credibility signals, such as LanguageTool library for detection of grammar & spelling errors or TextBlob for subjectivity detection [234]. Even further, Qureshi and Malick [240] detected the selected credibility signals by means of the classification methods proposed in the previous research works.

Surprisingly, adoption of LLMs for detection of multiple categories of credibility signals is only very rare despite a great potential of LLMs to target a wide range of various credibility signals categories. To this end, Leite et al. [163] used 3 instruction-tuned LLMs (GPT-3.5-Turbo, Alpaca-LoRA-30B and OpenAssistant-LLaMa-30B) with a specific prompt designed for each credibility signals (e.g., "Does the article make use of sensationalist claims?").

Credibility score prediction. To estimate overall credibility from the credibility signals, Weerkamp and De Rijke [311] opted for simple weighting schemata. The resulting credibility score was subsequently incorporated into a blog post retrieval model. Kang et al. [146] proposed three computational models (based on a weighted combination, and on a probabilistic language-based approach) for assessing tweet's credibility, using not only content-based but also social and hybrid strategies. Experiments on the test set consisting of 1,023 instances (while using 10-fold cross-validation) revealed that social model was able to outperform content-based and hybrid model. This can be explained by a short textual content of tweets, as well as very shallow linguistic and possibly noisy credibility signals, such as *a presence of* *exclamation mark*. Podgurski et al. [234] combined individual signals' subscores with signal weights through a linear combination function. Signal weights were determined by authors empirically by considering prior research results, signal measurement accuracy, and experimental calculations on test data.

By advancing from simple weighting and heuristics, consequent works started to employ feature engineering with traditional machine learning approaches. El Ballouli et al. [84] trained a random forest decision tree classifier on the top of detected credibility signals – each serving as a feature. To predict website credibility score, Esteves et al. [89] used Gradient Boosting and AdaBoost classification algorithms on the top of Microsoft Credibility and Content Credibility Corpus (C3).

For the purpose of feature selection, Qureshi and Malick [240] utilized the Kolmogorov–Smirnov (KS) test to identify discriminating features. Subsequently, a wide range of machine learning algorithms (12 regression and 10 classification ones) were used to predict original 5-scale as well as simplified binary credibility score. Results revealed that context-bases signals, such as high user influence and medium-to-high spread count, are very indicative of credible tweets. Credible tweets are normally spread by topic experts (the topic of the tweet is in the top 3 topics for that user), typically do not contain deceptive information and are not posted by malicious accounts. In terms of content-based signals, credible tweets have high informativeness and tent to not contain hate speech or links to low-credible sources.

Analogically to other ML/NLP areas, also credibility assessment underwent a shift towards deep learning methods. In this direction, Wu et al. [325] introduced an ANSP model based on adversarial networks and multi-task learning to capture differential credibility features for information credibility evaluation. The input to the model consists of the concatenations of word embeddings (Word2Vec) and meta-data embeddings (various credibility signals provided by the datasets). Evaluation on English and Chinese datasets, LIAR and Weibo respectively, showed that the combination of all features performs best for both datasets, while content-based features were consistently outperformed by contextual ones. Among these, some signals are particularly powerful on their own, such as the speaker, state info, party affiliation, or credit history.

Finally, Leite et al. [163] used Prompted Weak Supervision (PWS) on the top of LLM-predicted credibility signals. This approach was compared with unsupervised and supervised baselines. In the supervised finetuning scenario, a RoBERTa-Base model was finetuned with the ground-truth labels. In the zero-shot scenario, an instruction-tuned LLM based on LLaMa2 was prompted without any signals or in-context examples. Experimental results on four datasets (FakeNewsNet - GossipCop, FakeNewsNet - PolitiFact, FakeNewsAMT, and Celebrity) showed that Prompted Weak Supervision outperformed the zero-shot baseline by 38.3%, and achieved 86.7% of the performance of the state-of-the-art supervised baseline. Furthermore, in cross-domain settings, where the domain of the train set differs from the domain of the test set (e.g., Politics and Gossip), Prompted Weak Supervision outperformed the supervised baseline by 63%. Lastly, the authors study the association between credibility signals and veracity through (i) a statistical test, where 12 out of the 19 signals were shown to have an association with veracity, and (ii) an ablation study, in which signals were individually removed from the train set in order to inspect their contribution to the model's performance. This showed that the method's strength is in the combination of a wide range of signals rather than relying on a small set of signals that could be strong predictors of veracity. Through these analyses, the authors verified that some credibility signals are domain-specific. For example, signals such as *Source Credibility* and *Misleading about Content* improve performance mostly for the political domain, while others such as *Expert Citation* and *Call to Action* show benefits in entertainment news. In addition to the efficiency of this approach in terms of being independent of the costly annotation of long news articles with each signal, the main advantage of this method is generalisation capability. The authors found that their approach outperforms the fine-tuned methods on the out-of-domain test data. This result is particularly important in the modern era of misinformation, when new fake news emerge every day, making it impossible to have up-to-date annotated data.

4.3 Tools and services

While the systematic paper review have not revealed any specific demo papers describing tools or services dedicated to credibility assessment, various publicly available solutions exist. Many of them are on a border with false information detection. From 2019, the Credibility Coalition maintain a *CredCatalog*[11] – a catalog of initiatives that have a stated aim to improve information quality. Besides various organizations (like fact-checking or academic institutions), it provides overview of several relevant tools.

Besides that, there are commercial projects aiming to help end-users to evaluate credibility of the online content directly in their browsers. NewsGuard[12] provides *News Reliability Ratings* for news outlets based on nonpartisan journalistic criteria. Similarly, GroundNews[13] rate news stories according to their bias distribution (left-/right-leaning bias) or factuality. The labelling is, however, done rather indirectly through sources writing about such news stories, instead of analysing a content of individual news articles themselves.

Finally, an automatic analysis of several credibility signals (persuasion techniques, subjectivity, or machine-generated text among others) is available as a part of the *Assistant* tool in the well-known *Verification plugin*[14] developed and further enhanced as a part of EU-funded projects (*InVID*, *WeVerify* and *vera.ai*). The Assistant tool allows users to provide an URL or a local file as input, from which the text is extracted and various text analysis are performed, credibility signals being a recent addition to them.

4.4 Discussion

Limited research focus with absence of LLM-based solutions. As already shown in Figure 2, the number of works explicitly addressing credibility assessment by taking multiple categories of credibility signals into account is considerably lower in comparison with other individual signal categories. This is in contrast to fake-news detection, which attracted a plethora of research (and also practitioners' and public) attention.

In parallel, our literature survey pointed out a surprising lack of approaches utilizing LLMs. At the same time, LLMs provide a great opportunity to automatize detection of multiple credibility signals even in a zero-shot settings, as the work by Leite et al. [163] clearly demonstrated. Even further improvement in the terms of accuracy can be achieved by employing in-context learning, or instruction-tuning LLMs. While such LLM-based approaches would inquire higher computational costs, existing Parameter-efficient Fine Tuning (PEFT) techniques [327] may employed if necessary.

Lack of multilingual datasets with annotated credibility signals. Similarly to a limited research focus, also the situation with credibility-annotated datasets is falling behind fake-news research. There are very few datasets with manual (human) annotations at the level of individual credibility signals as well as overall credibility [40, 337]. These datasets are, unfortunately, very small preventing employment of language models even in few-shot settings. Despite the declared original intentions to extend them in future, to the best of our knowledge, no follow-up datasets have been created so far.

The remaining datasets are larger (containing hundreds or thousands of instances), nevertheless, annotated only at the level of overall credibility (see Table 2). Absence of datasets providing annotations of multiple categories of

[11] https://credibilitycoalition.org/credcatalog/
[12] https://www.newsguardtech.com/
[13] https://ground.news/
[14] https://www.veraai.eu/category/verification-plugin

credibility signals for the same set of instances, prevents multi-task training of models. Multi-task approaches may provide an interesting potential, especially since many credibility signals inherently share some underlying similarities that can contribute to a better performance (as the results of Prompted Weak Supervision (PWS) already demonstrated in [163]).

Robustness of credibility signals. For reliable credibility assessment, it is necessary that credibility signals and their relation to overall credibility remains the same in time, or at least is not influenced by significant domain and data drifts. To shed more light on the robustness of credibility signals, Ji et al. [139] evaluated how the signals and their prediction capability change over time. By utilizing the posts from the Weibo platform, two topic-specific datasets were created, one for climate change and second one for genetically modified organisms (GMO). The authors analysed how credibility signals evolve over time by splitting these datasets year-wise. They found that certain signals remain more stable over time, however, this stability is topic-dependent. For example, post sentiment performed better for the climate change dataset in most of the years compared with other features, while the post topic was the best predictor for the GMO dataset. In general, results showed that content-based features remained effective across time. On the other hand, contextual features, such as activeness and gender, were not correlated with veracity in the climate change dataset, while popularity showed no significant correlation with GMO misinformation in any given year from 2010 to 2020. In summary, content-based signals (especially sentiment and topic) seem to be more robust and less prone to the temporal data drift compared to context-based signals.

5 FACTUALITY, SUBJECTIVITY AND BIAS

Factuality is a signal reflecting a degree of certainty regarding the possibility of events. The literature on factuality traditionally distinguishes between individual event factuality detection (EFD) [168] and document-level event factuality identification (DEFI) [52] tasks. The main distinction between these two sub-tasks is in the scope of information concerned – while the EFD task considers individual event mentions, DEFI amis to aggregate several mentions of a certain event within a document.

Subjectivity represents an indicator of the overall subjectivity (or objectivity) of information. The task of identifying subjective information is modelled as either a binary or a more fine-grained problem that distinguishes between various degrees of subjectivity. Over the past years, there have been many shared tasks that focus on specific types of subjectivity. For example, Piskorski et al. [232] discriminate between objective reporting, opinionated news and satire. Derczynski et al. [79] and Gorrell et al. [99] organised the challenges for rumour detection. Irony [79] and sarcasm [99] is another common type of non-factual information that received close attention during the last years, with shared tasks and challenges dedicated to this problem.

Finally, *bias* represents a similar but more complex phenomenon of imbalance in terms of opinions or facts. As highlighted by [208], there is no one single concept of bias among scholars. However, in many cases bias is seen as a systematic favouring of a certain ideology when covering information [304]. This can be expressed in deliberately withholding certain part of information that contradict the favourable point of view [279] or vice versa, specifically searching for the facts that cover information from a certain political or ideological viewpoint [122]. In some cases, even if the choice of information sources is unbiased, the presentation of information in those sources may be performed in a biased manner that highlights the importance of only certain facts. Some instances of biased presentation include framing [88, 232], opinionated reporting style [232, 282] and even visual clues [31] used to distort the perception of information.

5.1 Datasets

Tables 3, 4 and 5 represent a summary of the papers that introduce datasets annotated for detecting factuality, bias and subjectivity respectively. As can be seen, the prevalent majority of the datasets are only available in English. Among the non-English for detecting factuality, subjectivity and bias, there are datasets in Urdu [203], Arabic [201] and German [8, 18]. Furthermore, multilingual subjectivity detection was a part of shared tasks at CheckThat! Labs of CLEF 2023 [32] and CLEF 2024 [33], covering Arabic, German, English, Italian, and Turkish languages.

The ontology proposed by Qian et al. [237] is the most widely accepted classification for detecting event factuality, with subsequent extension of the dataset with document-level event factuality annotations [238]. More recently, Li et al. [167] built the first large-scale annotated dataset based on this classification, by using LLM predictions with human judgements as a final step.

Table 3. Datasets for factuality detection

Dataset	Language(s)	# Instances	Classes	Content type
Qian et al. [237]	English, Chinese	1,948 (English) 4649 (Chinese)	CT-: negated events PS+: speculative events PS-: speculative negative events Uu: events appear in question CT+: factual events	News articles from China Daily, Sina Bilingual News, and Sina News. Sentence-level event factuality
Saurí and Pustejovsky [258]	English	9,488 events	Possible, Probable, Certain, Underspecified	TimeBank corpus [235]
Qian et al. [238]	English, Chinese	1,948 (English) 4649 (Chinese)	CT-: negated events PS+: speculative events PS-: speculative negative events Uu: events appear in question CT+: factual events	Extension of Qian et al. [237] with individual and document-level event annotations
Li et al. [167]	English	112,276 events	CT-: negated events PS+: speculative events PS-: speculative negative events Uu: events appear in question CT+: factual events	Documents and events

MPQA opinion corpus [317] is a particularly widespread benchmarking dataset for subjectivity detection, appearing in the majority of studies covered by the systematic review [42, 172, 306, 318]. The dataset is manually annotated with three frames, *objective*, *expressive subjective* and *direct subjective*. The dataset contains span-level annotations along with the annotation of the source (author, specific person, etc.) who expresses the subjective frame and the intensity of subjectivity. Direct subjective expressions are typically more explicit than expressive subjective. Originally created in English for detecting a phrase-level subjectivity, it is widely used in multilingual tasks by utilising parallel translations into other languages, such as Arabic, French, German, Romanian and Spanish [28, 197]. In addition, some of the datasets adopted the ontology of MPQA to create comparable corpora in other languages [18].

In addition to the datasets for bias detection provided in Table 4, Wessel et al. [314] introduced Media Bias Identification Benchmark (MBIB) collection, which is the most comprehensive set of the benchmark corpora for media bias detection consisting of 22 datasets. The types of biases covered include *hate speech*, *lexical*, *contextual*, *linguistic*, *gender*, *cognitive*, *racial* and *political* biases. 3 out of 22 datasets, however, represent a more high-level annotation of information into *fake* and *trustworthy* rather than biases.

Table 4. Datasets for bias detection

Dataset	Language(s)	# Instances	Classes	Content type
Piskorski et al. [232]	English, German, French, Italian, Russian, Polish	1592 news articles	14 presentation frames	News articles
Spinde et al. [289]	English	1,700	Biased vs non-biased	Short statements
Spinde et al. [286]	English	3,700	Biased vs non-biased Word-level bias annotation.	Sentences collected from news organizations with different political leaning.
Liu et al. [177]	English	2,990	**4 general frames**: Politics; Public opinion; Society/Culture Economic consequences **5 issue-specific frames:** Race/Ethnicity 2nd Amendment (Gun Rights); Gun control; Mental health; School/Public space safety	US news articles from 2018 annotated in terms of frames based on headlines only
Fan et al. [90]	English	300 news articles	Informational and Lexical Bias (Sentence, token level)	News articles from FOX, NYT and HPO
Chen et al. [60]	English	6964 news articles	Bias Detection Unfairness Detection (different levels of text granularity)	Articles from 41 publishers Labels derived from AllSides
Aksenov et al. [8]	German	47362 news articles	Fine-grained Bias Detection (different levels of text granularity)	Articles from 36 publishers

5.2 Methods and models

The methods discussed in this section can be categorized into three groups – factuality, subjectivity and bias/framing detection techniques. The methods in the first group can be further divided into two subcategories: (i) those targeting the factuality of individual event mentions and (ii) those assessing the document-level factuality of events.

Event factuality detection (EFD). The methods falling into this category aim to identify the author's level of certainty regarding the possibility of the individual mentioned event [51, 162, 168, 236, 331]. The most common architectures for event factuality detection are LSTM and bi-LSTM models trained on BERT representations [51, 168, 331]. A few studies also employ traditional machine learning approaches, such as SVM with LOSSO regression [162] and a combination of rule-based and maximum entropy methods [236]. All the reviewed methods for event factuality detection are trained on either English [162, 236] or Chinese [168, 331], or a combination of both languages [51].

Document-level event factuality identification (DEFI). Sentence-level event factuality often results in conflicts within a document, as different mentions of the same event may reflect varying degrees of factuality. Therefore, the methods in this sub-group aim to conclude the overall event factuality in a document based on the various sentence-level factuality values within that document. Cao et al. [52] propose an Uncertain Local-to-Global Network (ULGN) that makes use of two important characteristics of event factuality, *local uncertainty* and *global structure*. Similarly, Qian et al. [237] address the challenge of multiple event factuality values within a document by employing an LSTM model trained with both intra- and inter-sequence attention mechanisms to assess document-level event factuality. The model incorporates two types of input features, syntactic and semantic. Syntactic features are based on dependency paths from negative or speculative references to the event, while semantic features are derived from the sentences containing the event. Another approach to estimating document-level factuality is the Sentence-to-Document Inference Network (SDIN) proposed by Zhang et al. [339]. This architecture features a multilayer interaction network that aggregates individual event mentions into a global prediction. The last step employs *gated aggregation* that uses a sigmoid function to generate a mask vector that captures the most critical semantic and factual features of the event mentions. The training process applies a multi-task learning approach, where individual and document-level event factuality prediction tasks share

Table 5. Datasets for subjectivity detection

Dataset	Language(s)	# Instances	Classes	Content type
Piskorski et al. [232]	English, German, French, Italian, Russian, Polish	1592 news articles	Objective, Opinionated, Satire	News articles
Spinde et al. [287]	English	2,800 news articles 175,807 comments and retweets referring to these articles	Hateful vs Neutral	News articles, tweets
Biyani et al. [42]	English	700	Subjective vs Non-subjective	Threads from two popular online forums, Trip Advisor–New York and Ubuntu.
Wiebe et al. [317]	English	10,657 sentences (535 documents)	Objective Expressive subjective Direct subjective	187 different news sources
Spinde et al. [286]	English	3,700	Opinionated, Factual, or Mixed	Sentences collected from news organizations with different political leaning.
Banea et al. [28]	English, Arabic, French, German, Romanian, Spanish	9,700 in each language	Objective vs Subjective	This is an extension of MPQA dataset [315] created by translating sentence-level data into other languages. All information is parallel.
Atalla et al. [18]	German	6,848	Objective vs Subjective	Sentence-level annotation of news articles. Created to be compatible with MPQA.
Wiebe and Riloff [315]	Urdu	500 articles: 700 sentences annotated with emotion and 4,000 unbiased sentences	Objective vs Subjective	News articles from BBC Urdu
Mourad and Darwish [201]	Arabic	2,300	Neutral, Positive, Negative, Both, Sarcastic	Tweets published 2012, randomly sampled.
Jeronimo et al. [138]	Portuguese	450 words	Argumentation, Presupposition Modalization, Sentiment Valuation	Discourse markers for subjectivity in Portuguese
Maks and Vossen [189]	Dutch	11,000–56,000 tokens	Actor and Speaker/Writer subjectivity	Lexicon for subjectivity in Dutch based on Wikipedia articles and user comments
Pang and Lee [217]	English	1,000	Subjective vs Non-subjective	Movie reviews

the same pretrained model and interaction network. As an input, the model receives all the event mentions encoded into BERT representations at the final hidden state of [CLS] token. This approach makes it possible to significantly outperform the models by Cao et al. [52] and Zhang et al. [339] described above on Chinese and English DLEF corpora [237]. Qian et al. [238] propose an approach called Document-level Event Factuality identification via Machine Reading Comprehension Frameworks with Transfer Learning (DEFI-MRC-TL). The authors use BERT as a backbone model to train on a number of large-scale MRC corpora and fine-tuned on the DEFI task. As a target dataset, Qian et al. construct their own MRC-style DEFI corpus called DLEFM by annotating both events and document-level event factuality. They achieve significantly higher results on their dataset than the models by Cao et al. [52] and Zhang et al. [339].

Subjectivity. Subjectivity detection is s well-established NLP task that emerged long before the advent of deep learning and transformer models. As a result, the vast majority of the reviewed methods rely on traditional machine learning models, such as Multinomial NaiveBayes, Convolution Kernels, Support Vector Machines, Logistic regression, Latent Dirichlet Allocation and Decision Trees over lexical, syntactic and semantic features [18, 28, 42, 68, 172, 197, 201, 203, 306, 318]. Rule- and lexicon-based approaches are also common for this task [16, 112, 138, 189, 316]. The only transformer-based method identified through the systematic review is that by Savinova and Del Prado [260]. The authors test the efficiency of RoBERTa model in predicting subjectivity. They found a very high correlation with the human annotators and a significantly better performance that the existing rule-based and regression methods.

Some studies formulate the task of subjectivity detection as a part of the sentiment analysis task, where the "neutral" sentiment corresponds to the non-subjective class, and both "positive" and "negative" sentiments indicate subjective information [57]. Barbosa and Feng [30] explore traditional ML approached to perform the tweet sentiment prediction as a two-step approach. The authors first distinguish between subjective and non-subjective tweets, and then further perform "positive and "negative" tweet prediction for subjective tweets.

Finally, Lu et al. [180] address the task of sarcasm detection, which is often seen as a specialised type of subjective information. The authors propose a novel multimodal Fact-Sentiment Incongruity Combination Network (FSICN) approach for that integrates the factual similarity and the sentiment information into the sarcasm detection task. The FSICN method incorporates a dynamic connection component that identifies the most relevant image-text pairs to detect fact incongruities between them.

Bias. The methods used for the task of bias detection can be categorized based on the type of bias concerned. At a high level, the task can be approached as a binary classification of biased vs non-biased information. The methods falling into this group typically involve traditional machine learning models and lexicon-based methods [4, 110, 161, 289, 328]. Over the last years, transformer models and LLMs became state-of-the-art approaches towards this task [37, 90, 158, 173, 177, 185, 286, 313].

More fine-grained bias analysis methods are predominantly focused on detecting the political leaning bias. Chen et al. [60] explore various levels of textual granularity (word, sentence, paragraph and discourse level) for political bias and unfairness detection. Their main method consists of a recursive neural network and uses GloVe embeddings [227] as an input. According to their analysis on the bias locations, last paragraph typically contains the most biased text segment, while all biased articles start with a neutral tone. Baly et al. [27] model the task of political leaning bias detection jointly with the task of general trustworthiness detection. The authors use multi-task ordinal regression framework as the main model and consider 7 scales of political leaning. They found that each of the two task benefits from the joint training. Sakketou et al. [253] look into the task of detecting user bias. They identify potential spreaders of biased information by applying Graph Attention Networks (GAT) over User2Vec embeddings. The latter is obtained through applying the SentenceBERT (SBERT) [244] over the user's history of posts. The majority of recent approaches for political bias detection are based on transfer learning using transformer models [8, 24–26, 209, 295].

Another group of fine-grained bias detection methods concerns the idea of framing, where frames are seen as perspectives used to discuss the same topic. While some of the frames can represent political leaning, they can also be seen as discussion sub-topics. Liu et al. [177] analyse the dataset of news headlines on the topic of gun violence in the United States. Within this topic, they predict 9 discussion frames, such as mental health, politics, ethnicity, economical consequences, legal regulations, gun rights, public space safety and society. The authors found that BERT model significantly outperforms deep learning aprroaches, such as RNNs, LSTMs, Bi-LSTMs and Bi-GRU, for each of 9 frames.

Finally, Wessel et al. [314] introduce the Media Bias Identification Benchmark (MBIB) task of diverse bias detection, by unifying the existing datasets annotated with gender, political, cognitive, racial, textual and linguistic biases. They benchmark their dataset on ConvBERT, Bart, RoBERTa-Twitter, ELECTRA, and GPT-2 models. The authors found that individual bias detection tasks benefit from different models. For example, ConvBERT is better at predicting political, textual and cognitive bias, while ELECTRA model shows better average performance at racial and gender bias detection based on macro-average F1-scores.

5.3 Tools and services

The in-depth analysis of the research papers covered by this survey revealed several publicly available tools and services used by the NLP methods aimed at bias and factuality analysis.

Media Bias/Fact Check (MBFC) website[15] is a widely used service for media factuality and bias detection. It provides systematic human-centered annotations, spanning over 8 years, regarding the factuality and bias of over 2,000 news websites. The annotators affiliated with International Fact-Checking Network (IFCN) conclude factuality in terms of a score between 0 and 10, which is in turn mapped into a 6-level scale, indicating a "very low", "low", "mixed", "mostly factual", "high" and "very high" rankings. To detect bias, the annotators perform an aggregation of 4 factors: biased wording, factual sourcing, story choice leaning (e.g., pro-liberal or pro-conservative), and political affiliation. Each of these categories is rated by experts on a scale of 0 to 10, from least to most biased. The final score is then calculated as a sum of the four respective scores and is mapped to the categories indicating a range of categories between extreme and moderate left and right bias and unbiased/centered reporting. The original goal of the service was to educate the public on media factuality bias and deceptive news practices. In addition, the API makes the tool highly useful for NLP classification tasks [6, 24, 25, 34, 147, 255].

AllSides Media Bias Ratings[16] is a service similar to MBFC in terms of covering the political leaning bias based on a scale from right to left. The source coverage is also comparable to MBFC, with over 2,400 websited analysed. However, unlike MBFC, the judgement is not limited to a closed group of experts, but also includes ordinary people across the political spectrum trained to perform the analysis. This provides useful insights into how people with certain political leaning biases rate media news without knowing the source. Additionally, there is an option of a community feedback to agree or disagree with the provided rating. The labelled data from AllSides has been used in a number of classification tasks for political bias detection [24, 166, 285, 285, 288, 289].

Ad Fonted Media[17] provides an interactive media bias chart where media sources are annotated based on two different dimensions, *reliability* and *political bias*. The bias dimension consists of most extreme left/right, hyperpartisan left/right, strong left/right, skews left/right and middle or balanced bias. The reliability dimension consists of numerical scores that are mapped to 8 categories: fabricated news, misleading information, selective or incomplete story/propaganda, opinionated of highly varied reliability news, simple fact reporting, and high effort original fact reporting. This resource is also commonly used by NLP approaches to analyse media bias with various degrees of granularity [60, 136].

Finally, The University of Sheffield as a part of their *GATE Cloud infrastructure* [72] provides public tools for multilingual news genre and framing detection based on its submission [323] to 2023 SemEval Shared Task 3 [232]. The models used in these tools achieved competitive results in the shared task, placing in top 3 for most of these languages. The genre detection tool[18] tackles the task of classifying the news into *opinionated*, *objective* and *satire*. The framing detection tool[19] allows the analysis of the 9 principal frames used to present the information.

5.4 Discussion

As noted from the survey results, the tasks of factuality, subjectivity and bias detection cover various subtasks and task formulations. Among the subtasks, while subjectivity detection distinguishes between subjective and non-subjective content, certain formulations also perceive it as a sub-task of sentiment analysis. In turn, under certain formulations,

[15] https://mediabiasfactcheck.com/
[16] https://www.allsides.com/
[17] https://adfontesmedia.com/
[18] https://cloud.gate.ac.uk/shopfront/displayItem/news-genre-classifier
[19] https://cloud.gate.ac.uk/shopfront/displayItem/news-framing-classifier

sentiment detection tasks can be seen as a specific case of bias detection. As an example, distinction between neutral and positive/negative sentiment is a type of subjectivity detection task. In turn, positive and negative sentiment towards a certain topic can be seen as a biased representation of information.

Lack of unified definitions and approaches. One of the main challenges of this category of credibility signals is a high diversity in how researchers perceive and address factuality, subjectivity and bias, resulting in many definitions and ontologies. This problem strengthened by the inherent subjective nature of the factuality/bias is present already in dataset annotation causing also a low inter-annotator agreement [287]. Additionally, the problem of factuality detection is often seen as a sentiment detection [201] or a hate speech detection task [174]. The problem of bias detection is sometimes seen as a specific case of fake news detection, where the biased presentation of facts is automatically seen as a fake information [174]. Finally, factuality, subjectivity and bias is a cumulative notion that is highly nuanced depending on how balanced the opinionated information is and how it is quoted and presented in text [232].

Data scarcity. The biggest challenge in multilingual subjectivity and bias detection is data scarcity. Many English studies are US-centric and only few datasets are available in other languages. For this reason, the multilingual datasets are not well-representative for benchmarking models in global analyses and multicultural settings, since bias detection requires cultural context and background knowledge about a country's political spectrum. For instance, the concepts of right- and left-leaning media can be different across European media and Middle East media. It is important for constructing more comprehensive datasets to benchmark multilingual models.

Focus on binary classification. Despite the explanatory advantage of a fine-grained approach preferred by media professionals, the majority of the models are trained to perform the binary classification of texts into biased and non-biased. This can be, at least partially, a result of the scarcity of datasets available for a fine-grained classification (as illustrated in Table 4).

6 PERSUASION TECHNIQUES AND LOGICAL FALLACIES

Over the past decade, propagandistic efforts have been widely used on social media platforms to shape public opinion and drive engagement on a massive scale. To address this issue, several studies have proposed computational methods aimed at automatically detecting such content [75]. Within the scope of propagandistic content, persuasion techniques and logical fallacies aim to deliberately influence others' opinions using rhetorical and psychological mechanisms [193]. Recently, the task of automatic detection of persuasion techniques has gained increased attention from the NLP research community, with several resources introduced over the years.

6.1 Datasets

The majority of the existing datasets were structured to address two related tasks: (i) classifying the input text as either containing persuasion techniques or not (binary task), and (ii) extracting all specific persuasion techniques present in the input text (multi-label task). The binary task offered a more straightforward approach, but it lacked the granularity necessary for a detailed analysis of propagandistic content. The multi-label task, on the other hand, was a more complex extension of the binary task, as it further required the model to identify the techniques individually, allowing a more fine-grained assessment of propaganda. For this reason, we focus our analysis on the multi-label task. A summary of existing datasets and their key attributes is presented in Table 6.

The NLP4IF-2019 shared task [73] introduced a dataset for fine-grained propaganda detection consisting of news articles from 36 propagandist and 12 non-propagandist news outlets, and annotated with 18 different propaganda techniques. Two subtasks were featured: sentence-level classification (binary task) and fragment-level classification

Table 6. Summary of datasets for detection of persuasion techniques.

Dataset	# PTs	Language(s)	# Instances	Persuasion techniques	Content type
SemEval-2019 [73]	18	English	7,485	Appeal to authority, Appeal to fear/prejudice, Bandwagon, Black-and-white fallacy/dictatorship, Causal oversimplification, Doubt, Exaggeration or minimization, Flag-waving, Loaded language, Name calling or labeling, Obfuscation/intentional vagueness/confusion, Red herring, Reductio ad Hitlerum, Repetition, Slogans, Straw man, Thought-terminating cliché, Whataboutism	News articles
Baisa et al. [23]	18	Czech	7,494	Blaming, Labelling, Argumentation, Emotions, Demonizing, Relativizing, Fear mongering, Fabulation, Opinion, Location, Source, Russia, Expert, Attitude to a politician, Topic, Genre, Focus, Overall sentiment	News articles
Lawson et al. [160]	4	English	90	Authority, Commitment/Consistency, Liking, Scarcity	Emails
PTC (SemEval-2020) [74]	14	English	8,981	Appeal to authority, Appeal to fear/prejudice, Bandwagon/Reductio ad Hitlerum, Black-and-white fallacy, Causal oversimplification, Doubt, Exaggeration/minimization, Flag-waving, Loaded language, Name calling/labeling, Repetition, Slogans, Thought-terminating cliché, Whataboutism/Straw man/Red herring	News articles
SemEval-2021 [81]	22	English	2,488	Appeal to authority, Appeal to (Strong) Emotions, Appeal to Fear or Prejudices, Bandwagon, Black-and-White Fallacy or Dictatorship, Causal Oversimplification, Flag-Waving Doubt, Exaggeration or Minimisation, Slogans, Glittering Generalities (Virtue), Loaded Language, Misrepresentation of Someone's Position (Straw Man), Name Calling or Labeling, Repetition, Obfuscation/Intentional Vagueness/Confusion, Presenting Irrelevant Data (Red Herring), Reductio ad Hitlerum, Thought-Terminating Cliché, Smears, Transfer, Whataboutism	Facebook posts
WANLP-2022 [13]	20	Arabic	1,942	Appeal to authority, Appeal to fear/prejudices, Bandwagon, Black-and-white fallacy/dictatorship, Flag-waving, Doubt Causal oversimplification, Exaggeration/minimisation, Glittering generalities (virtue), Loaded language, Misrepresentation of someone's position (straw man), Name calling or labeling, Repetition, Slogans Obfuscation/intentional vagueness/confusion, Presenting irrelevant data (red herring), Reductio ad hitlerum, Smears, Thought-terminating cliché, Whataboutism	Tweets
Macagno [186]	9	English, Italian, Portuguese	2,657	Straw man, False dichotomy, Ignoring qualifications, Question begging epithets, Post hoc ergo propter hoc, Hasty generalization, Slippery Slope, Persuasive definition, Quasi-definition	Tweets
ArAIEval-2023 [119]	23	Arabic	5,919	Appeal to Authority, Appeal to Fear/Prejudice, Appeal to Hypocrisy, Appeal to Popularity, Appeal to Time, Appeal to Values, Casting Doubt, Causal Oversimplification, Consequential Oversimplification, Conversation Killer, Exaggeration or Minimisation, False Dilemma or No Choice, Flag Waving, Guilt by Association, Loaded Language, Name Calling or Labelling, Red Herring, Repetition Obfuscation/intentional vagueness/confusion, Slogans Questioning the Reputation, Strawman, Whataboutism	News articles Tweets
SemEval-2023 [232]	23	English, French, German, Georgian, Greek, Italian, Polish, Russian, Spanish	49,444	Appeal to Authority, Appeal to Fear/Prejudice, Appeal to Hypocrisy, Appeal to Popularity, Appeal to Time, Appeal to Values, Casting Doubt, Causal Oversimplification, Consequential Oversimplification, Conversation Killer, Exaggeration or Minimisation, False Dilemma or No Choice, Flag Waving, Guilt by Association, Loaded Language, Name Calling or Labelling, Red Herring, Repetition Obfuscation/intentional vagueness/confusion, Questioning the Reputation, Slogans, Strawman, Whataboutism	News articles
Almotairy et al. [14]	15	Arabic	2,100	Flag-waving, Smears, Name-calling, Loaded language, Exaggeration, Whataboutism, Glittering, Doubt, Causal oversimplification, Dictatorship, Appeal to fear, Slogan, Thought-terminating cliché, Appeal to authority, Reductio ad Hitlerum	Tweets

(multi-label span identification task). Similarly, the SemEval-2020 shared task introduced the PTC-SemEval20 dataset [74], which considered the same techniques used in the NLP4IF-2019 task, however, certain techniques were merged or removed due to their low frequency: *Red Herring* and *Straw man* were combined with *Whataboutism*, and *Bandwagon* was merged with *Reductio ad Hitlerum*, while *Obfuscation, Intentional vagueness, Confusion* was removed entirely. In SemEval-2021 [81], a dataset of multimodal persuasion techniques was introduced, containing Facebook posts with images and texts representing memes shared by users from 26 public groups. The visual modality provided additional context that was not present in the text modality alone. For instance, techniques such as *Smears*, *Doubt*, and *Appeal to Fear/Prejudice* appeared more frequently when considering the image along with the text.

Other datasets were introduced for non-English languages. Baisa et al. [23] introduced a dataset for detection of persuasion techniques in the Czech language with around 7, 000 news articles. The techniques considered in their dataset differed substantially from others, with the presence of domain-specific techniques such as ‘Russia’ (indicating that Russia was a topic discussed in the document). WANLP-2022 [13], ArAIEval-2023 [119], and Almotairy et al. [14] introduced datasets for persuasion detection in Arabic texts. The three Arabic datasets share several characteristics such as number of instances (between 2, 000 and 6, 000), content type (mainly tweets, with Hasanain et al. [119] additionally containing news articles), and label scheme (at least 10 techniques are shared between the three datasets).

Macagno [186] introduced the first multilingual dataset for detection of persuasion techniques, containing tweets in English, Italian, and Portuguese. However, the dataset is small (< 3, 000 instances in total) for the purpose of training deep learning models, and similarly to [23], the set of persuasion techniques considered differs substantially from other datasets. SemEval-2023 [232] introduced a large-scale multilingual dataset covering 9 languages (English, French, German, Georgian, Greek, Italian, Polish, Russian, and Spanish), with almost 50, 000 news articles labelled with 23 different persuasion techniques, therefore being the largest dataset currently available in terms of number of instances, languages, and persuasion techniques. Three languages were considered "surprise languages" (Georgian, Greek, and Spanish), for which training sets were not available during the competition, thus encouraging systems to deal with out-of-domain data. Furthermore, the set of 23 persuasion techniques were grouped into 6 coarse-grained categories. For example, the techniques of *Loaded Language*, *Obfuscation, Intentional Vagueness, Confusion*, *Exaggeration or Minimisation*, and *Repetition*, were grouped into the umbrella of *Manipulative Wording*.

6.2 Methods and models

The majority of methods and models developed for the task of automatic detection of persuasion techniques were introduced in the shared tasks discussed in the previous section. The joint effort of multiple different attempts at producing the best-scoring system allows to identify which methodological decisions are key to producing accurate models to detect persuasion. Table 7 summarises the highest-scoring[20] systems across the shared tasks aimed at automatic detection of persuasion techniques.

Transformer-based models were employed in all top-scoring systems for automatic detection of persuasion techniques due to their ability to capture nuanced contextual information [299]. In the NLP4AI-2019 shared task [73], 5 out of the 6 submissions included transformer-based models, while more recently in SemEval-2023 [232], all 16 submissions were comprised of transformer-based models. In earlier shared tasks, BERT was initially preferred over other architectures such as RoBERTa, ALBERT, and DeBERTa, which were adopted more frequently later, specially RoBERTa. In SemEval-2019 BERT was used in all transformer-based submissions. In SemEval-2020 and SemEval-2021, BERT was used in a total

[20] Shared task systems that did not publish a description of their approach are not considered in our analysis.

Table 7. Top-scoring systems for automatic detection of persuasion techniques across different datasets.

System	Dataset	Model	Approach	F1 Micro
newspeak [332]	SemEval-2019 [77]	BERT base uncased	- Token-level classification with 20 classes: No PTs, one of the 18 PTs, and an auxiliary class to handle word-level tokenisation. - Oversampling and class weighting.	0.2488
stalin [83]		GROVER large [336]	- Linear projection of contextual embeddings - SMOTE oversampling - BiLSTM classifier	0.1453
Hitachi [199]	SemEval-2020 [74]	Ensemble (BERT, GTP-2, RoBERTa, XLM XLM-RoBERTa, XLNet)	- BIO encodings - Contextual embeddings with POS tags and named entities. - Three training objectives: (i) BIO tag, (ii) token-level, and (iii) sentence-level classification. - Two BiLSTMs, one for objectives (i) and (ii), and another for (iii). - Class weighting	0.5155
ApplicaAI [144]		RoBERTa large	- Self-training using additional data (500k sentences from OpenWebText). - Added a continional random field (CRF) layer.	0.4915
MinD [294]	SemEval-2021 [81]	Ensemble (BERT, RoBERTa, XLNet, DeBERTa, ALBERT)	- Uses additional data from Da San Martino et al. [77]. - Model ensemble - Custom rules for the Repetition technique. - Character-level n-grams	0.593
Volta [106]		RoBERTa	- Used backtranslation as data augmentation	0.57
NGU CNLP [131]	WANLP-2022 [13]	AraBERT	- Translated the PTC dataset [77] to Arabic and used as additional training data - Stacking-based model ensemble	0.649
IITD [195]		XLM-RoBERTa large	- Simply fine-tuned the model using the task dataset	0.609
UL & UM6P [159]	ArAIEval-2022 [119]	AraBERT-Twitter-v2	- Used an asymmetric multi-label loss objective [245]	0.5666
rematchka [2]		AraBERT-v2	- Class weighting - Balanced data sampler	0.5658
Razuvayevskaya et al. [243]	SemEval-2023 [232]	XLM-RoBERTa large	- Multilingual joint fine-tuning. - LoRA - Class weighting	0.429
KInITVeraAI [126]		XLM-RoBERTa large	- Multilingual joint fine-tuning. - Carefully chosen classification threshold for each language (around 0.2).	0.42
Ampa [221]		XLM-RoBERTa large	- Oversampling - Ensemble with models trained with one and multiple languages.	0.395

of 25 systems, while RoBERTa is used in 14 systems. Nevertheless, systems using RoBERTa achieved the top 2 submissions in both shared tasks [104, 144, 199, 294]. For tasks in Arabic (WANLP-2022 and ArAIEval-2022), monolingual models (AraBERT and variations) outperformed multilingual models (e.g., XLM-RoBERTa and mBERT) [119]. In SemEval-2023 where 9 languages were available, multilingual models (e.g., mBERT and XLM-RoBERTa) outperformed monolingual models trained with each language separately [323]. Unsurprisingly, the larger variations of the models generally outperformed the smaller versions. For reference, RoBERTa large has almost triple the size of RoBERTa base (355 and 125 million parameters, respectively). Also, model ensembles combining different pretrained models were widely and effectively employed, although at the cost of requiring more computational resources [62, 104, 125, 199, 294].

In addition to the contextual embeddings generated by transformer-based models, some works have experimented with supplementary input features such as part-of-speech (POS) tags [152, 199, 251], named entity recognition encodings

[152, 199], word-level n-grams [97, 152, 154], character-level n-grams [62, 294], and sentiment scores [152, 251]. However, apart from the contextual embeddings, it is not clear if other input representation methods play a significant role in improving performance. In fact, apart from Morio et al. [199] who used named entity recognition encodings, all other 1st and 2nd placing systems in shared tasks have not used other supplementary features apart from the contextual embeddings.

Arguably the most relevant methods employed in top-performing systems are aimed towards dealing with data skewness. Most datasets for this task suffer from data skewness, meaning the majority of persuasion techniques are underrepresented, while a small set of persuasion techniques comprise the majority of the dataset. To deal with this issue, several different methods are employed, such as oversampling underrepresented techniques [83, 221, 332], scaling the contribution of the techniques to the loss function according to their proportion (i.e., class weighting) [2, 199, 243, 332], and using supplementary data obtained with (i) data augmentation [106], (ii) semi-supervised methods [144], or (iii) similar datasets [3, 131, 195, 294].

6.3 Tools and services

In terms of production grade tools to detect persuasion techniques in texts, to the best of our knowledge, two services are currently available.

The *Propaganda Persuasion Techniques Analyzer (PRTA)* [76] was designed to detect instances of propaganda in texts by highlighting the spans where specific techniques were used. It provided users with the ability to compare texts based on their use of propaganda techniques, offering detailed statistics on the prevalence of these techniques, both overall and over time. The tool allowed for filtering based on time intervals, keywords, or political orientation of the media. Users could analyse texts either through a dedicated interface or via an API[21]. PRTA used a BERT model trained on the SemEval-2019 dataset, fine-tuned for both fragment-level and sentence-level classification tasks. To gather data, PRTA crawled a growing list of over 250 RSS feeds, Twitter accounts, and websites, extracting text via the Newspaper3k library and performing deduplication using a hash function. The system then identified sentences containing propaganda, organised the articles into topics (such as COVID-19 or Brexit), and allowed users to compare the use of propaganda techniques across various media sources.

The *GATE Cloud* infrastructure [72] operated by the University of Sheffield provides a service to detect 23 different persuasion techniques across multiple languages using a multilingual BERT (mBERT) model[22]. The model was fine-tuned using data from SemEval-2023 in English, French, German, Italian, Polish, and Russian. Class weighting was applied during fine-tuning to account for the label skewness issue. The tool was capable of processing up to $1,200$ documents per day free of charge through its API, with an average processing rate of 2 documents per second. Researchers could request higher quotas if needed for larger-scale analysis.

6.4 Discussion

Most commonly adopted persuasion techniques. Table 6 shows that 7 out of the 10 datasets present a similar set of persuasion techniques [13, 14, 74, 77, 81, 119, 232], while Baisa et al. [23], Lawson et al. [160], and Macagno [186] largely diverge from them. Considering these seven similar datasets, we observe a set of 12 persuasion techniques that are consistent among them: *Appeal to authority*, *Appeal to fear/prejudice*, *Causal oversimplification*, *Doubt*, *Exaggeration or minimization*, *Flag-waving*, *Loaded language*, *Name calling or labeling*, *Slogans*, and *Whataboutism* appear in all seven

[21] https://www.tanbih.org/prta
[22] https://cloud.gate.ac.uk/shopfront/displayItem/persuasion-classifier

datasets, while *Repetition* appears in six datasets (except Almotairy et al. [14]), and *Reductio ad Hitlerum* appears in five datasets (except Hasanain et al. [119], and Piskorski et al. [232]). Combining existing datasets with overlapping persuasion techniques could be a promising research direction to develop more robust models and benchmarking resources. For example, Tian et al. [294] achieved 1st place in SemEval-2021 Task 6 by leveraging the dataset from Da San Martino et al. [77] as additional training data.

Label skewness. A common characteristic inherent to this label scheme is the significant skewness in distribution of the persuasion techniques. For example, in Piskorski et al. [232] (the largest multilingual dataset, with 50,000 instances), a small set of 6 techniques represent 71.8% of the entire dataset - *Loaded Language* (18.5%), *Name Calling-Labelling* (23.7%), *Casting Doubt* (12.5%), *Questioning the Reputation* (7.6%), *Appeal to Fear-Prejudice* (4.8%), and *Exageration/Minimisation* (4.7%) - while the remaining 17 techniques account for only 28.2% of the dataset. The same trend can be seen in Da San Martino et al. [77], with 6 techniques representing 68.2% of the dataset - *Loaded language* (34%), *Name calling, labeling* (17.3%), *Repetition* (10.2%), *Exaggeration, minimization* (7.6%), *Doubt* (7.5%), and *Appeal to fear/prejudice* (4.9%) - while the remaining 12 techniques account for 31.2% of the dataset. This sparse label distribution makes training and evaluating models more challenging, as the majority of persuasion techniques are considerably underrepresented.

Evaluation. Shared tasks on persuasion techniques typically employ the F1-Micro measure as the official evaluation metric [13, 74, 81, 119, 232], however, the F1-Micro does not account for skewed class distributions, as opposed to F1-Macro[23]. This evaluation strategy encourages models that excel at predicting a few densely distributed persuasion techniques. For example, in ArAIEval-2023 [119], the baseline model (majority vote classifier that always predicts the most common persuasion technique), achieves an F1-Micro of 0.3599, and an F1-Macro of 0.0279, representing a difference of 92% between the two scores. Similarly, the best submission by Lamsiyah et al. [159] achieves F1-Micro and F1-Macro scores of 0.5666 and 0.2156, a difference of 62%. Similar figures are seen in other shared tasks such as SemEval-2023 [232], with the average F1-Micro score for first place submissions across all languages resulting in 0.454, and the F1-Macro in 0.211, a difference of 53%. These figures highlight how current state-of-the-art models trained for this task still lack the ability to predict underrepresented persuasion techniques effectively.

Multilinguality. Despite efforts to create multilingual datasets such as SemEval-2023 [232] and Macagno [186], most datasets focus on English or Arabic. For languages without annotated datasets, machine-translation and zero-shot classification approaches may be viable alternatives (if some level of noise introduced by machine translation is acceptable). Nevertheless, joint multilingual training often result in more accurate models. For instance, Razuvayevskaya et al. [243] experimented training monolingual models with translated data from SemEval-2023, which improved performance only for English, but not for other languages. Therefore, multilingual datasets with wider variety of languages are still required to produce state-of-the-art models and evaluation benchmarks.

7 CLAIMS AND VERACITY

Given the impossibility to verify the credibility of every single piece of information posted online, two credibility signals and corresponding tasks are particularly crucial in supporting fact-checkers and other players involved in the fight against disinformation and misinformation [102]: *check-worthiness detection* and the *retrieval of previously fact-checked claims*. The former is aimed at identifying which claims are check-worthy because of their relevance and interest to the general public [121]. The second, which ideally takes place after the first one, is meant to ease claim verification by

[23]Generally F1-Macro is also reported as a secondary metric, but for the purpose of the competition, the F1-Micro determines which system is best.

checking whether the given claim was already verified before by searching in existing repositories of fact-checked claims [266], both monolingual and multilingual.

Outputs of both tasks can also be used (usually in combination) as credibility signals to indicate whether the check-worthy (central) claims contained in a piece of content have been fact-checked and if so, what was the given veracity value or values. They are recognized as such both in the credibility signals list created by the *Credible Web Community Group* (signals 'article has a central claim' and 'fact-check status of a claim'; see [204]) as well as in the related works [337]. The check-worthiness of a claim can be considered a content-based signal, while fact-check status of a claim is a context-based one, since it requires additional external sources to be detected.

7.1 Datasets

Check-worthy claim detection is a popular task within the NLP community thanks to the series of CheckThat! shared tasks organised at CLEF, which led to the release of related datasets. Indeed, check-worthy claim detection is the only subtask that has been proposed at all seven CheckThat! editions [11, 19, 20, 33, 35, 205, 207]. Through those editions, datasets for training check-worthy claim detection models have been constantly extended to cover additional languages, starting from English and Arabic at CheckThat! 2018 [19] to Arabic, Dutch, English and Spanish at CheckThat! 2024 [33] and even multimodal and multi-genre data in the 2023 edition [11]. Beside CheckThat!, additional datasets have been created and made available for research, as shown in Table 8. The sources covered by such datasets are typically social media, news and political debates, i.e. three relevant areas where the presence of disinformation may have detrimental effects on a large audience. Particular relevance was given to COVID-related content during the pandemic [108, 254], when the continuous flow of misleading information could negatively affect public health. Although English is the most represented language, datasets for check-worthiness detection have been created in several other languages, probably because being relevant and of interest to the general public is something that is time- and geographically-bounded, with claims referring to specific events for each specific country or area.

Concerning the retrieval of previously fact-checked claims, there are several datasets containing verified (fact-checked) claims collected from professional fact-checking organizations, such as X-Fact [103], MultiFC [21] or ClaimsKG [293]. Being usually collected for the task of automatic fact-checking, these are not directly usable for the task of retrieval of previously fact-checked claims by themselves, as they lack the input claims (e.g., social media posts) and the pairs of the input and the verified claims. The first datasets designed specifically for the task appeared in 2020 in the works by Shaar et al. [266] and Vo and Lee [301] who independently prepared datasets based on both Snopes and PolitiFact. The former became the basis for a series of CheckThat! Lab shared tasks (Task 2) organised at CLEF in 2020 [35], 2021 [268] and most recently 2022 [206]. The tasks gradually expanded the original dataset in size and also by adding an additional language (Arabic). The political debates part of the dataset was additionally expanded in [267].

The datasets for retrieval of previously fact-checked claims are usually collected by using one of the following approaches: (i) looking into the fact-checking articles for links to the original content making the claim that is being verified, e.g., [35, 206, 231, 266, 268, 277] or (ii) searching for (social media) content (such as discussion threads) that contains URL of or semantic links to the fact-checking articles, e.g. [117, 210, 301]. The former typically achieves high precision at the cost of lower number of pairs and possible missing connections (i.e., many false negatives), while the latter can maximise recall at the cost of introducing noise unless manual checking is applied. Typical example of a high noise are CrowdChecked [117] or MuMiN [210] which both contain a large number of social media posts which distinguishes them from other datasets.

Table 8. Selected datasets used for check-worthiness detection. Note that the datasets used in the different CLEF CheckThat! editions often overlap.

Dataset	Language	# Claims	Content type	Classes
TR-Claim19 [149]	Turkish	2,287	Tweets	Check-worthiness 26 rationale categories
CW-USPD-2016 [96]	English	5,415	Political debates	Check-worthiness
Dhar and Das [80]	Bengali, Hindi	2,402	Political news, Twitter	Check-worthiness
MM-Claims [58]	English	3,400	tweets (image + text)	Claim detection, check-worthiness, visual relevance
Sheikhi et al. [271]	Norwegian	4,885	News	Check-worthiness
AraCOVID19-MFH [108]	Arabic	10,828	COVID-related tweets	Check-worthiness, factual, hateful
Faramarzi et al. [254]	English	7,017	COVID-related tweets	Check-worthiness, claim extraction
ClaimBuster dataset [17]	English	22,281	Presidential debates	Check-worthiness, factual
CLEF-2018 CheckThat! Lab Task 1 [19]	English, Arabic	17,300	Political debates	Check-worthiness
CLEF-2019 CheckThat! Lab Task 1 [20]	English	24,000	Debates, speeches, press conferences	Check-worthiness
CLEF-2020 CheckThat! Lab Task 1 [35]	English, Arabic	962 (EN) 7,500 (A),	Tweets, political debates, speeches	Check-worthiness
CLEF-2021 CheckThat! Lab Task 1 [207]	Arabic, Bulgarian, English, Spanish, Turkish	18,014 (tweets), 50,123 (sentences)	Debates, speeches, tweets	Check-worthiness
CLEF-2022 CheckThat! Lab Task 1 [205]	Arabic, Bulgarian, Dutch, English, Spanish, Turkish	30,363	Tweets	Check-worthiness, verifiable, harmful
CLEF-2023 CheckThat! Lab Task 1 [11]	Arabic, English, Spanish	70,806	Tweets, political debates, speeches	Multimodal and multigenre check-worthiness
CLEF-2024 CheckThat! Lab Task 1 [33]	Arabic, Dutch, English, Spanish	64,700	Tweets, political debates, speeches	Check-worthiness

While many available datasets focus on English only, there are several newer ones supporting other languages (e.g., Arabic [206, 268] or Spanish [191]) or even a range of languages [151, 210, 231, 277]. The most notable among these are MultiClaim [231] and MMTweets [277] due to the number of included languages, high precision of identified pairs and their amount as well as due to the fact that they both introduced a task of crosslingual retrieval, in which input claims are in different language than that of the verified claims.

A full list of selected relevant datasets is reported in Table 9. Besides these, it is also worth mentioning two datasets focusing on COVID-related claims, namely CoAID [70] and MM-COVID [169]; however, these focus on a different task of fake news/disinformation detection. A dataset of COVID-related tweets and fact-checks is also presented in [116, 142]. In this case, although the dataset was introduced for a different task, it could also be useful for previously fact-checked claim retrieval.

7.2 Methods and models

Existing works try to first of all define what *check-worthiness* is. In a recent survey, Panchendrarajan and Zubiaga [216] identify two main aspects making a claim check-worthy: its *verifiability* and its *priority*. The first item refers to the possibility of determining the veracity of a claim, which can be likely supported by evidence. A claim is verifiable when it contains a factual statement that can be checked, which means that personal opinions or events presented as uncertain are excluded. Second, given that verifying all statements about the world is impossible, it is important to prioritize claims which are considered timely, interesting to the general public and whose verification might have a broader impact [78]. In this latter aspect it differs from a related, but a distinct task of *claim detection* which solely aims to identify what constitutes a claim in a text (either using a binary classification or by identification of spans) [105, 196, 326].

Table 9. Selected datasets used for retrieval of previously fact-checked claims

Dataset	Language(s)	# Input claims	# Verified claims	# Pairs	Content type
That is a Known Lie – Snopes [266]	English	1,000	10,396	1,000	social media posts (Twitter)
That is a Known Lie – PolitiFact [266]	English	768	16,636	768	political debates
Snopes (Vo and Lee [301])	English	11,167	1,703	11,202	social media posts (Twitter)
PolitiFact (Vo and Lee [301])	English	2,026	467	2,037	social media posts (Twitter)
Kazemi et al. [151]	English, Hindi, Bengali, Malayalam, Tamil	NA	NA	2,343	instant messages (WhatsApp)
CLEF-2022 CheckThat! Lab Task 2A [206]	English, Arabic	2,518	44,214	2,699	social media posts (Twitter)
CLEF-2022 CheckThat! Lab Task 2B [206]	English	752	20,771	869	political debates and speeches
CrowdChecked [117]	English	316,564	10,340	332,660	social media posts (Twitter)
MuMiN [210]	41 languages	21,565,018	12,914	NA	social media posts (Twitter)
NLI19-SP (FacTeR-Check) [191]	Spanish	40,000	61	NA	social media posts (Twitter)
MultiClaim [231]	posts in 27 languages, fact-checks in 39 languages	28,092	205,751	31,305	social media posts (Twitter, Facebook, Instagram)
MMTweets [277]	posts in 4 languages, fact-checks in 11 languages	1,600	30,452	4,258	social media posts (Twitter)

Given this need for prioritization, check-worthiness detection has been cast as a ranking problem since the first editions of the CheckThat! Lab within the CLEF Evaluation initiative,[24] which made the task well-known within the NLP community, although some related works had already been presented before [121]. In particular, given a political debate, the first CheckThat! task was aimed at predicting which claims should be prioritized for fact checking. This is reflected in the evaluation approach proposed for the CheckThat! series, which has become the *de facto* standard in check-worthiness detection: systems should output the list of input claims ranked by check-worthiness, which is usually evaluated using *Mean Average Precision (MAP)*, reciprocal rank and *P@k* for $k \in \{1, 3, 5, 10, 20, 30\}$. More recently, however, in the shared task check-worthiness has been evaluated using F1 in a binary classification task rather than ranking [11, 33].

Regarding methods, in the CheckThat! editions up to 2023 state-of-the-art results for check-worthy claim detection were reached by methods that rely on fine-tuned transformer-based methods such as BERT, RoBERTa, DistilBERT, [93, 261, 321] and language-specific variants, [11], often combined with data manipulation or model ensembling strategies. For example, the best results on the English portion of the CheckThat 2022 dataset were achieved by a RoBERTa model that leveraged a back-translation-driven data augmentation process [259]. The last edition in 2024, instead, has seen an increased interest in using also generative LLMs for the task such as LLama 2 and 3, GPT3.4 and 5, Mixtral and Mistral [33]. For instance, the best performing system on English fine-tuned Llama2 7b on the original training data, using prompts generated by ChatGPT [171]. Nevertheless, despite the notable progress in data and methods for check-worthy claim detection, there is currently a lack of an extended coverage across languages and topics.

Concerning previously fact-checked claims, the task is also formulated as a retrieval one, although the name of the task may vary across the literature; fact-checking URL recommendation [300], detection of previously fact-checked claims [266], verified claim retrieval [35], searching for fact-checked information [301], claim matching [151] or retrieval of previously fact-checked claims [231] have all been previously used to denote it.

Being a retrieval task, the existing methods apply one or a series of (re-)rankers and *mean average precision (MAP)*, *mean reciprocal rank (MRR)*, *P@k* or a *Success@k* (*Hit@k*) are used as evaluation metrics. In case of a series of rankers, the works tend to use a baseline ranker that is easy to compute and has a good recall and then one or more subsequent

[24] https://www.clef-initiative.eu/

rerankers that work over a progressively smaller subset of results retrieved by a previous ranker in the pipeline, see, e.g., [117, 266]. The rerankers' task is to improve precision by moving the relevant results to the top; since they are working with a smaller set of results, they can be more computationally demanding. Alternatively, the rankers could be used in combination as an ensemble to improve the precision at the cost of higher computational demands, but this is not observed in the surveyed works. In most of the works, BM25 [249] or similar information retrieval algorithms are used as a baseline. Various neural text embedding models are used as either sole rankers (e.g., in [231]), rerankers (e.g., in [266]) or as ensembles [191], especially sentence transformers [244], which use Siamese networks to pre-train text representations, usually on various sentence similarity datasets.

The approaches also use several other techniques to improve the retrieval performance, such as text embedding models fine-tuning [151, 231], distance supervision to work with noisy data [117], key sentences extraction [272], extended context of the input and verified claims (especially for political debates [265]), extraction of text from images [231, 301], using multimodal representation combining text and images [301], or query (input claim) modification or rewriting to be more easily matched with the fact-checked claims [39, 150, 290]. The approaches may further differ by the use of loss, selection of negative examples and other (hyper-)parameters when fine-tuning the neural models. All surveyed approaches (including solutions submitted to the CheckThat! Lab challenge [35, 206, 268]) rely on smaller languages models, such as BERT, XLMRoBERTa, etc. One exception is the work of Sundriyal et al. [290], where the authors employ LLMs to normalize the claims for the purpose of query rewriting for the retrieval task; however, they do not perform experiments on the retrieval task, thus focusing only on the first step (claim detection) in the pipeline.

7.3 Tools and services

Given that automatizing check-worthiness detection can greatly support and speed up fact-checking activities, a number of systems has been already developed for the task, some of which are based on insights and databases actually used by fact-checking organisations.

ClaimBuster [120] was the first end-to-end system for computer-assisted fact-checking trained on a human-labeled dataset of check-worthy factual claims from the U.S. general election debate transcripts. The first component of the pipeline is a detector of check-worthy factual claims which, given a sentence, first labels it as being 'non-factual', 'factual and unimportant' or 'factual and check-worthy'. In case of the latter, a ranking score is assigned based on SVM decision function. Patwari et al. [220] present *Tathya*, a tool focusing only on check-worthiness detection, which compared to ClaimBuster can yield a significant performance improvement, particularly on recall. It is based on a multi-classifier system using features such as topics, entity history and PoS tuples.

ClaimRank [135] performs check-worthy claim detection and supports English and Arabic texts. Its strength is that it was trained on actual annotations from nine reputable fact-checking organizations, therefore mimicking their real strategy for claim selection. The ranking is based on a number of lexical, structural and semantic features, used to train a neural network with two hidden layers as proposed by Gencheva et al. [96]. Another system, focusing specifically on tweets in Arabic, is *Tahaqqaq* [270], which includes the possibility to identify check-worthy claims, estimate the user credibility in terms of spreading fake news, and find authoritative accounts.

dEFEND [71] is another end-to-end system that, given a link to a post or a news, detects check-worthy sentences by assigning them a score, with the goal to distinguish between check-worthy factual claims from subjective ones. The system displays also the propagation network of the text as well as an analysis of the news comments and the textual evidence supporting the classifier decision. Another similar end-to-end platform, providing evidence snippets to credibility classification and check-worthiness decisions, is *BRENDA* [45], which provides also the possibility to collect users' feedback about wrong predictions. The tool is released as Google Chrome extension.

Among the few systems dealing with languages other than English, *FactRank* [38] was the first system able to process check-worthy texts in Dutch. The classification algorithm was developed iteratively, combining expert fact-checker input, a codebook to support reliable human labelling, and active-learning. Check-worthiness classification performance is comparable to results obtained on English with ClaimBuster.

Concerning retrieval of previously fact-checked claims, it is supported by some of the end-to-end verification systems mentioned above, namely ClaimBuster [120], Tahaqqaq [270] or BRENDA [45]. Besides these, *Google Fact-Check Explorer*[25] is often used to perform the task since it indexes a large corpus of fact-checks. Other specialized tools include *Fact-Check Finder*[26] built on models developed in [231].

7.4 Discussion

Multilinguality. We observe in both check-worthiness detection as well as in retrieval of previously fact-checked claims stronger shift towards multilinguality. This is, on one hand, reflected in newer datasets containing also languages other than English (either multilingual ones or focused on a specific language), on the other by the more prevalent use of multilingual models. Since the amount of data in other languages is often limited, approaches for transfer learning [151, 182], low-resource fine-tuning [55] or adapter fusion [262] are explored. However, using translation to English in combination with an English language model can still sometimes outperform a multilingual approach, as was observed, e.g., in the previously fact-checked claim retrieval task [231]. This is likely to change in the future with the employment and/or development of larger and better balanced multilingual models.

Multimodality. Although most available datasets are mostly textual, if they contain links to the original content where the claim was made, it is sometime possible to get to other modalities, such as images, videos or audio, which can be contained in a piece of content (e.g., a social media post). These can be important, because in many cases it is there where the actual claim is being made or the claim requires multiple modalities to be properly interpreted. Multimodal content can also be perceived as more credible by users, has a higher engagement and is increasingly easier to produce [7]. Thus, multimodal approaches continue to be increasingly prevalent as well as important. At the moment, most approaches transcribe the modality to text by either using OCR or image description approaches [190, 231], however there are already some approaches that process the other modalities directly, be it images [301] or speech [132]. The future advances will likely lie in advancements of the latter category of approaches.

Availability of datasets. Although there are available resources for both check-worthiness detection and retrieval of previously fact-checked claims, both tasks have their own (sometimes overlapping) sets of challenges. In case of check-worthiness detection, it is relatively easy to collect check-worthy claims – these are all claims verified by fact-checkers. However, collecting non-check-worthy ones is much more challenging. In case of retrieval of previously fact-checked claims, the challenge lies in collecting input claims (e.g., in the form of social media posts) and in identification of pairs between the input and the fact-checked claims. As discussed above, existing methods either lead to too strict matching with many unidentified (false negative) pairs or to too much noise in the data. Another issue for both tasks is that many datasets were previously built using Twitter. If only the IDs of the tweets have been published, it is now very expensive for researchers to use them due to the X's current API limitations and pricing, thus making their use impractical or completely unfeasible.

[25] https://toolbox.google.com/factcheck/explorer
[26] https://fact-check-finder.kinit.sk

Combination of check-worthy claim detection and retrieval of previously fact-checked claims. As can be seen from the surveyed works, most of them approach the tasks in separation. This is reasonable from the scientific perspective, but more end-to-end (combined) approaches capable of first detecting a check-worthy claim and then retrieving previously fact-checked claims are needed for practitioners to use.

Adoption of LLMs. Although we observe first approaches using LLMs in check-worthiness detection [33], their potential is still not explored in the surveyed works on the retrieval of fact-checked claims. Their benefit could be in input claim normalization [290], using retrieval augmented generation [129], in results reranking or in providing summaries of retrieved fact-checks.

8 ADDITIONAL CREDIBILITY SIGNALS

8.1 Machine-generated text

The task of *machine-generated text* detection (also called synthetic or neural text detection or authorship attribution [297]) has already been researched since about 2018 when the arrival of GPT-1 and transformer architecture have enabled generation of reasonably coherent texts[297]. However, it has been only with the arrival of generative LLMs that the task gained upon practical importance due to increased fidelity and quality of the machine-generated texts in English [298], but also in many other languages [187, 309] and due to low costs of such generation. Since then, there has been recognised potential of the machine-generated texts to be misused for influence operations [98], disinformation [49, 183, 303], spam or unethical authorship [69]. Although there have been voices tempering down fears of massive misuse of LLMs for disinformation generation [276], actual field studies are missing and the rapid advancement of the new models makes the misuse of LLMs to generate and/or amplify disinformation increasingly easier. Nevertheless, when considering using machine-generated text detection as a credibility signal, it has to be noted that there are also many benign and legitimate uses (e.g., machine translation, use of LLMs to improve stylistics or grammar, etc.), so the context of such a use needs to be taken into account as well.

Due to increased interest in the topic, there have been lately several new datasets released extending the task to a range of languages and domains [187, 309], often supported by specialized data challenges and tasks, such as SemEval-2024 Task 8 [308]. As to the detection methods and models, these range from statistical, e.g., Binoculars [114], Fast-DetectGPT [29], etc. to fine-tuned language models, such as Longformer [36, 170], RoBERTa [280] or MDeBERTa [187], but also increasingly including fine-tuned LLMs or their combinations [284].

Despite the undeniable progress in the task, there are several remaining challenges, such as robustness of detectors to the out of distribution data and adversarial attacks (such as authorship obfuscation [188]), their interpretability and explainability or detection of not purely generated texts [137, 297, 324].

8.2 Text quality

Text quality is a broad category of credibility signals measuring text's linguistic accuracy, such as readability, grammatical correctness, or spelling mistakes. It is strongly related to perceived credibility, since high-quality, more professional, content is often seen as more trustworthy. Research on statin-related websites found that more readable and accurate information significantly improves users' perceptions of credibility [176]. A similar study by Kiili et al. [153] highlighted that professionalism in text, such as proper grammar and clear structure, plays a crucial role in how credibility is judged.

The similar relation to credibility can be observed for a low-quality content. Harris [118] showed that poor grammar or frequent spelling errors can be a signal of lack of credibility, prompting readers to question the reliability of the information presented. Greškovičová et al. [100] investigated how various editorial elements such as superlatives, clickbaits, boldface and poor grammar affect the quality and credibility of online messages. Thus, the quality of the text directly enhances trust in it.

Various NLP techniques have been developed to rate text quality and, by extension, its credibility. Mosquera and Moreda [200] evaluates text by extracting features like contractions, slang, misspellings, emoticons and readability (using the Readability Index). It also measures information content through entropy and emotional tone using emotional distance. Finally, it applies the Expectation-Maximization (EM) algorithm to cluster texts by informality levels. The tool obtained a F1 score of 60.6%.

In Pitler and Nenkova [233], the model combines lexical, syntactic, and discourse features to predict human judgments of text readability. It evaluates vocabulary difficulty using unigram language models, syntactic complexity by measuring parse tree height, and discourse relations through annotated markers. Additionally, entity coherence is examined by analyzing semantic continuity between sentences. When all features are used to feed a linear regression, the accuracy on readability results in 88.88% . The model proposed by Mesgar and Strube [192] uses neural networks to evaluate the quality of the text. It captures semantic connections between adjacent sentences by representing sentences as vectors and identifying the most similar states between them. The model uses Recurrent Neural Networks to account for word context within sentences and a Convolutional Neural Network to identify patterns of coherence across the text. This allows the analysis of sentence-to-sentence transitions to predict the readability with a 97.77% accuracy.

For the purpose of text quality classification, datasets of various content types originating in a different sources have been used so far. To name a few, Mosquera and Moreda [200] used a dataset of 50 Yahoo! Answers posts, rated by 6 people in 4 informality level. Pitler and Nenkova [233] utilized 30 Wall Street Journal articles from the Penn Treebank, rated by college students for readability; and Mesgar and Strube [192] used a dataset of 105 texts human labeled from the British National Corpus and Wikipedia.

8.3 References and citations

When speaking of credibility, elements such as *references*, *citations* and partially also *quotes* can influence the confidence of the user in the content and naturally represent important credibility signals [5, 22, 219, 337]. It should be noted that the role and influence of references or citations is different to the influence of the quotes although they can occasionally have a similar role [22, 337].

References and citations are traditionally analysed in the context of scientific articles, where they give support by acknowledging prior work and providing a knowledge basis, and thus enhancing their reliability and supporting the validity of their findings [22, 337]. Commonly, explicit references make the sources more transparent, incrementing credibility perception. Quotes from outside experts further enhance the credibility of scientific articles by offering validation and expert perspectives [5, 337]. Analogously, the role of references and citations can be extrapolated to other domains and online content, such as journalism and newspapers, social media content and marketing [22, 337].

In Baier et al. [22], it is discussed how the explicitness of references and the depth of assurance provided by citations influence the perceived credibility of the information with experimental results showing that news articles with explicit citations are perceived as more credible by readers; whereas Hamborg et al. [109] highlights the importance of citations in presenting diverse perspectives. This is done leveraging the aggregating news from various sources and examining their references and the methodology involves analyzing citation patterns to identify biases and credibility in news articles. References, citations and quotes also can be used in network analysis to detect biases [5]. Patricia Aires et al.

[219] used references and citations in link analysis and bias detection, while quotes are discussed in regards to the effectiveness for bias detection.

The datasets used in these studies are obtained from multiple domains, including journalism, social media, and scientific content. In Baier et al. [22], the dataset comprises articles with varying levels of citation explicitness to experimentally assess reader perceptions of credibility. Hamborg et al. [109] aggregate news articles from diverse sources, examining their citation patterns to identify biases and provide multiple perspectives. Similarly, works [5, 219] derived the datasets from media outlets, focusing on link analysis and citation patterns to detect bias and assess credibility. These datasets reflect a broad range of content types, enhancing the generalizability of the findings across different informational contexts.

8.4 Clickbaits and title representativeness

The term *clickbait* refers to content designed to raise curiosity and attract users to click on links, often by using sensationalized or misleading headlines. The goal is typically to increase web traffic and advertising revenue [338]. Clickbait headlines often blur the lines between fact and fiction, contributing to the spread of fake news online [283].

To detect clickbait, various NLP methods can be employed, focusing on analyzing lexical and semantic features of headlines. For instance, frequent use of sensationalist language, unresolved pronouns, and forward-referencing structures can indicate clickbait [338]. Traditional machine learning models, such as Support Vector Machines (SVM) and Naïve Bayes classifiers, are often used to identify these features by assigning probabilities to words and phrases, which are then used to classify headlines as clickbait or non-clickbait [61]. In this direction, Zhang and Clough [338] focused on Chinese social media and found significant regional variations in clickbait prevalence. Their studies highlight the effectiveness of SVM and Naïve Bayes classifiers, achieving an F1-measure of 0.834. Combining these approaches with other models like Long Short-Term Memory (LSTM) networks can further enhance classification accuracy, with the SVM model achieving 98.53% accuracy [322].

Title representativeness, which measures how accurately a title reflects the content of the article, is another crucial aspect. Spezzano et al. [283] examined the credibility of news when provided with different combinations of title, image, and source bias. Their study found that combining these elements yielded the best accuracy for automated detectors 0.83, underscoring the importance of integrating multiple meta-data elements for improved accuracy in detecting misleading content.

The datasets used in these studies vary in their scope and origin. For example, one of them focused on Chinese social media platforms like WeChat, where they analyzed regional clickbait patterns using a dataset of social media posts [338]; whereas the study by Winarto et al. [322] leveraged YouTube titles in their study, creating a large, labeled dataset of video headlines to train their machine learning models. Additionally, Spezzano et al. [283] collected news articles with varying degrees of title representativeness, image content, and source bias, analyzing how these elements impacted credibility perceptions. Together, these datasets provide a diverse foundation for training and evaluating models aimed at detecting clickbait across different platforms and content types.

8.5 Originality and content reuse

Originality and content reuse have a direct impact on the credibility of information. When the content is original, it reflects the author's unique insights, analysis, or research, which strengthens its authenticity and trustworthiness. However, when the content is reused, whether through plagiarism, replication without attribution, or even subtle forms of paraphrasing, it can undermine the perceived credibility of the information. Plagiarism is defined as taking intellectual property from another and passing it off as one's own without citation [41].

In the context of Natural Language Processing, plagiarism detection has become more difficult due to the amount of texts available both in traditional print publications and now online. The challenge lies in detecting not only direct copying but also other forms of plagiarism, such as paraphrasing, translations or using hired writers to produce content. The exploration of originality and credibility signals in text reuse involves a variety of approaches combining NLP techniques with advanced machine learning methods. One prominent strategy is leveraging sentence segmentation, tokenization, and syntactic parsing to dissect the structure of text [64]. The proposed framework integrates these NLP techniques with trigram similarity measures and dependency relations matching, showing a detection accuracy of 70.53% for plagiarized short excerpts of text. This proves how traditional linguistic tools can be enhanced by statistical models to detect subtle patterns in content reuse. Similarly, trigram similarity alongside language model probability and longest common subsequence methods to deal with plagiarism detection can be adopted [215]. Use of a Naive Bayes classifier resulted in the same accuracy rate of 70.53%, underscoring the importance of pairing conventional text analysis techniques with machine learning algorithms to capture nuanced instances of content duplication.

Other works advanced further by incorporating deeper layers of analysis through methods like Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) [86]. The proposed solution employs lexical, semantic, and syntactic analysis to detect similarities, achieving an accuracy of 89%. By focusing on semantic relationships between words and phrases, it makes it particularly effective at discovering more sophisticated forms of content reuse.

Beyond general plagiarism detection, more specialized approaches concentrate themselves on verifying the authenticity of authorship. Enriquez et al. [87] developed an Authorship Verification model aimed at identifying hired plagiarism, where someone other than the credited author produces the work. By analyzing stylistic features, such as writing patterns and comparing them with the claimed author's profile, their approach achieved an accuracy of 85%. This sheds light on the possibilities of stylistic analysis as a signal of originality, distinguishing between genuine and outsourced authorship.

The datasets used in these studies on plagiarism detection can be found from a variety of sources and text types, reflecting the different challenges posed by plagiarism. For example, Bin-Habtoor and Zaher [41] focus on general plagiarism detection systems, analyzing publicly available datasets from academic publications. Chong et al. [64] utilize a dataset of academic articles to explore plagiarism through techniques like trigram similarity and dependency relations, applying them to short passages to measure detection accuracy. Enriquez et al. [87] use a dataset of student writing to train their Authorship Verification model, particularly aimed at detecting plagiarism in commissioned work. Elngar et al. [86] apply their algorithm to a dataset comprising academic and non-academic texts, with an emphasis on lexical, semantic, and syntactic analysis. Lastly, Pal et al. [215] use a dataset of online content, including articles and essays, to experiment with automatic plagiarism detection using advanced NLP techniques, combining feature extraction with machine learning models. These datasets represent a range of content, allowing for the robust evaluation of plagiarism detection methods across different textual formats.

8.6 Offensive language

The use of *offensive language* (and its related phenomena, such as hate speech, abusive or toxic language; see, e.g., [218] for definitions), particularly in online discourse, often correlates with the low credibility of the content, containing false information, prejudices, biases, and what in general is considered as toxic language. For example, Botella-Gil et al. [44] show that there is a direct association between the credibility of media sources and the presence of hate in online comments, while Bourgeade et al. [46] show how fake stories are used to spread hate against immigrants by analyzing racial hoaxes on Twitter. Also some of the most widely used datasets for check-worthy claim detection already include information about the presence of hateful [108] or harmful [205] content, showing the importance of adopting a multi-faceted view on the problem of misleading information. Indeed, the use of abusive or toxic language, has also been included in the W3C list of credibility signals [204].

Offensive language detection has been extensively investigated within the NLP community for at least a decade, starting from English [310] and then comprising more and more languages (for example the Hate Speech Datasets repository[27] that lists more than one hundred datasets in 25 languages). The creation of resources, which led in turn to the development of several approaches to offensive language detection, has been fostered by the shared tasks on hate or offensive language detection organised throughout the years at SemEval and other evaluation campaigns [43, 92, 222, 320, 334, 335]. Among the different approaches implemented for offensive language detection, transformer-based architectures based on RoBERTa [178] and its multilingual variants have proven to be very effective for the task when fine-tuned on offensive data, with F1 scores > 0.90 [335]. Nevertheless, the problem of online content moderation is far from being solved given the amount of toxic content still circulating on social media, which suggests that research on the topic still needs to address several understudied aspects of the phenomenon. Current works have identified some research directions worth studying, such as the problem of human label variation when creating datasets for offensive language detection [164], the role of annotators' biases [257], the presence of spurious correlations in existing datasets [242, 319] and the lack of robustness when classifying data from different online platforms and domains [256]. Concerning generative LLMs, recent experiments showed that classifying offensive language using zero- and few-shot learning LLMs yields considerably lower results in comparison to smaller models fine-tuned on the entire training set [82]. Generative LLMs have been alternatively used to create synthetic datasets for hate speech detection, with the goal to address problems such as data decay and privacy concerns related to social media [54]. However, the fact that major generative LLMs have been adjusted through so-called *alignment* to avoid generating hateful content may represent an obstacle for future research in this direction.

8.7 Non-NLP research on credibility signals

While the goal of this survey is to give an overview of the credibility assessment and credibility signals from the perspective of Natural Language Processing, additional credibility signals that can be extracted from ancillary elements are not of lesser relevance. These non-NLP credibility signals can range from the multimedia attached or linked to the textual information [109, 115, 333], to other cues based on different metrics, such as the impact of an author [47, 134, 340], or even the acceptance of the engaged public [157, 175, 342]. For example, it can be expected that a fake video would unlikely be a part of credible content; likewise, a biased author with an untrustworthy historic record could lead to the same outcome, the same as the refutation of an accepted authority in the field.

Several taxonomies can be used to categorize non-NLP credibility signals, for example, a distinction can be made on the signals coming from different content such as complementary videos or images, or external ones, like author or source information [5, 219, 330]. A general classification of non-NLP credibility signals is summarized in Table 10 with corresponding signals' examples and references.

As presented, a variety of factors could be leveraged in order to perform a credibility assessment, and the potential usage of them is a decision to make depending on the efficiency and the availability of the signals in the experiment. It

[27] https://hatespeechdata.com/

Table 10. Classification of Non-NLP signals

Category	Examples of non-NLP credibility signals	References
Image-based Signals	Image quality (resolution), Relevance of images to the rest of the content, Source attribution, Image authenticity (fake or generated), Presence of alternative text metadata	[213, 283, 337]
Video-based Signals	Video quality (high definition, video production), Relevance of video to the rest of the content, Source attribution, Video authenticity, Presence of transcriptions and subtitles	[213, 337]
Audio-based Signals	Audio quality, Relevance of audio to the rest of the content, Source attribution, Audio authenticity, Presence of transcriptions	[67]
Interactive and Embedded Content Signals	Quality and usability of interactive elements (interactive frameworks, maps, utility elements or forms), Trustworthiness of embedded elements (e.g. YouTube, newspapers, software providers)	[66]
Source Signals	Domain authority (reputation of the website, domain authority scores), Publisher reputation (historical credibility of publisher/organization), Website technology and design (quality, usability, loading speed, SEO implementation, consistency in design, branding/marketing), Traffic statistics (website analytics)	[283, 330]
Ownership and Transparency	Author information (availability, detail depth, verifiability of author), Editorial policies (regulations and moderation/revision processes of medium), Funding and sponsorship disclosure	[213]
External References	External references or reviews from renowned sites, External fact-checking from trustworthy sites	[134, 157, 179, 330]
User Interaction	Comments and discussion or readers' profiles, User reviews and engagement, Social media activity (e.g. virality, influencers)	[139, 157, 175, 179, 281, 333, 340]
Security of Website	HTTPS protocol, Legal policies and service terms, Compliance with standards (GDPR or local regulations), Cybersecurity practices	[213]
Content Presentation	Consistency in design and structure, Visual and audio aesthetics	[283]

is noticeable that the credibility signals are from a large variety of sources and formats, and some of them have been presenting good research results recently [67, 246, 330]. Additionally, it is worthy of noticing that this pattern is also shown through the extensive variety of credibility signals present in the text, where a large variety of credibility signals and much active presence in the research field can also be found, due to the big potential of the NLP techniques in this regard.

Lastly, a mention of the possibility of applying both textual and non-textual signals combined in the process of credibility assessment should be given, as multiple studies support these types of methodologies with good experimental results [47, 134, 329]. For example, one common approach to tackle low credible content detection is the usage of fake news propagation graphs [66, 281, 333], where the social activity of the sharing and propagation of a piece of news in a social network is the subject of study. There exists another aspect of interest in these studies where there is also a common practice of combining both the user and community engagement together with the actual content (including using NLP techniques to process the text) to detect the trustworthiness of the piece of news in particular [175, 340, 342].

The practice of applying both NLP and non-NLP based procedures, jointly with the notable variety of credibility signals [139, 213, 248] shows the interest and the open possibilities to aid in the field of exploiting these findings. The usage of one or the other, or both, will be dependent on the specific case of usage.

9 CHALLENGES AND OPEN PROBLEMS

Following the analysed research works as well as our own experience in the area of credibility assessment and credibility signals, we follow up with an overarching and orthogonal (to credibility assessment and categories of credibility signals described so far) discussion of challenges and open problems.

9.1 Adoption and potential of generative LLMs

Generative Large Language Models (LLMs) have demonstrated substantial improvements in complex tasks that require reasoning abilities [239]. Brown et al. [48] showed that pretrained LLMs are capable of few-shot learning, meaning they can learn to perform new tasks with only a few training examples. Similarly, Petroni et al. [229] highlighted the strong ability of LLMs to recall relational knowledge acquired during pretraining to perform various tasks without further annotated labels or human supervision (i.e., zero-shot learning). Additionally, several key advancements, such as Retrieval-Augmented Generation (RAG) [165], Reinforcement Learning with Human Feedback (RLHF) [212], and robust prompting techniques [239, 312], have further enhanced the capabilities of LLMs. In this context, recent generative LLMs operate as dialogue systems, where the model is prompted by the user with instructions to perform specific tasks.

Such models offer several opportunities to address challenges related to the automatic detection of credibility signals. One of the key advantages of using prompting with LLMs is the flexibility in adapting a single foundational model to handle multiple subtasks associated with credibility assessment. With a carefully designed framework, LLMs can be guided to focus on different aspects of content analysis, such as detecting persuasion techniques, evaluating the veracity of claims, identifying potential bias, and recognizing patterns of misinformation. This flexibility reduces the need to develop and fine-tune separate models for each task, allowing practitioners to use the same model across various credibility-related tasks. In fact, a promising research direction is to explore multi-task learning, as in verifying if the capacity of performing certain credibility-assessment tasks can aid in other related ones (e.g., persuasion and bias).

Moreover, the capability of learning with zero/few examples is particularly valuable for tasks where domain-specific data is scarce or constantly changing, as is the case with credibility assessment. An enormous amount of human effort is required to curate high-quality annotated datasets for the different subtasks involved in assessing credibility. Specially since tasks such as labeling misinformation or identifying biased content often demand the expertise of domain specialists such as fact-checkers, journalists and social scientists. Adding to this challenge, credibility indicators can be highly context-dependent, varying across cultural and temporal dimensions. In this context, LLMs offer a more scalable approach by drawing on vast amounts of unsupervised pretraining data, and by adjusting to specific end-tasks through careful prompting strategies, which require far less human effort than manual data labeling. As an example, in Leite et al. [163], a generative LLM was employed to predict 19 different credibility signals present in textual content without using any training data (i.e., in a zero-shot setting).

Finally, the generative capabilities of large language models can be leveraged to produce more interpretable[28] outputs, which is crucial for subject-matter experts that may leverage the model's predictions. Instead of providing only binary or scalar outputs (e.g., true/false, misinformation/non-misinformation) as in usual classification tasks, generative LLMs can produce detailed explanations or summaries that can highlight the reasoning behind their predictions. This transparency allows human experts to critically assess the model's outputs, cross-check them with external information, and ensure that (i) incorrect predictions (in this context, known as *model hallucinations* [140]) are properly mitigated, and (ii) any credibility assessment aligns with the context of the content being investigated. This property of interpretable outputs can significantly increase trust in model-driven decisions and reduce the likelihood of over-reliance on machine predictions, ensuring that humans remain in the loop for final judgments.

We acknowledge that LLM adoption (despite providing a lot of potential) is also accompanied with several challenges. Fine-tuning as well as deployment of LLMs require a considerably higher computing power which directly translates to higher costs. Moreover, learning techniques commonly used in limited labelled data scenarios (prompting, in-context

[28] Here, the concept of interpretability differs from explainability, which is often used in the field of machine learning to refer to specific methods to analyse how intermediate states of the model lead to certain outcomes [50].

learning, fine-tuning, meta-learning, or few-shot learning) are known to be sensitive to various randomness factors [226], what can result in undesired performance instability. Beside randomness factors, also systematic choices, such as the format of the prompt or how many (in-context) samples are used, have significant effect on the overall performance and the stability of these approaches [225, 302]. Nevertheless, ongoing research in the area of LLMs is already providing suitable solutions to such problems, such as Parameter-efficient Fine Tuning (PEFT) techniques [327], which have been demonstrated to perform well also on credibility signal detection tasks [243]; or instability mitigation techniques, such as ensembling, noise regularisation and model interpolation [223]. At the same time, we would like to stress that employing LLMs does not necessarily provide benefit in all cases. When a sufficient amount of labelled samples is available (100-1000 depending on the specific task and model), fine-tuned smaller language models can outperform larger general ones [224].

9.2 Dataset availability and multilinguality

Dataset availability heavily differs across various categories of credibility signals. Firstly, we observed a significant lack of large-enough datasets providing overall (expertly-determined) annotation of credibility that can be used for training and evaluating solutions on credibility assessment task (Section 4.1). Furthermore, datasets containing annotations for overall credibility and (at least some) credibility signals at the same time are even more scarce. As a result, many researchers opted to use (more available) fake news datasets as a replacement. At this place, we would like to highlight again that fake news annotations cannot reliably replace credibility annotations (non-credible content is not necessarily only false content, since credibility is a broader concept going beyond content truthfulness).

On the other hand, the situation with dataset availability for individual categories of credibility signals is much better. This is especially thanks to data challenges (particularly SemEval tasks and CLEF CheckThat! Labs) in which either new datasets were introduced or existing datasets were extended (with more data, new languages, or additional types of annotations). Unfortunately, in some cases (e.g., [232]), the datasets from data challenges are not published completely and a hidden/testing set is not shared with researchers not even when the competition is over (and thus only the training/validation sets remain available for further research).

Especially for international news or the investigation of global claims, journalists and fact-checkers need to verify the credibility of information by cross-checking sources in different languages. Furthermore, emerging disinformation in one country could be spread to other countries, especially when its topic is global (e.g., pandemic, wars, international relations). Therefore, it is important to implement credibility analysis tools that support multiple languages. However, due to the scarcity of multilingual datasets, trained models can exhibit biases towards some languages and cultures and hence can underperform on content in low-resourced languages.

To overcome this issue, global collaborations could be initiated for creating multilingual resources. In this direction, we can already observe a positive trend within data challenges. Many of them introduced multilingual datasets, commonly considering some languages as surprise ones (i.e., languages that are present in a test set, but missing in a train/validation set). Such approach motivates participants to develop multilingual solutions that are capable to make a prediction in a zero-shot setting (considering a language a predicted content is written in).

Besides datasets availability and multilinguality, a quality of annotations remain another challenge. Annotation of overall credibility as well as individual credibility signals is many times highly subjective (as also shown in [40]), especially in cases such as persuasion techniques where presence of signal and borders between various signals can be blurred. The situation is getting even more challenging in multilingual settings, where typically native speakers are needed to annotate data. Firstly, acquiring human experts fluent in several languages is challenging itself. Secondly, organizing and consolidating annotation process (including post-annotation verification) is a complex task. Considering also our own experience working with such datasets, the provided labels cannot be easily verified and many times we identified (inevitable) incorrect labels.

Last but not least, credibility assessment and automatic detection of credibility signals naturally happen in very dynamically evolving (online) environment. New topics and global events constantly emerge, causing significant data and concept drifts. In some cases, such drifts can cause that the existing datasets may become obsolete and non-representative. Secondly, the list of credibility signals evolve itself. We can take a machine-generated text as an illustrative example. This kind of signal become highly relevant only recently with generative LLMs becoming easily available for a large end user base (and unfortunately also bad actors). Such new/redefined signals thus naturally result into a demand for new datasets. Finally, dynamics of this area also lies in its adversarial character. Bad actors (e.g., the ones who are spreading propaganda or disinformation) will always try to get their content undetected as low-credible one by employing various obfuscation techniques. This must be taken into consideration when introducing new datasets. By continuing with an illustrative example of machine-generated text, there are already datasets (e.g., [188]) providing besides machine-generated texts also their alternative versions after applying several authorship obfuscation techniques, what allows to train and evaluate more robust detection models.

9.3 Ethical and legal issues

Credibility assessment of online (primarily social media) content is from its nature an area that must address several ethical and legal issues. Firstly, similarly to other related research areas (e.g., on fake news detection), it may be potentially misused by bad actors in order to create a content that appears to be more credible. Also by following open-research spirit, publishing credibility assessment systems can theoretically result into misusing such system in the adversarial manner to tune disinformation/propaganda generation systems and allowing them to stay undetected. This kind of potential threat is, however, an analogical issue to the security domain and the principle of security by obscurity. Nothing prevents bad actors to develop their own credibility assessments systems and apply then in adversarial training scenario. Moreover, positive outcomes of credibility assessment research are more tangible, with many practical tools (as also showed across this survey) already put into the hands of media professionals or general public.

Besides potential misuse, additional ethical considerations must be addressed thoroughly during the research activities. First of all, training various classification/detection systems is inherently a subject of potential biases. Such biases can come directly from the datasets used for the training purposes (in terms of data selection, data pre-/post-processing, or data annotation itself) or can be introduced during training the models, especially when fine-tuning pre-trained LLMs that have incorporated biases by themselves (including biases between high- and low-resource languages). Secondly, the authors should always clearly formulate intended use and failure cases of the trained models/deployed tools. In this way, we may prevent media professionals to over-rely on the predicted (potentially incorrect) values. Fortunately, the two-step approach to credibility assessment (i.e., detect more granular credibility signals and then aggregate them) makes the whole process more transparent. Last but not least, explainability and interpretability of the models' predictions plays a crucial role in this area, since credibility assessment must be credible itself, otherwise it would not provide expected level of trust to its end users.

Besides ethical issues, the research in this area must address also several legal issues, many of them are shared with other research works on social media. Firstly, as we also witness recently, social media platforms can change their data processing policies and restrict access to their data, which may delay progress in research and development of credibility tools. Additionally, social media data must be anonymized to protect user privacy before being used as training data or for the model inference. Media organizations also impose limitations due to copyright laws, with some not permitting their content to be used in AI tool development. LLMs, especially closed LLMs such as ChatGPT and GPT-4, lack transparency regarding pre-training data and LLMs can memorize content in their pretraining data [148, 202]. Therefore, anonymization and removal of copyrighted content is crucial to credibility tools, even when they serve as foundational or backbone models. Paradigms such as unlearning [59] or LLM editing could be potential research directions to tackle these issues.

9.4 Practical deployment

Credibility assessment is naturally very challenging when it comes into the practical deployment. Also by proceeding from our own experience, during a deployment in real-world settings, the performance of trained models many times does not achieve the sufficient thresholds. The root of performance decrease between the offline experiments (on the testing data) vs. in the real-world world scenarios can be traced into the out-of-distribution nature of data (from any of relevant aspects, such as topic, format, language, etc.). This imposes a need for more robust solutions, which are, however, difficult to achieve considering already-discussed limited dataset availability and dynamics of online environments.

Furthermore, there is commonly a lack of mutual understanding. Technical practitioners and media professionals often use different terminology for technical concepts and tasks. This can pose challenges on interpretation of AI-predicted outputs, understanding limitations of AI models or interpretation of task results. As a similar challenge, the prior datasets and benchmarks were mostly created by computational linguists or computer scientists, while media professionals were not actively involved in labeling. This may easily result into different perspectives what is credible or not; or where some credibility signals should be detected as present and where not. Additionally, as seen from the survey results, some tasks are already subjective such as bias detection and check-worthy detection tasks. These issues can result in model outputs that can fail to align with the perspectives of media professionals, or only succeed in identifying cherry-picked (unambiguous) examples.

Although it is desirable (from the research perspective) to use open-source models for transparency and reproducibility, hosting of several models can be costly for media organizations or academic institutions. Also researchers, when deploying their models, must consider a dilemma regarding the size of the models. Larger language models can typically achieve a higher classification accuracy. At the same time, they would need to be deployed at costly GPU-machines (in order to make inference feasible and fast enough). As a result, a compromise between computation costs and acceptable performance drop must be found. While methods on distilling large models to smaller version can become helpful in some cases, the performance of distilled models can again be worse in comparison to their original versions.

Last but not least, end users (especially general public), even together with their peers, lack the capability to determine the correctness of AI-based credibility indicators [181]. This implies serious risks of people being misled by AI-based credibility indicators that are wrong, and consequently believing in or even spreading misinformation. For this reason, already mentioned explainability/interpretability as well as clear communication of expected accuracy of the AI-based systems to end users represent a critical aspect of the whole solution.

9.5 Multimodal approaches

When tackling credibility assessment, the usage of multimodal approaches grows accordingly to the adjacent increasing complexity of the platforms and the evolution of the communication methods. While text is still one of the most common methods to spread information, it can also be combined with other media types, such as images, videos and audios. Multimodal techniques integrate various forms of data, offering a richer understanding of the content and its credibility signals. The main potential of these methods lies in their ability to overcome limitations inherent in text-only analysis. While Natural Language Processing excels at detecting linguistic patterns that may signal misinformation (e.g., sentiment analysis, linguistic complexity, semantic coherence), it can overlook cues that are often found in non-textual elements. Additionally, as previously commented in Section 8.7, there are promising results of non-NLP methodologies to extract information and perform analysis in the credibility context. Thus, having proven the potential of NLP, non-NLP techniques can be combined with it at the same time.

In this regard, the more the interconnectivity of the media formats and the evolution of online sources (such as newspapers, social networks, blogs and other types of communication frameworks), the more interest in using multimodal approaches exists. Non-credible content often spans beyond text, employing visual, auditive, and interactive elements to manipulate perceptions. Studies have shown that combining textual analysis with multimedia analysis leads to significant improvements in detecting and understanding misinformation [12, 65, 113].

Multimodal systems rely on complementary features extracted from text alongside non-NLP methods discussed earlier. For example, low credibility content such a fake news article can back-up its misleading text with an altered image or video to enhance its perceived credibility. In such cases, the integration of visual and textual analysis is critical when flagging the content as suspicious.

Multimodal approaches involve processing multiple data streams simultaneously, combining traditional NLP methods with analysis of visual, auditive, and interactive content to form a wider comprehension of credibility. This integration is typically achieved through ensemble models or architectures that unify the outputs of various sub-models focused on different media types. A common method involves using transformer-based models such as BERT or GPT for text analysis alongside convolutional neural networks (CNNs) for image or video analysis. These models work in tandem, the text models identifying linguistic signals of credibility while the CNNs assess visual features like image authenticity, quality, and relevance. Additional layers of analysis might focus on audio quality, transcription accuracy, or even the relationship between the text and accompanying media (e.g., whether an image truly matches the content described in an article). For example, a multimodal model might analyze a tweet's text for sentiment and coherence, while also assessing the attached image's metadata and authenticity. If the image shows signs of manipulation or does not align with the textual claim, the credibility for that piece of content would be lowered.

Even in multimodal systems, textual analysis remains a central component. Text provides rich semantic information, often containing the explicit claims or arguments being made, which serve as the basis for evaluating credibility. Without text, it would be difficult to discern the specific intent behind multimedia content and a more specific tool would be needed to perform the evaluation under certain conditions. Thus, NLP techniques like semantic parsing, sentiment analysis, and entity recognition form the ground of these multimodal approaches, supplying the primary framework for understanding the broader context in which non-textual signals are embedded.

Studies in multimodal credibility assessments consistently show that while non-textual cues (e.g., image authenticity, video manipulation detection) significantly enhance the detection process, they are most effective when paired with deep text analysis. For instance, recent work on fake news propagation graphs combines user engagement patterns with textual signals and visual content to identify disinformation [240]. Another promising approach is the integration of social media metadata (such as virality metrics and user behavior) alongside textual content to detect the spread of fake news [283, 333]. The application of multimodal techniques in credibility assessments is expanding rapidly, particularly in social media, news aggregation platforms, and user-generated content sites like YouTube, Twitter, and Reddit.

In future developments, there is potential to enhance multimodal models by integrating additional modalities, such as biometric signals (e.g. eye-tracking data), interactive content quality, or real-time user engagement metrics. Research on multimodal disinformation detection is ongoing, with several promising methods emerging, such as ensemble techniques that combine the strengths of multiple machine learning models across different modalities [9, 115].

10 CONCLUSIONS

With the rapid development of generative AI, the threat of misusing LLMs to hybrid operations is one of the commonly cited risks of their future development [49, 98]. The recent research demonstrated especially a high potential of LLMs to generate disinformation content in various global or local narratives [303]. In this context, a large-scale ability to automatically determine the credibility of online social media content becomes even more crucial as it was in the pre-LLM era.

Potential negative impact on society and democratic values was recognized by the research community and previously led to emergence of many fake news detection approaches. At this place, we would like to emphasize that credibility assessment, while being a sibling task to fake-news detection, provides several considerable advantages, such as better and deeper explainability or wider opportunities for application in research as well as in practice.

In addition, being able to reliably automatically recognize credible content, is also a way how to tackle with extensive spreading and consumption of false information. Detected high-credible content can be promoted by recommender systems or search engines, or highlighted in a user interface of social media platforms, while low-credible content can be accompanied with a low-credibility warning. With a breakdown of predicted credibility score to individual credibility indicators, end users (either media professionals or a general public) are able to explore the evidence leading to the credibility assessment or manually evaluate individual detected credibility signals and make a final assessment by themselves. Such level of explainability and pro-active human involvement in the decision process is vital and unfortunately lacks in many fake news detection approaches that commonly result into a single predicted value (commonly a binary one) with challenging or even impossible explanation caused by a black-box nature of the employed techniques (e.g., deep learning approaches).

Despite these considerable advantages of credibility assessment, the current state-of-the-art research works suffer from multiple drawbacks. Most crucially, credibility assessment field can be characterized as highly fragmented. On one side, there are credibility assessment approaches that automatically detect credibility signals and aggregate them to make a final prediction about the content credibility. The utilized credibility signals are, however, mostly shallow linguistic ones (such as a number of hashtags), their automatic detection relies on simple methods (like rule- or heuristic-based techniques), and also prediction utilizes mostly basic weighting schemata. There are only very few approaches that are in line with the current state of the art (deep learning, LLMs), such as [163]; or using the previous research results to detect more complex credibility signals, such as [240].

On the other side, there are automatic approaches detecting various categories of credibility signals, like factuality, biases, persuasion techniques, or previously fact-checked claims. The prevalence of state-of-the-art techniques (including LLMs and various fine-tuning approaches, including PEFTs) is much higher in these works. However, such approaches remain isolated from credibility assessment, many times even not mentioning that their prediction can be considered as one of more advanced and more reliable credibility signals.

To contribute in closing this undesired gap between research works, lack of interconnection of research results, as well as hindering application of the outcomes in practice, we conducted this systematic survey on automatic credibility assessment and detection of credibility signals from the NLP perspective. By collecting and describing 175 research papers, we not only systematically summarized the current state of the research in (currently fragmented) research areas, but also identified challenges and potential for future research. Our thorough analyses and discussions aim to point to an interesting avenues for future research – out of them, we would like to specifically highlight the adoption of latest highly-multilingual and multimodal LLMs. Their capabilities, currently remaining largely undiscovered, represent a way how to make a significant step forward in the area of credibility assessment. Moreover, they provide an unprecedented chance for deployment of the research results and providing novel credibility assessment tools to media professionals as well as general public. Such tools can serve both purposes – to detect false non-credible content as well as identify credible one that is worth further reading and sharing in social media environment.

ACKNOWLEDGMENTS

This work was partially supported by the European Union under the Horizon Europe projects: *vera.ai* (GA No. 101070093), *AI-CODE* (GA No. 101135437); *AI4TRUST* (GA No. 101070190); by the UK's innovation agency (Innovate UK) grant 10039055; and by EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V03-00020.

REFERENCES

[1] Athira A.b., S. D. Madhu Kumar, and Anu Mary Chacko. 2023. A Systematic Survey on Explainable AI Applied to Fake News Detection. *Engineering Applications of Artificial Intelligence* 122 (June 2023), 106087. https://doi.org/10.1016/j.engappai.2023.106087

[2] Reem Abdel-Salam. 2023. rematchka at ArAIEval Shared Task: Prefix-Tuning & Prompt-tuning for Improved Detection of Propaganda and Disinformation in Arabic Social Media Content. In *Proceedings of ArabicNLP 2023*, Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham (Eds.). Association for Computational Linguistics, Singapore (Hybrid), 536–542. https://doi.org/10.18653/v1/2023.arabicnlp-1.52

[3] Dia Abujaber, Ahmed Qarqaz, and Malak A. Abdullah. 2021. LeCun at SemEval-2021 Task 6: Detecting Persuasion Techniques in Text Using Ensembled Pretrained Transformers and Data Augmentation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 1068–1074. https://doi.org/10.18653/v1/2021.semeval-1.148

[4] Swati Aggarwal, Tushar Sinha, Yash Kukreti, and Siddarth Shikhar. 2020. Media bias detection and bias short term impact assessment. *Array* 6 (2020), 100025. https://doi.org/10.1016/j.array.2020.100025

[5] Victoria Aires, Altigran da Silva, Fabiola Nakamura, and Eduardo Nakamura. 2020. An Evaluation of Structural Characteristics of Networks to Identify Media Bias in News Portals. In *Proceedings of the Brazilian Symposium on Multimedia and the Web* (São Luís, Brazil) *(WebMedia '20)*. Association for Computing Machinery, New York, NY, USA, 225–232. https://doi.org/10.1145/3428658.3431077

[6] Victoria Patricia Aires, Juliana Freire, Fabiola G Nakamura, Altigran Soares da Silva, and Eduardo F Nakamura. 2020. An information theory approach to detect media bias in news websites.

[7] Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal Automated Fact-Checking: A Survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5430–5448. https://doi.org/10.18653/v1/2023.findings-emnlp.361

[8] Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem (Eds.). Association for Computational Linguistics, Online, 121–131. https://doi.org/10.18653/v1/2021.woah-1.13

[9] Mohammed Al-alshaqi, Danda B. Rawat, and Chunmei Liu. 2024. Ensemble Techniques for Robust Fake News Detection: Integrating Transformers, Natural Language Processing, and Machine Learning. *Sensors* 24, 18 (2024). https://doi.org/10.3390/s24186062

[10] Rasha M. BinSultan Al-Eidan, Hend S. Al-Khalifa, and AbdulMalik S. Al-Salman. 2009. Towards the measurement of Arabic Weblogs credibility automatically. In *Proceedings of the 11th International Conference on Information Integration and Web-Based Applications & Services* (Kuala Lumpur, Malaysia) *(iiWAS '09)*. Association for Computing Machinery, New York, NY, USA, 618–622. https://doi.org/10.1145/1806338.1806455

[11] Firoj Alam, Alberto Barrón-Cedeño, Gullal S. Cheema, Gautam Kishore Shahi, Sherzod Hakimov, Maram Hasanain, Chengkai Li, Rubén Míguez, Hamdy Mubarak, Wajdi Zaghouani, and Preslav Nakov. 2023. Overview of the CLEF-2023 CheckThat! Lab Task 1 on Check-Worthiness in Multimodal and Multigenre Content. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023 (CEUR Workshop Proceedings, Vol. 3497)*, Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos (Eds.). CEUR-WS.org, 219–235. https://ceur-ws.org/Vol-3497/paper-019.pdf

[12] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A Survey on Multimodal Disinformation Detection. (2022). arXiv:2103.12541 [cs.MM] https://arxiv.org/abs/2103.12541

[13] Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 Shared Task on Propaganda Detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, and Wajdi Zaghouani (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 108–118. https://doi.org/10.18653/v1/2022.wanlp-1.11

[14] Bodor Moheel Almotairy, Manal Abdullah, and Dimah Hussein Alahmadi. 2024. Dataset for detecting and characterizing Arab computation propaganda on X. *Data in Brief* 53 (2024), 110089. https://doi.org/10.1016/j.dib.2024.110089

[15] Majed Alrubaian, Muhammad Al-Qurishi, Atif Alamri, Mabrook Al-Rakhami, Mohammad Mehedi Hassan, and Giancarlo Fortino. 2019. Credibility in Online Social Networks: A Survey. *IEEE Access* 7 (2019), 2828–2855. https://doi.org/10.1109/ACCESS.2018.2886314

[16] Wazib Ansar and Saptarsi Goswami. 2021. Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights* 1, 2 (2021), 100052. https://doi.org/10.1016/j.jjimei.2021.100052

[17] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A Benchmark Dataset of Check-Worthy Factual Claims. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles (Eds.). AAAI Press, 821–829. https://ojs.aaai.org/index.php/ICWSM/article/view/7346

[18] Malik Atalla, Christian Scheel, Ernesto William De Luca, and Sahin Albayrak. 2011. Investigating the applicability of current machine-learning based subjectivity detection algorithms on german texts. In *Proceedings of Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*. 17–24.

[19] Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 1: Check-Worthiness. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018 (CEUR Workshop Proceedings, Vol. 2125)*, Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-2125/invited_paper_13.pdf

[20] Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 1: Check-Worthiness. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019 (CEUR Workshop Proceedings, Vol. 2380)*, Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-2380/paper_269.pdf

[21] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4685–4697. https://doi.org/10.18653/v1/D19-1475

[22] Christian Baier, Max Göttsche, Axel Hellmann, et al. 2022. Too Good To Be True: Influencing Credibility Perceptions with Signaling Reference Explicitness and Assurance Depth. In *Journal of Business Ethics*, Vol. 178. 695–714. https://doi.org/10.1007/s10551-020-04719-7

[23] Vít Baisa, Ondřej Herman, and Ales Horak. 2019. Benchmark Dataset for Propaganda Detection in Czech Newspaper Texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Varna, Bulgaria, 77–83. https://doi.org/10.26615/978-954-452-056-4_010

[24] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4982–4991. https://doi.org/10.18653/v1/2020.emnlp-main.404

[25] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3528–3539. https://doi.org/10.18653/v1/D18-1389

[26] Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3364–3374. https://doi.org/10.18653/v1/2020.acl-main.308

[27] Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-Task Ordinal Regression for Jointly Predicting the Trustworthiness and the Leading Political Ideology of News Media. 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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 2109–2116. https://doi.org/10.18653/v1/N19-1216

[28] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual Subjectivity: Are More Languages Better?. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Chu-Ren Huang and Dan Jurafsky (Eds.). Coling 2010 Organizing Committee, Beijing, China, 28–36. https://aclanthology.org/C10-1004

[29] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. arXiv:2310.05130 [cs.CL] https://arxiv.org/abs/2310.05130

[30] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Coling 2010: Posters*. 36–44.

[31] Andrew W. Barrett and Lowell W. Barrington. 2005. Bias in Newspaper Photograph Selection. *Political Research Quarterly (formerly WPQ)* 58 (2005), 609 – 618. https://api.semanticscholar.org/CorpusID:154387178

[32] Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, Gullal S. Cheema, Dilshod Azizov, and Preslav Nakov. 2023. The CLEF-2023 CheckThat! Lab: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III* (Dublin, Ireland). Springer-Verlag, Berlin, Heidelberg, 506–517. https://doi.org/10.1007/978-3-031-28241-6_59

[33] Alberto Barrón-Cedeño, Firoj Alam, Tanmoy Chakraborty, Tamer Elsayed, Preslav Nakov, Piotr Przybyła, Julia Maria Struß, Fatima Haouari, Maram Hasanain, Federico Ruggeri, Xingyi Song, and Reem Suwaileh. 2024. The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 449–458. https://doi.org/10.1007/978-3-031-56069-9_62

[34] Alberto Barrón-Cedeño, Firoj Alam, Andrea Galassi, Giovanni Da San Martino, Preslav Nakov, Tamer Elsayed, Dilshod Azizov, Tommaso Caselli, Gullal S. Cheema, Fatima Haouari, Maram Hasanain, Mucahid Kutlu, Chengkai Li, Federico Ruggeri, Julia Maria Struß, and Wajdi Zaghouani. 2023. Overview of the CLEF–2023 CheckThat! Lab on Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and Their Source. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer Nature Switzerland, Cham, 251–275.

[35] Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12260)*, Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro (Eds.). Springer, 215–236. https://doi.org/10.1007/978-3-030-58219-7_17

[36] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150 [cs.CL] https://arxiv.org/abs/2004.05150

[37] Seth P Benson and Iain J Cruickshank. 2024. Developing a Natural Language Understanding Model to Characterize Cable News Bias. *IEEE Access* (2024).

[38] Bettina Berendt, Peter Burger, Rafael Hautekiet, Jan Jagers, Alexander Pleijter, and Peter Van Aelst. 2021. FactRank: Developing automated claim detection for Dutch-language fact-checkers. *Online Social Networks and Media* 22 (2021), 100113. https://doi.org/10.1016/j.osnem.2020.100113

[39] Varad Bhatnagar, Diptesh Kanojia, and Kameswari Chebrolu. 2022. Harnessing Abstractive Summarization for Fact-Checked Claim Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2934–2945. https://aclanthology.org/2022.coling-1.259

[40] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 93 (oct 2020), 26 pages. https://doi.org/10.1145/3415164

[41] A. Bin-Habtoor and Mahmoud Zaher. 2012. A Survey on Plagiarism Detection Systems. *International Journal of Computer Theory and Engineering* (01 2012), 185–188. https://doi.org/10.7763/IJCTE.2012.V4.447

[42] Prakhar Biyani, Sumit Bhatia, Cornelia Caragea, and Prasenjit Mitra. 2014. Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowledge-Based Systems* 69 (2014), 170–178. https://doi.org/10.1016/j.knosys.2014.04.048

[43] Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

[44] Beatriz Botella-Gil, Alba Bonet-Jover, Robiert Sepúlveda-Torres, Patricio Martínez-Barco, and Estela Saquete. 2024. Exploring the Relationship between News Reliability and Violent Comments in Digital Media. In *Proceedings of the 40th International Conference of the Spanish Society for Natural Language Processing*.

[45] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. BRENDA: Browser Extension for Fake News Detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2117–2120. https://doi.org/10.1145/3397271.3401396

[46] Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads.

In *Findings of the Association for Computational Linguistics: EACL 2023*. 674–684.

[47] Erica J. Briscoe, D. Scott Appling, Rudolph L. Mappus, and Heather Hayes. 2013. Determining credibility from social network structure. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Niagara, Ontario, Canada) *(ASONAM '13)*. Association for Computing Machinery, New York, NY, USA, 1418–1424. https://doi.org/10.1145/2492517.2492574

[48] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[49] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. *Truth, Lies, and Automation: How Language Models Could Change Disinformation.* Technical Report. Center for Security and Emerging Technology. https://doi.org/10.51593/2021CA003

[50] Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

[51] Jinjuan Cao, Zhong Qian, and Peifeng Li. 2022. End-to-End Event Factuality Identification with Cross-Lingual Information. In *2022 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892869

[52] Pengfei Cao, Yubo Chen, Yuqing Yang, Kang Liu, and Jun Zhao. 2021. Uncertain Local-to-Global Networks for Document-Level Event Factuality Identification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2636–2645. https://doi.org/10.18653/v1/2021.emnlp-main.207

[53] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 675–684. https://doi.org/10.1145/1963405.1963500

[54] Camilla Casula, Elisa Leonardelli, and Sara Tonelli. 2024. Don't Augment, Rewrite? Assessing Abusive Language Detection with Synthetic Data. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 11240–11247. https://doi.org/10.18653/v1/2024.findings-acl.669

[55] Recep Firat Cekinel, Çağrı Çöltekin, and Pinar Karagoz. 2024. Cross-Lingual Learning vs. Low-Resource Fine-Tuning: A Case Study with Fact-Checking in Turkish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 4127–4142. https://aclanthology.org/2024.lrec-main.368

[56] Yung-Sheng Chang, Yan Zhang, and Jacek Gwizdka. 2021. The effects of information source and eHealth literacy on consumer health information credibility evaluation behavior. *Computers in Human Behavior* 115 (2021), 106629. https://doi.org/10.1016/j.chb.2020.106629

[57] Iti Chaturvedi, Erik Cambria, Roy E. Welsch, and Francisco Herrera. 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion* 44 (2018), 65–77. https://doi.org/10.1016/j.inffus.2017.12.006

[58] Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-Claims: A Dataset for Multimodal Claim Detection in Social Media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 962–979. https://doi.org/10.18653/v1/2022.findings-naacl.72

[59] Jiaao Chen and Diyi Yang. 2023. Unlearn What You Want to Forget: Efficient Unlearning for LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12041–12052. https://doi.org/10.18653/v1/2023.emnlp-main.738

[60] Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, David Bamman, Dirk Hovy, David Jurgens, Brendan O'Connor, and Svitlana Volkova (Eds.). Association for Computational Linguistics, Online, 149–154. https://doi.org/10.18653/v1/2020.nlpcss-1.16

[61] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading Online Content: Recognizing Clickbait as "False News". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection* (Seattle, Washington, USA) *(WMDD '15)*. Association for Computing Machinery, New York, NY, USA, 15–19. https://doi.org/10.1145/2823465.2823467

[62] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at SemEval-2020 Task 11: It Takes Three to Tango: RoBERTa, CRF, and Transfer Learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, Barcelona (online), 1462–1468. https://doi.org/10.18653/v1/2020.semeval-1.191

[63] Wonchan Choi, Besiki Stvilia, and Hyun Seung Lee. 2023. Developing a platform-specific framework for web credibility assessment: A case of social Q&A sites. *Information Processing & Management* 60, 3 (2023), 103321. https://doi.org/10.1016/j.ipm.2023.103321

[64] Miranda Chong, Lucia Specia, and Ruslan Mitkov. 2010. Using natural language processing for automatic plagiarism detection. In *Proc. of 4th International Plagiarism Conference, Northrumbia University Newcastle-upon-Tyne, UK*.

[65] Monika Choudhary, Satyendra Singh Chouhan, and Santosh Singh Rathore. 2024. Beyond Text: Multimodal Credibility Assessment Approaches for Online User-Generated Content. *ACM Trans. Intell. Syst. Technol.* (June 2024). https://doi.org/10.1145/3673236 Just Accepted.

[66] Hui Na Chua, Qaisar Khan, Muhammed Basheer Jasser, and Richard T. K Wong. 2023. Problem Understanding of Fake News Detection from a Data Mining Perspective. In *2023 IEEE 13th International Conference on Control System, Computing and Engineering (ICCSCE)*. 297–302. https://doi.org/10.1109/ICCSCE58721.2023.10237152

[67] Amelia Ciobanu, Tudor Culda, Cristian Negrescu, and Dumitru Stanomir. 2014. Analysis of reverberation time blind estimation used in audio forensics. In *2014 11th International Symposium on Electronics and Telecommunications (ISETC)*. 1–4. https://doi.org/10.1109/ISETC.2014.7010810

[68] Keith Cortis and Brian Davis. 2022. Baseline English and Maltese-English Classification Models for Subjectivity Detection, Sentiment Analysis, Emotion Analysis, Sarcasm Detection, and Irony Detection. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*. 161–168.

[69] Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *IEEE Access* 11 (2023), 70977–71002. https://doi.org/10.1109/ACCESS.2023.3294090

[70] Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv:2006.00885 [cs.SI] https://arxiv.org/abs/2006.00885

[71] Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: A System for Explainable Fake News Detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 2961–2964. https://doi.org/10.1145/3357384.3357862

[72] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an architecture for development of robust HLT applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 168–175. https://doi.org/10.3115/1073083.1073112

[73] Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov (Eds.). Association for Computational Linguistics, Hong Kong, China, 162–170. https://doi.org/10.18653/v1/D19-5024

[74] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, Barcelona (online), 1377–1414. https://doi.org/10.18653/v1/2020.semeval-1.186

[75] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Yokohama, Yokohama, Japan) *(IJCAI'20)*. Article 672, 7 pages.

[76] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Prta: A System to Support the Analysis of Propaganda Techniques in the News. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Asli Celikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, Online, 287–293. https://doi.org/10.18653/v1/2020.acl-demos.32

[77] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5636–5646. https://doi.org/10.18653/v1/D19-1565

[78] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management* 60, 2 (2023), 103219. https://doi.org/10.1016/j.ipm.2022.103219

[79] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 69–76.

[80] Rudra Dhar and Dipankar Das. 2021. Leveraging Expectation Maximization for Identifying Claims in Low Resource Indian Languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, Sivaji Bandyopadhyay, Sobha Lalitha Devi, and Pushpak Bhattacharyya (Eds.). NLP Association of India (NLPAI), National Institute of Technology Silchar, Silchar, India, 307–312. https://aclanthology.org/2021.icon-main.37

[81] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 70–98. https://doi.org/10.18653/v1/2021.semeval-1.7

[82] Aleksandra Edwards and Jose Camacho-Collados. 2024. Language Models for Text Classification: Is In-Context Learning Enough?. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 10058–10072. https://aclanthology.org/2024.lrec-main.879

[83] Adam Ek and Mehdi Ghanimifard. 2019. Synthetic Propaganda Embeddings To Train A Linear Projection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov (Eds.). Association for Computational Linguistics, Hong Kong, China, 155–161. https://doi.org/10.18653/v1/D19-5023

[84] Rim El Ballouli, Wassim El-Hajj, Ahmad Ghandour, Shady Elbassuoni, Hazem Hajj, and Khaled Shaban. 2017. CAT: Credibility Analysis of Arabic Content on Twitter. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, Nizar Habash, Mona Diab, Kareem Darwish, Wassim El-Hajj, Hend Al-Khalifa, Houda Bouamor, Nadi Tomeh, Mahmoud El-Haj, and Wajdi Zaghouani (Eds.). Association for Computational Linguistics, Valencia, Spain, 62–71. https://doi.org/10.18653/v1/W17-1308

[85] Houda Elmimouni, Andrea Forte, and Jonathan Morgan. 2022. Why People Trust Wikipedia Articles: Credibility Assessment Strategies Used by Readers. In *Proceedings of the 18th International Symposium on Open Collaboration* (<conf-loc>, <city>Madrid</city>, <country>Spain</country>, </conf-loc>) *(OpenSym '22)*. Association for Computing Machinery, New York, NY, USA, Article 9, 10 pages. https://doi.org/10.1145/3555051.3555052

[86] Ahmed A Elngar, Mohamed Gamal, Amar Fathy, Basma Moustafa, Omar Mahmoud, and Mohamed Shaban. 2021. Plagiarism detection algorithm model based on NLP technology. *Journal of Cybersecurity and Information Management* 5, 1 (2021), 43–61.

[87] Daniel Enriquez, Gage Christensen, Hayden Donovan, Jared Lam, Noah Wong, Sergiu Dascalu, David Feil-Seifer, and Emily Hand. 2023. Authorship Verification for Hired Plagiarism Detection. In *Proceedings of the 9th International Conference on Applied Computing & Information Technology* (Virtual Event, USA) *(ACIT '22)*. Association for Computing Machinery, New York, NY, USA, 19–24. https://doi.org/10.1145/3543895.3543928

[88] Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication* 57, 1 (2007), 163–173.

[89] Diego Esteves, Aniketh Janardhan Reddy, Piyush Chawla, and Jens Lehmann. 2018. Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. 50–59.

[90] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In Plain Sight: Media Bias Through the Lens of Factual Reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 6343–6349. https://doi.org/10.18653/v1/D19-1664

[91] KJ Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X Zhang. 2023. Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–42.

[92] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@SEPLN (CEUR Workshop Proceedings, Vol. 2150)*. CEUR-WS.org, 214–228.

[93] Raphael Antonius Frick, Inna Vogel, and Jeong-Eun Choi. 2023. Fraunhofer SIT at CheckThat!-2023: Enhancing the Detection of Multimodal and Multigenre Check-Worthiness Using Optical Character Recognition and Model Souping.. In *CLEF (Working Notes)*. 337–350.

[94] Norbert Fuhr, Anastasia Giachanou, Gregory Grefenstette, Iryna Gurevych, Andreas Hanselowski, Kalervo Jarvelin, Rosie Jones, YiquN Liu, Josiane Mothe, Wolfgang Nejdl, Isabella Peters, and Benno Stein. 2018. An Information Nutritional Label for Online Documents. *SIGIR Forum* 51, 3 (Feb. 2018), 46–66. https://doi.org/10.1145/3190580.3190588

[95] Qin Gao, Ye Tian, and Mengyuan Tu. 2015. Exploring factors influencing Chinese user's perceived credibility of health and safety information on Weibo. *Computers in Human Behavior* 45 (2015), 21–31. https://doi.org/10.1016/j.chb.2014.11.071

[96] Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Varna, Bulgaria, 267–276. https://doi.org/10.26615/978-954-452-049-6_037

[97] Erfan Ghadery, Damien Sileo, and Marie-Francine Moens. 2021. LIIR at SemEval-2021 task 6: Detection of Persuasion Techniques In Texts and Images using CLIP features. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 1015–1019. https://doi.org/10.18653/v1/2021.semeval-1.139

[98] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. https://doi.org/10.48550/arXiv.2301.04246 arXiv:2301.04246 [cs].

[99] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*. Association for Computational Linguistics, 845–854.

[100] Katarína Greškovičová, Radomír Masaryk, Nikola Synak, and Vladimíra Čavojová. 2022. Superlatives, clickbaits, appeals to authority, poor grammar, or boldface: Is editorial style related to the credibility of online health messages? *Frontiers in Psychology* 13 (2022). https://doi.org/10.3389/fpsyg.2022.940903

[101] Jack Grieve and Helena Woodfield. 2023. *The Language of Fake News*. Cambridge University Press.

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

[103] Ashim Gupta and Vivek Srikumar. 2021. X-Fact: A New Benchmark Dataset for Multilingual 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 2: Short Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 675–682. https://doi.org/10.18653/v1/2021.acl-short.86

[104] Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Volta at SemEval-2021 Task 6: Towards Detecting Persuasive Texts and Images using Textual and Multimodal Ensemble. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 1075–1081. https://doi.org/10.18653/v1/2021.semeval-1.149

[105] Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic Encapsulation and Semantic Amalgamation Based Generalised Claim Detection from Online Content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 3178–3188. https://doi.org/10.18653/v1/2021.eacl-main.277

[106] Vansh Gupta and Raksha Sharma. 2021. NLPIITR at SemEval-2021 Task 6: RoBERTa Model with Data Augmentation for Persuasion Techniques Detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 1061–1067. https://doi.org/10.18653/v1/2021.semeval-1.147

[107] Neal Haddaway, Matthew Page, Chris Pritchard, and Luke McGuinness. 2022. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews* 18 (03 2022). https://doi.org/10.1002/cl2.1230

[108] Mohamed Seghir Hadj Ameur and Hassina Aliane. 2021. AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset. *Procedia Computer Science* 189 (2021), 232–241. https://doi.org/10.1016/j.procs.2021.05.086 AI in Computational Linguistics.

[109] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2017. Matrix-based news aggregation: exploring different news perspectives. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (Toronto, Ontario, Canada) *(JCDL '17)*. IEEE Press, 69–78. https://doi.org/10.1109/JCDL.2017.7991561

[110] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2020. Automated identification of media bias by word choice and labeling in news articles. In *Proceedings of the 18th Joint Conference on Digital Libraries* (Champaign, Illinois) *(JCDL '19)*. IEEE Press, 196–205. https://doi.org/10.1109/JCDL.2019.00036

[111] Suhaib Kh Hamed, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. 2023. A Review of Fake News Detection Approaches: A Critical Analysis of Relevant Studies and Highlighting Key Challenges Associated with the Dataset, Feature Representation, and Data Fusion. *Heliyon* 9, 10 (Oct. 2023), e20382. https://doi.org/10.1016/j.heliyon.2023.e20382

[112] Hugo Lewi Hammer, Per Erik Solberg, and Lilja Øvrelid. 2014. Sentiment classification of online political discussions: a comparison of a word-based and dependency-based method. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Alexandra Balahur, Erik van der Goot, Ralf Steinberger, and Andres Montoyo (Eds.). Association for Computational Linguistics, Baltimore, Maryland, 90–96. https://doi.org/10.3115/v1/W14-2616

[113] Sakshini Hangloo and Bhavna Arora. 2022. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia Systems* 28, 8 (2022), 2391–2422. https://doi.org/10.1007/s00530-022-00966-y

[114] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. arXiv:2401.12070 [cs.CL] https://arxiv.org/abs/2401.12070

[115] Christian Hansen, Casper Hansen, Jakob Grue Simonsen, Birger Larsen, Stephen Alstrup, and Christina Lioma. 2020. Factuality Checking in News Headlines with Eye Tracking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2013–2016. https://doi.org/10.1145/3397271.3401221

[116] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb (Eds.). Association for Computational Linguistics, Kyiv, Ukraine (Virtual), 72–81. https://aclanthology.org/2021.wanlp-1.8

[117] Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. CrowdChecked: Detecting Previously Fact-Checked Claims in Social Media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 266–285. https://aclanthology.org/2022.aacl-main.22

[118] Robert Harris. 1997. Evaluating Internet Research Sources. http://www.virtualsalt.com/evaluating-internet-research-sources/. Accessed: 2024-10-01.

[119] Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Freihat. 2023. ArAIEval Shared Task: Persuasion Techniques and Disinformation Detection in Arabic Text. In *Proceedings of ArabicNLP 2023*, Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham (Eds.). Association for Computational Linguistics, Singapore (Hybrid), 483–493. https://doi.org/10.18653/v1/2023.arabicnlp-1.44

[120] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 1803–1812. https://doi.org/10.1145/3097983.3098131

[121] Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting Check-worthy Factual Claims in Presidential Debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1835–1838. https://doi.org/10.1145/2806416.2806652

[122] Hans J. G. Hassell, John B. Holbein, and Matthew R. Miles. 2020. There is no liberal media bias in which news stories political journalists choose to cover. *Science Advances* 6 (2020). https://api.semanticscholar.org/CorpusID:215516592

[123] Sandro Hawke. [n. d.]. Technological Approaches to Improving Credibility Assessment on the Web: Final Community Group Report 11 October 2018. https://www.w3.org/2018/10/credibility-tech/.

[124] Hendrik Heuer and Elena L. Glassman. 2024. Reliability Criteria for News Websites. *ACM Trans. Comput.-Hum. Interact.* 31, 2, Article 21 (jan 2024), 33 pages. https://doi.org/10.1145/3635147

[125] Tashin Hossain, Jannatun Naim, Fareen Tasneem, Radiathun Tasnia, and Abu Nowshed Chy. 2021. CSECU-DSG at SemEval-2021 Task 6: Orchestrating Multimodal Neural Architectures for Identifying Persuasion Techniques in Texts and Images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 1088–1095. https://doi.org/10.18653/v1/2021.semeval-1.151

[126] Timo Hromadka, Timotej Smolen, Tomas Remis, Branislav Pecher, and Ivan Srba. 2023. KInITVeraAI at SemEval-2023 Task 3: Simple yet Powerful Multilingual Fine-Tuning for Persuasion Techniques Detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, 629–637. https://doi.org/10.18653/v1/2023.semeval-1.86

[127] Bo Hu, Zhendong Mao, and Yongdong Zhang. 2024. An Overview of Fake News Detection: From a New Perspective. *Fundamental Research* (Feb. 2024). https://doi.org/10.1016/j.fmre.2024.01.017

[128] Dongchen Huang, Yige Zhu, and Eni Mustafaraj. 2019. How Dependable are "First Impressions" to Distinguish between Real and Fake NewsWebsites?. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media* (Hof, Germany) *(HT '19)*. Association for Computing Machinery, New York, NY, USA, 201–210. https://doi.org/10.1145/3342220.3343670

[129] Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving Cross-lingual Fact-checking with Cross-lingual Retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1024–1035. https://aclanthology.org/2022.coling-1.86

[130] Emelia May Hughes, Renee Wang, Prerna Juneja, Tony W Li, Tanushree Mitra, and Amy X. Zhang. 2024. Viblio: Introducing Credibility Signals and Citations to Video-Sharing Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 807, 20 pages. https://doi.org/10.1145/3613904.3642490

[131] Ahmed Samir Hussein, Abu Bakr Soliman Mohammad, Mohamed Ibrahim, Laila Hesham Afify, and Samhaa R. El-Beltagy. 2022. NGU CNLP atWANLP 2022 Shared Task: Propaganda Detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, and Wajdi Zaghouani (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 545–550. https://doi.org/10.18653/v1/2022.wanlp-1.66

[132] Petar Ivanov, Ivan Koychev, Momchil Hardalov, and Preslav Nakov. 2024. Detecting Check-Worthy Claims in Political Debates, Speeches, and Interviews Using Audio Data. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 12011–12015. https://doi.org/10.1109/ICASSP48485.2024.10447064

[133] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 946, 21 pages. https://doi.org/10.1145/3613904.3642473

[134] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 524 (nov 2022), 40 pages. https://doi.org/10.1145/3555637

[135] Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting Check-Worthy Claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Yang Liu, Tim Paek, and Manasi Patwardhan (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 26–30. https://doi.org/10.18653/v1/N18-5006

[136] Israa Jaradat, Haiqi Zhang, and Chengkai Li. 2024. On Detecting Cherry-picking in News Coverage Using Large Language Models. *arXiv e-prints* (2024), arXiv–2401.

[137] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic Detection of Machine Generated Text: A Critical Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 2296–2309. https://doi.org/10.18653/v1/2020.coling-main.208

[138] Caio LM Jeronimo, Claudio EC Campelo, Leandro Balby Marinho, Allan Sales, Adriano Veloso, and Roberta Viola. 2020. Computing with subjectivity lexicons. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 3272–3280.

[139] Jiaojiao Ji, Yuqi Zhu, and Naipeng Chao. 2023. A comparison of misinformation feature effectiveness across issues and time on Chinese social media. *Information Processing & Management* 60, 2 (2023), 103210. https://doi.org/10.1016/j.ipm.2022.103210

[140] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. https://doi.org/10.1145/3571730

[141] Chenyan Jia, Alexander Boltz, Angie Zhang, Anqing Chen, and Min Kyung Lee. 2022. Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 371 (nov 2022), 27 pages. https://doi.org/10.1145/3555096

[142] Ye Jiang, Xingyi Song, Carolina Scarton, Iknoor Singh, Ahmet Aker, and Kalina Bontcheva. 2023. Categorising Fine-to-Coarse Grained Misinformation: An Empirical Study of the COVID-19 Infodemic. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 556–567. https://aclanthology.org/2023.ranlp-1.61

[143] Prerna Juneja, Wenjuan Zhang, Alison Marie Smith-Renner, Hemank Lamba, Joel Tetreault, and Alex Jaimes. 2024. Dissecting users' needs for search result explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 841, 17 pages. https://doi.org/10.1145/3613904.3642059

[144] Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 Task 11: On RoBERTa-CRF, Span CLS and Whether Self-Training Helps Them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, Barcelona (online), 1415–1424. https://doi.org/10.18653/v1/2020.semeval-1.187

[145] Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. 2017. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management* 53, 5 (2017), 1043–1061.

[146] Byungkyu Kang, John O'Donovan, and Tobias Höllerer. 2012. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (Lisbon, Portugal) *(IUI '12)*. Association for Computing Machinery, New York, NY, USA, 179–188. https://doi.org/10.1145/2166966.2166998

[147] Uku Kangur, Roshni Chakraborty, and Rajesh Sharma. 2024. Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes. *arXiv preprint arXiv:2406.12444* (2024).

[148] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright Violations and Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 7403–7412.

[149] Yavuz Selim Kartal and Mucahid Kutlu. 2020. TrClaim-19: The First Collection for Turkish Check-Worthy Claim Detection with Annotator Rationales. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Raquel Fernández and Tal Linzen (Eds.). Association for Computational Linguistics, Online, 386–395. https://doi.org/10.18653/v1/2020.conll-1.31

[150] Ashkan Kazemi, Artem Abzaliev, Naihao Deng, Rui Hou, Scott Hale, Veronica Perez-Rosas, and Rada Mihalcea. 2023. Query Rewriting for Effective Misinformation Discovery. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 398–407. https://doi.org/10.18653/v1/2023.ijcnlp-main.26

[151] Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4504–4517. https://doi.org/10.18653/v1/2021.acl-long.347

[152] Sopan Khosla, Rishabh Joshi, Ritam Dutt, Alan W Black, and Yulia Tsvetkov. 2020. LTIatCMU at SemEval-2020 Task 11: Incorporating Multi-Level Features for Multi-Granular Propaganda Span Identification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, Barcelona (online), 1756–1763. https://doi.org/10.18653/v1/2020.semeval-1.230

[153] Carita Kiili, Leena Laurinen, and Miika Marttunen. 2007. How students evaluate credibility and relevance of information on the internet? *IADIS International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2007* (01 2007).

[154] Moonsung Kim and Steven Bethard. 2020. TTUI at SemEval-2020 Task 11: Propaganda Detection with Transfer Learning and Ensembles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, Barcelona (online), 1829–1834. https://doi.org/10.18653/v1/2020.semeval-1.240

[155] Spiro Kiousis. 2001. Public Trust or Mistrust? Perceptions of Media Credibility in the Information Age. *Mass Communication and Society* 4, 4 (Nov. 2001), 381–403. https://doi.org/10.1207/S15327825MCS0404_4

[156] V. Klemperer and M. Brady. 2006. *Language of the Third Reich: LTI: Lingua Tertii Imperii*. Bloomsbury Academic. https://books.google.sk/books?id=kwsleqxx_SMC

[157] Akihiro Kondo and Hironobu Abe. 2023. Travel blogger credibility scoring method for recommendation system. In *2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. 1–4. https://doi.org/10.1109/IMCOM56909.2023.10035643

[158] Jan-David Krieger, Timo Spinde, Terry Ruas, Juhi Kulshrestha, and Bela Gipp. 2022. A domain-adaptive pre-training approach for language bias detection in news. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 7 pages. https://doi.org/10.1145/3529372.3530932

[159] Salima Lamsiyah, Abdelkader El Mahdaouy, Hamza Alami, Ismail Berrada, and Christoph Schommer. 2023. UL & UM6P at ArAIEval Shared Task: Transformer-based model for Persuasion Techniques and Disinformation detection in Arabic. In *Proceedings of ArabicNLP 2023*, Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham (Eds.). Association for Computational Linguistics, Singapore (Hybrid), 558–564. https://doi.org/10.18653/v1/2023.arabicnlp-1.55

[160] Patrick Lawson, Carl J. Pearson, Aaron Crowson, and Christopher B. Mayhorn. 2020. Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy. *Applied Ergonomics* 86 (2020), 103084. https://doi.org/10.1016/j.apergo.2020.103084

[161] Konstantina Lazaridou, Ralf Krestel, and Felix Naumann. 2017. Identifying Media Bias by Analyzing Reported Speech. In *2017 IEEE International Conference on Data Mining (ICDM)*. 943–948. https://doi.org/10.1109/ICDM.2017.119

[162] Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event Detection and Factuality Assessment with Non-Expert Supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lluís Màrquez, Chris Callison-Burch, and Jian Su (Eds.). Association for Computational Linguistics, Lisbon, Portugal, 1643–1648. https://doi.org/10.18653/v1/D15-1189

[163] João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. Weakly Supervised Veracity Classification with LLM-Predicted Credibility Signals. arXiv:2309.07601 [cs.CL] https://arxiv.org/abs/2309.07601

[164] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 10528–10539.

[165] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[166] Chang Li and Dan Goldwasser. 2021. MEAN: Multi-head Entity Aware Attention Networkfor Political Perspective Detection in News Media. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, Anna Feldman, Giovanni Da San Martino, Chris Leberknight, and Preslav Nakov (Eds.). Association for Computational Linguistics, Online, 66–75. https://doi.org/10.18653/v1/2021.nlp4if-1.10

[167] Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024. MAVEN-Fact: A Large-scale Event Factuality Detection Dataset. *arXiv preprint arXiv:2407.15352* (2024).

[168] Xiaojia Li, Yun Zhang, and Zhong Qian. 2020. An End-to-End Approach for Document-level Event Factuality Identification in Chinese. In *2020 International Conference on Asian Language Processing (IALP)*. 215–220. https://doi.org/10.1109/IALP51396.2020.9310484

[169] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A Multilingual and Multimodal Data Repository for Combating COVID-19 Disinformation. arXiv:2011.04088 [cs.SI] https://arxiv.org/abs/2011.04088

[170] Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated Text Detection in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 36–53. https://doi.org/10.18653/v1/2024.acl-long.3

[171] Yufeng Li, Rrubaa Panchendrarajan, and Arkaitz Zubiaga. 2024. FactFinders at CheckThat! 2024: Refining Check-worthy Statement Detection with LLMs through Data Pruning. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024 (CEUR Workshop Proceedings, Vol. 3740)*, Guglielmo Faggioli, Nicola Ferro, Petra Galuscáková, and Alba García Seco de Herrera (Eds.). CEUR-WS.org, 520–537. https://ceur-ws.org/Vol-3740/paper-47.pdf

[172] Chenghua Lin, Yulan He, and Richard Everson. 2011. Sentence Subjectivity Detection with Weakly-Supervised Learning. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, Haifeng Wang and David Yarowsky (Eds.). Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 1153–1161. https://aclanthology.org/I11-1129

[173] Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. 2024. IndiVec: An Exploration of Leveraging Large Language Models for Media Bias Detection with Fine-Grained Bias Indicators. In *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1038–1050. https://aclanthology.org/2024.findings-eacl.70

[174] Szu-Yin Lin, Yun-Ching Kung, and Fang-Yie Leu. 2022. Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis. *Information Processing & Management* 59, 2 (2022), 102872. https://doi.org/10.1016/j.ipm.2022.102872

[175] Xialing Lin, Patric R. Spence, and Kenneth A. Lachlan. 2016. Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior* 63 (2016), 264–271. https://doi.org/10.1016/j.chb.2016.05.002

[176] Eunice Ling, Domenico de Pieri, Evenne Loh, Karen M Scott, Stephen C H Li, and Heather J Medbury. 2024. Evaluation of the Accuracy, Credibility, and Readability of Statin-Related Websites: Cross-Sectional Study. *Interact J Med Res* 13 (14 Mar 2024), e42849. https://doi.org/10.2196/42849

[177] Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Mohit Bansal and Aline Villavicencio (Eds.). Association for Computational Linguistics, Hong Kong, China, 504–514. https://doi.org/10.18653/v1/K19-1047

[178] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[179] Hao Lu, Yanwu Yang, Runsheng Gan, and Nan Zhang. 2012. The research on micro-blog public opinion index and the application of prototype system. In *Proceedings of 2012 9th IEEE International Conference on Networking, Sensing and Control*. 405–410. https://doi.org/10.1109/ICNSC.2012.6204953

[180] Qiang Lu, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. 2024. Fact-sentiment incongruity combination network for multimodal sarcasm detection. *Information Fusion* 104 (2024), 102203. https://doi.org/10.1016/j.inffus.2023.102203

[181] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The effects of ai-based credibility indicators on the detection and spread of misinformation under social influence. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.

[182] Jason Lucas, Limeng Cui, Thai Le, and Dongwon Lee. 2022. Detecting False Claims in Low-Resource Regions: A Case Study of Caribbean Islands. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, Tanmoy Chakraborty, Md. Shad Akhtar, Kai Shu, H. Russell Bernard, Maria Liakata, Preslav Nakov, and Aseem Srivastava (Eds.). Association for Computational Linguistics, Dublin, Ireland, 95–102. https://doi.org/10.18653/v1/2022.constraint-1.11

[183] Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14279–14305. https://doi.org/10.18653/v1/2023.emnlp-main.883

[184] Emma Lurie and Eni Mustafaraj. 2018. Investigating the Effects of Google's Search Engine Result Page in Evaluating the Credibility of Online News Sources. In *Proceedings of the 10th ACM Conference on Web Science* (Amsterdam, Netherlands) *(WebSci '18)*. Association for Computing Machinery, New York, NY, USA, 107–116. https://doi.org/10.1145/3201064.3201095

[185] Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. An Effective Approach for Informational and Lexical Bias Detection. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 66–77. https://doi.org/10.18653/v1/2023.fever-1.7

[186] Fabrizio Macagno. 2022. Argumentation profiles and the manipulation of common ground. The arguments of populist leaders on Twitter. *Journal of Pragmatics* 191 (2022), 67–82. https://doi.org/10.1016/j.pragma.2022.01.022

[187] Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9960–9987. https://doi.org/10.18653/v1/2023.emnlp-main.616

[188] Dominik Macko, Robert Moro, Adaku Uchendu, Ivan Srba, Jason Samuel Lucas, Michiharu Yamashita, Nafis Irtiza Tripto, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2024. Authorship Obfuscation in Multilingual Machine-Generated Text Detection. arXiv:2401.07867 [cs.CL] https://arxiv.org/abs/2401.07867

[189] Isa Maks and Piek Vossen. 2012. Building a fine-grained subjectivity lexicon from a web corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 3070–3076. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1018_Paper.pdf

[190] Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2022. Did I See It Before? Detecting Previously-Checked Claims over Twitter. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 367–381.

[191] Alejandro Martín, Javier Huertas-Tato, Álvaro Huertas-García, Guillermo Villar-Rodríguez, and David Camacho. 2022. FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems* 251 (2022), 109265. https://doi.org/10.1016/j.knosys.2022.109265

[192] Mohsen Mesgar and Michael Strube. 2018. A Neural Local Coherence Model for Text Quality Assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 4328–4339. https://doi.org/10.18653/v1/D18-1464

[193] Clyde R Miller. 1939. The techniques of propaganda. From "how to detect and analyze propaganda," an address given at town hall. *The Center for learning* (1939).

[194] Tanushree Mitra and Eric Gilbert. 2021. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (Aug. 2021), 258–267. https://doi.org/10.1609/icwsm.v9i1.14625

[195] Shubham Mittal and Preslav Nakov. 2022. IITD at WANLP 2022 Shared Task: Multilingual Multi-Granularity Network for Propaganda Detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, and Wajdi Zaghouani (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 529–533. https://doi.org/10.18653/v1/2022.wanlp-1.63

[196] Shubham Mittal, Megha Sundriyal, and Preslav Nakov. 2023. Lost in Translation, Found in Spans: Identifying Claims in Multilingual Social Media. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3887–3902. https://doi.org/10.18653/v1/2023.emnlp-main.236

[197] Aditya Mogadala and Vasudeva Varma. 2012. Language independent sentence-level subjectivity analysis with feature selection. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*. 171–180.

[198] Maria D Molina, S Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "Fake News" Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. *American Behavioral Scientist* 65, 2 (2021), 180–212. https://doi.org/10.1177/0002764219878224

[199] Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 Task 11: An Empirical Study of Pre-Trained Transformer Family for Propaganda Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (Eds.). International Committee for Computational Linguistics, Barcelona (online), 1739–1748. https://doi.org/10.18653/v1/2020.semeval-1.228

[200] Alejandro Mosquera and Paloma Moreda. 2021. SMILE: An Informality Classification Tool for Helping to Assess Quality and Credibility in Web 2.0 Texts. *Proceedings of the International AAAI Conference on Web and Social Media* 6, 3 (Aug. 2021), 2–7. https://doi.org/10.1609/icwsm.v6i3.14347

[201] Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*. 55–64.

[202] Felix B Mueller, Rebekka Görge, Anna K Bernzen, Janna C Pirk, and Maximilian Poretschkin. 2024. LLMs and Memorization: On Quality and Specificity of Copyright Compliance. *arXiv preprint arXiv:2405.18492* (2024).

[203] Smruthi Mukund and Rohini K Srihari. 2010. A vector space model for subjectivity classification in Urdu aided by co-training. In *Coling 2010: Posters*. 860–868.

[204] Multiple authors and Sandro Hawke. 2019. Credibility Signals: Unofficial Draft 26 November 2019. https://credweb.org/signals-20191126.

[205] Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mücahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, Alex Nikolov, and Yavuz Selim Kartal. 2022. Overview of the CLEF-2022 CheckThat! Lab Task 1 on Identifying Relevant Claims in Tweets. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022 (CEUR Workshop Proceedings, Vol. 3180)*, Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast (Eds.). CEUR-WS.org, 368–392. https://ceur-ws.org/Vol-3180/paper-28.pdf

[206] Preslav Nakov, Giovanni Da San Martino, Firoj Alam, Shaden Shaar, Hamdy Mubarak, and Nikolay Babulkov. 2022. Overview of the CLEF-2022 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, Vol. 3180. CEUR-WS, 11. https://ceur-ws.org/Vol-3180/paper-29.pdf

[207] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mücahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 12880)*, K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro (Eds.). Springer, 264–291. https://doi.org/10.1007/978-3-030-85251-1_19

[208] Preslav Nakov, Husrev Taha Sencar, Jisun An, and Haewoon Kwak. 2021. A survey on predicting the factuality and the bias of news media. *arXiv preprint arXiv:2103.12506* (2021).

[209] Eric Ness, Arooj Fatima, and Mahdi Maktabdar Oghaz. 2023. Data Driven Model to Investigate Political Bias in Mainstream Media. *IEEE Access* 11 (2023), 41880–41893. https://doi.org/10.1109/ACCESS.2023.3270630

[210] Dan S. Nielsen and Ryan McConville. 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* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3141–3153. https://doi.org/10.1145/3477495.3531744

[211] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. 2013. Web Credibility: Features Exploration and Credibility Prediction. In *Advances in Information Retrieval*, Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 557–568.

[212] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[213] Laura O'Grady. 2006. Future directions for depicting credibility in health care web sites. *International Journal of Medical Informatics* 75, 1 (2006), 58–65. https://doi.org/10.1016/j.ijmedinf.2005.07.035 Health and the Internet for All.

[214] Matthew Page, David Moher, Patrick Bossuyt, Isabelle Boutron, Tammy Hoffmann, Cynthia Mulrow, Larissa Shamseer, Jennifer Tetzlaff, Elie Akl, Sue Brennan, Roger Chou, Julie Glanville, Jeremy Grimshaw, Asbjørn Hróbjartsson, Manoj Lalu, Tianjing Li, Elizabeth Loder, Evan Mayo-Wilson, Steve Mcdonald, and Joanne Mckenzie. 2021. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ* 372 (03 2021), n160. https://doi.org/10.1136/bmj.n160

[215] Surya Kant Pal, Omar Junior Raffik, Rita Roy, Vishwakarma Bipin Lalman, Shristee Srivastava, and Bhoomi Sharma. 2023. Automatic Plagiarism Detection Using Natural Language Processing. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*. 218–222.

[216] Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal* 7 (2024), 100066. https://doi.org/10.1016/j.nlp.2024.100066

[217] Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. Barcelona, Spain, 271–278. https://doi.org/10.3115/1218955.1218990

[218] Jana Papcunová, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. 2023. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & Intelligent Systems* 9, 3 (01 Jun 2023), 2827–2842. https://doi.org/10.1007/s40747-021-00561-0

[219] Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. 2019. A Link-based Approach to Detect Media Bias in News Websites. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 742–745. https://doi.org/10.1145/3308560.3316460

[220] Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. TATHYA: A Multi-Classifier System for Detecting Check-Worthy Statements in Political Debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 2259–2262. https://doi.org/10.1145/3132847.3133150

[221] Amalie Pauli, Rafael Sarabia, Leon Derczynski, and Ira Assent. 2023. TeamAmpa at SemEval-2023 Task 3: Exploring Multilabel and Multilingual RoBERTa Models for Persuasion and Framing Detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, 847–855. https://doi.org/10.18653/v1/2023.semeval-1.117

[222] John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 Task 5: Toxic Spans Detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurélie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, 59–69. https://doi.org/10.18653/V1/2021.SEMEVAL-1.6

[223] Branislav Pecher, Jan Cegin, Robert Belanec, Jakub Simko, Ivan Srba, and Maria Bielikova. 2024. Fighting Randomness with Randomness: Mitigating Optimisation Instability of Fine-Tuning using Delayed Ensemble and Noisy Interpolation. arXiv:2406.12471 [cs.CL] https://arxiv.org/abs/2406.12471

[224] Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. Comparing Specialised Small and General Large Language Models on Text Classification: 100 Labelled Samples to Achieve Break-Even Performance. arXiv:2402.12819 [cs.CL] https://arxiv.org/abs/2402.12819

[225] Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. On Sensitivity of Learning with Limited Labelled Data to the Effects of Randomness: Impact of Interactions and Systematic Choices. arXiv:2402.12817 [cs.CL] https://arxiv.org/abs/2402.12817

[226] Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. A Survey on Stability of Learning with Limited Labelled Data and its Sensitivity to the Effects of Randomness. *ACM Comput. Surv.* 57, 1, Article 19 (Oct. 2024), 40 pages. https://doi.org/10.1145/3691339

[227] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[228] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. https://aclanthology.org/C18-1287

[229] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[230] Lara Piccolo, Azizah C Blackwood, Tracie Farrell, and Martino Mensio. 2021. Agents for fighting misinformation spread on Twitter: design challenges. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–7.

[231] Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual Previously Fact-Checked Claim Retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 16477–16500. https://doi.org/10.18653/v1/2023.emnlp-main.1027

[232] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, 2343–2361. https://doi.org/10.18653/v1/2023.semeval-1.317

[233] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: a unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii) *(EMNLP '08)*. Association for Computational Linguistics, USA, 186–195.

[234] León Podgurski, Karolina Zaczynska, and Georg Rehm. 2022. Evaluating Web Content Using the W3C Credibility Signals. https://doi.org/10.3233/SSW220005

[235] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, Vol. 2003. Lancaster, UK, 40.

[236] Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2015. A two-step approach for event factuality identification. In *2015 International Conference on Asian Language Processing (IALP)*. 103–106. https://doi.org/10.1109/IALP.2015.7451542

[237] Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2019. Document-Level Event Factuality Identification via Adversarial Neural Network. 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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 2799–2809. https://doi.org/10.18653/v1/N19-1287

[238] Zhong Qian, Heng Zhang, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2022. Document-level event factuality identification via machine reading comprehension frameworks with transfer learning. In *Proceedings of the 29th International Conference on Computational Linguistics*. 2622–2632.

[239] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with Language Model Prompting: A Survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5368–5393. https://doi.org/10.18653/v1/2023.acl-long.294

[240] Khubaib Ahmed Qureshi and Rauf Ahmed Shams Malick. 2023. Explainable tweet credibility ranker: A comprehensive credibility solution. *Computers and Electrical Engineering* 112 (2023), 109028. https://doi.org/10.1016/j.compeleceng.2023.109028

[241] Khubaib Ahmed Qureshi, Rauf Ahmed Shams Malick, and Muhammad Sabih. 2021. Social Media and Microblogs Credibility: Identification, Theory Driven Framework, and Recommendation. *IEEE Access* 9 (2021), 137744–137781. https://doi.org/10.1109/ACCESS.2021.3114417

[242] Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3027–3040.

[243] Olesya Razuvayevskaya, Ben Wu, João A. Leite, Freddy Heppell, Ivan Srba, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. *PLOS ONE* 19, 5 (05 2024), 1–26. https://doi.org/10.1371/journal.pone.0301738

[244] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[245] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric Loss for Multi-Label Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 82–91.

[246] Jens Riegelsberger, M. Angela Sasse, and John D. McCarthy. 2005. Do people trust their eyes more than ears? media bias in detecting cues of expertise. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1745–1748. https://doi.org/10.1145/1056808.1057012

[247] Soo Young Rieh and David R. Danielson. 2007. Credibility: A Multidisciplinary Framework. *Annual Review of Information Science and Technology* 41, 1 (Jan. 2007), 307–364. https://doi.org/10.1002/aris.2007.1440410114

[248] Angela Rijo and Sven Waldzus. 2023. That's interesting! The role of epistemic emotions and perceived credibility in the relation between prior beliefs and susceptibility to fake-news. *Computers in Human Behavior* 141 (2023), 107619. https://doi.org/10.1016/j.chb.2022.107619

[249] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[250] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2023. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications* (2023), 121641.

[251] Cees Roele. 2021. WVOQ at SemEval-2021 Task 6: BART for Span Detection and Classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 270–274. https://doi.org/10.18653/v1/2021.semeval-1.32

[252] Victoria L. Rubin and Elizabeth D. Liddy. 2006. Assessing Credibility of Weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. https://api.semanticscholar.org/CorpusID:28691245

[253] Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri-Jacques Geiss, Paolo Rosso, and Lucie Flek. 2022. Factoid: A new dataset for identifying misinformation spreaders and political bias. *arXiv preprint arXiv:2205.06181* (2022).

[254] Noushin Salek Faramarzi, Fateme Hashemi Chaleshtori, Hossein Shirazi, Indrakshi Ray, and Ritwik Banerjee. 2023. Claim Extraction and Dynamic Stance Detection in COVID-19 Tweets. In *Companion Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 1059–1068. https://doi.org/10.1145/3543873.3587643

[255] Allan Sales, Albin Zehe, Leandro Balby Marinho, Adriano Veloso, Andreas Hotho, and Janna Omeliyanenko. 2021. Assessing Media Bias in Cross-Linguistic and Cross-National Populations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 561–572.

[256] Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10 (2020), 1–34.

[257] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5884–5906.

[258] Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language resources and evaluation* 43 (2009), 227–268.

[259] Aleksandar Savchev. 2022. AI Rational at CheckThat!-2022: Using transformer models for tweet classification. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022 (CEUR Workshop Proceedings, Vol. 3180)*, Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast (Eds.). CEUR-WS.org, 656–659. https://ceur-ws.org/Vol-3180/paper-52.pdf

[260] Elena Savinova and Fermin Moscoso Del Prado. 2023. Analyzing Subjectivity Using a Transformer-Based Regressor Trained on Naïve Speakers' Judgements. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. 305–314.

[261] Marcin Sawinski, Krzysztof Wecel, Ewelina Ksiezniak, Milena Strózyna, Wlodzimierz Lewoniewski, Piotr Stolarski, and Witold Abramowicz. 2023. OpenFact at CheckThat!-2023: Head-to-Head GPT vs. BERT - A Comparative Study of Transformers Language Models for the Detection of Check-worthy Claims. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023 (CEUR Workshop Proceedings, Vol. 3497)*, Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro, and Michalis Vlachos (Eds.). CEUR-WS.org, 453–472. https://ceur-ws.org/Vol-3497/paper-040.pdf

[262] Ipek Baris Schlicht, Lucie Flek, and Paolo Rosso. 2023. Multilingual Detection of Check-Worthy Claims Using World Languages and Adapter Fusion. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I* (Dublin, Ireland). Springer-Verlag, Berlin, Heidelberg, 118–133. https://doi.org/10.1007/978-3-031-28244-7_8

[263] Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1245–1254.

[264] Charles C. Self and Chris Roberts. 2019. Credibility. In *An Integrated Approach to Communication Theory and Research* (3 ed.). Routledge.

[265] Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The Role of Context in Detecting Previously Fact-Checked Claims. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1619–1631. https://doi.org/10.18653/v1/2022.findings-naacl.122

[266] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3607–3618. https://doi.org/10.18653/v1/2020.acl-main.332

[267] Shaden Shaar, Nikola Georgiev, Firoj Alam, Giovanni Da San Martino, Aisha Mohamed, and Preslav Nakov. 2022. Assisting the Human Fact-Checkers: Detecting All Previously Fact-Checked Claims in a Document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2069–2080. https://doi.org/10.18653/v1/2022.findings-emnlp.151

[268] Shaden Shaar, Fatima Haouari, Watheq Mansour, Maram Hasanain, Nikolay Babulkov, Firoj Alam, and Preslav Nakov. 2021. Overview of the CLEF-2021 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims in Tweets and Political Debates. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, Vol. 2936. CEUR-WS, 13. http://ceur-ws.org/Vol-2936/paper-29.pdf

[269] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, Yan Liu, Feng Qian was, Feng Qian, He Jiang, Natali Ruchansky, Yan Liu, Ming Zhang, Laks V. S. Lakshmanan, Michael Simpson, Saravanan Thirumuruganathan, Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology* 10, 3 (2019), 21. https://doi.org/10.1145/3305260 arXiv:1901.06437

[270] Zien Sheikh Ali, Watheq Mansour, Fatima Haouari, Maram Hasanain, Tamer Elsayed, and Abdulaziz Al-Ali. 2023. Tahaqqaq: A Real-Time System for Assisting Twitter Users in Arabic Claim Verification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 3170–3174. https://doi.org/10.1145/3539618.3591815

[271] Ghazaal Sheikhi, Samia Touileb, and Sohail Khan. 2023. Automated Claim Detection for Fact-checking: A Case Study using Norwegian Pre-trained Language Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Tanel Alumäe and Mark Fishel (Eds.). University of Tartu Library, Tórshavn, Faroe Islands, 1–9. https://aclanthology.org/2023.nodalida-1.1

[272] Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article Reranking by Memory-Enhanced Key Sentence Matching for Detecting Previously Fact-Checked Claims. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5468–5481. https://doi.org/10.18653/v1/2021.acl-long.425

[273] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* 8, 3 (2020), 171–188. https://doi.org/10.1089/big.2020.0062 arXiv:https://doi.org/10.1089/big.2020.0062 PMID: 32491943.

[274] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (Sept. 2017), 22–36. https://doi.org/10.1145/3137597.3137600 arXiv:1708.01967

[275] Erica Shusas and Andrea Forte. 2024. Trust and Transparency: An Exploratory Study on Emerging Adults' Interpretations of Credibility Indicators on Social Media Platforms. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 343, 7 pages. https://doi.org/10.1145/3613905.3650801

[276] Felix M Simon, Sacha Altay, and Hugo Mercier. 2023. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review* 4, 5 (2023).

[277] Iknoor Singh, Carolina Scarton, Xingyi Song, and Kalina Bontcheva. 2024. Breaking Language Barriers with MMTweets: Advancing Cross-Lingual Debunked Narrative Retrieval for Fact-Checking. arXiv:2308.05680 [cs.CL] https://arxiv.org/abs/2308.05680

[278] Niraj Sitaula, Chilukuri K. Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. *Credibility-Based Fake News Detection*. Springer International Publishing, Cham, 163–182. https://doi.org/10.1007/978-3-030-42699-6_9

[279] Jackie Smith, John D. McCarthy, Clark Mcphail, and Boguslaw Augustyn. 2001. From Protest to Agenda Building: Description Bias in Media Coverage of Protest Events in Washington, D.C. *Social Forces* 79 (2001), 1397 – 1423. https://api.semanticscholar.org/CorpusID:145456088

[280] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. arXiv:1908.09203 [cs.CL] https://arxiv.org/abs/1908.09203

[281] Aastik Solanki, Adesh Kumar Jaiswal, Aditi Mishra, Anant Prakash Mishra, and Suruchi Sabherwal. 2021. Machine Learning-Based Identification of Collusive Users in Twitter Stream. In *2021 International Conference on Computational Performance Evaluation (ComPE)*. 797–802. https://doi.org/10.1109/ComPE53109.2021.9752090

[282] Stuart N Soroka. 2012. The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics* 74, 2 (2012), 514–528.

[283] Francesca Spezzano, Anu Shrestha, Jerry Alan Fails, and Brian W. Stone. 2021. That's Fake News! Reliability of News When Provided Title, Image, Source Bias & Full Article. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 109 (apr 2021), 19 pages. https://doi.org/10.1145/3449183

[284] Michal Spiegel and Dominik Macko. 2024. KInIT at SemEval-2024 Task 8: Fine-tuned LLMs for Multilingual Machine-Generated Text Detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 558–564. https://doi.org/10.18653/v1/2024.semeval-1.84

[285] Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021. Do You Think It's Biased? How To Ask For The Perception Of Media Bias. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 61–69. https://doi.org/10.1109/JCDL52503.2021.00018

[286] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 1166–1177. https://doi.org/10.18653/v1/2021.findings-emnlp.101

[287] Timo Spinde, Elisabeth Richter, Martin Wessel, Juhi Kulshrestha, and Karsten Donnay. 2023. What do Twitter comments tell about news article bias? Assessing the impact of news article bias on its perception on Twitter. *Online Social Networks and Media* 37-38 (2023), 100264. https://doi.org/10.1016/j.osnem.2023.100264

[288] Timo Spinde, Lada Rudnitckaia, Felix Hamborg, and Bela Gipp. 2021. Identification of Biased Terms in News Articles by Comparison of Outlet-Specific Word Embeddings. In *Diversity, Divergence, Dialogue*, Katharina Toeppe, Hui Yan, and Samuel Kai Wah Chu (Eds.). Springer International Publishing, Cham, 215–224.

[289] Timo Spinde, Lada Rudnitckaia, Jelena Mitrović, Felix Hamborg, Michael Granitzer, Bela Gipp, and Karsten Donnay. 2021. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management* 58, 3 (2021), 102505. https://doi.org/10.1016/j.ipm.2021.102505

[290] Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. 2023. From Chaos to Clarity: Claim Normalization to Empower Fact-Checking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6594–6609. https://doi.org/10.18653/v1/2023.findings-emnlp.439

[291] Abhijit Suprem, Sanjyot Vaidya, and Calton Pu. 2022. Exploring Generalizability of Fine-Tuned Models for Fake News Detection. In *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*. IEEE Computer Society, 82–88. https://doi.org/10.1109/CIC56439.2022.00022

[292] Bahi Takkouche and Guy Norman. 2011. PRISMA statement. *Epidemiology* 22, 1 (2011), 128.

[293] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In *The Semantic Web – ISWC 2019*, Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon (Eds.). Springer International Publishing, Cham, 309–324.

[294] Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, and Wenming Xiao. 2021. MinD at SemEval-2021 Task 6: Propaganda Detection using Transfer Learning and Multimodal Fusion. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu (Eds.). Association for Computational Linguistics, Online, 1082–1087. https://doi.org/10.18653/v1/2021.semeval-1.150

[295] Minh Tran. 2020. How biased are American media outlets? A framework for presentation bias regression. In *2020 IEEE International Conference on Big Data (Big Data)*. 4359–4364. https://doi.org/10.1109/BigData50022.2020.9377987

[296] Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A Comprehensive Survey of Multimodal Fake News Detection Techniques: Advances, Challenges, and Opportunities. *International Journal of Multimedia Information Retrieval* 12, 2 (Aug. 2023), 28. https://doi.org/10.1007/s13735-023-00296-3

[297] Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and Obfuscation of Neural Text Authorship: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 25, 1 (July 2023), 1–18. https://doi.org/10.1145/3606274.3606276

[298] Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TuringBench: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics,

Punta Cana, Dominican Republic, 2001–2016. https://doi.org/10.18653/v1/2021.findings-emnlp.172

[299] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[300] Nguyen Vo and Kyumin Lee. 2018. The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3209978.3210037

[301] Nguyen Vo and Kyumin Lee. 2020. Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7717–7731. https://doi.org/10.18653/v1/2020.emnlp-main.621

[302] Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind Your Format: Towards Consistent Evaluation of In-Context Learning Improvements. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 6287–6310. https://doi.org/10.18653/v1/2024.findings-acl.375

[303] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation Capabilities of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 14830–14847. https://doi.org/10.18653/v1/2024.acl-long.793

[304] Paul Waldman and James Devitt. 1998. Newspaper photographs and the 1996 presidential election: The question of bias. *Journalism & Mass Communication Quarterly* 75, 2 (1998), 302–311.

[305] Ace Wang, Liz Maylin De Jesus Sanchez, Anya Wintner, Yuanxin Zhu, and Eni Mustafaraj. 2023. Assessing Google Search's New Features in Supporting Credibility judgments of Unknown Websites. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval (CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 303–307. https://doi.org/10.1145/3576840.3578277

[306] Dong Wang and Yang Liu. 2011. A cross-corpus study of unsupervised subjectivity identification based on calibrated em. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*. 161–167.

[307] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 422–426.

[308] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2057–2079. https://doi.org/10.18653/v1/2024.semeval-1.279

[309] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4GT-Bench: Evaluation Benchmark for Black-Box Machine-Generated Text Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3964–3992. https://doi.org/10.18653/v1/2024.acl-long.218

[310] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[311] Wouter Weerkamp and Maarten De Rijke. 2008. Credibility improves topical blog post retrieval. In *Proceedings of ACL-08: HLT*. 923–931.

[312] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[313] Martin Wessel and Tomáš Horych. 2024. Beyond the Surface: Spurious Cues in Automatic Media Bias Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*. 21–30.

[314] Martin Wessel, Tomás Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. Introducing MBIB - The First Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2765–2774. https://doi.org/10.1145/3539618.3591882

[315] Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International conference on intelligent text processing and computational linguistics*. Springer, 486–497.

[316] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics* 30, 3 (2004), 277–308.

[317] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39 (2005), 165–210.

[318] Michael Wiegand and Dietrich Klakow. 2011. Convolution kernels for subjectivity detection. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*. 254–261.

[319] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*. 602–608.

[320] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.

[321] Evan M. Williams, Paul Rodrigues, and Sieu Tran. 2021. Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021 (CEUR Workshop Proceedings, Vol. 2936)*, Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi (Eds.). CEUR-WS.org, 659–669. https://ceur-ws.org/Vol-2936/paper-55.pdf

[322] Tora Sangputra Yopie Winarto, Kevin Wijaya, Muhammad Abdullah Faqih, Simeon Yuda Prasetyo, and Yohan Muliono. 2023. Tackling Clickbait with Machine Learning: A Comparative Study of Binary Classification Models for YouTube Title. *Procedia Computer Science* 227 (2023), 282–290. https://doi.org/10.1016/j.procs.2023.10.526 8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023).

[323] Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A. Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, 1995–2008. https://doi.org/10.18653/v1/2023.semeval-1.275

[324] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2024. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. arXiv:2310.14724 [cs.CL] https://arxiv.org/abs/2310.14724

[325] Lianwei Wu, Yuan Rao, Ambreen Nazir, and Haolin Jin. 2020. Discovering differential features: Adversarial learning for information credibility evaluation. *Inf. Sci.* 516, C (apr 2020), 453–473. https://doi.org/10.1016/j.ins.2019.12.040

[326] Amelie Wuehrl, Lara Grimminger, and Roman Klinger. 2023. An Entity-based Claim Extraction Pipeline for Real-world Biomedical Fact-checking. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 29–37. https://doi.org/10.18653/v1/2023.fever-1.3

[327] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. arXiv:2312.12148 [cs.CL] https://arxiv.org/abs/2312.12148

[328] Hao Yan, Sanmay Das, Allen Lavoie, Sirui Li, and Betsy Sinclair. 2019. The Congressional Classification Challenge: Domain Specificity and Partisan Intensity. In *Proceedings of the 2019 ACM Conference on Economics and Computation* (Phoenix, AZ, USA) *(EC '19)*. Association for Computing Machinery, New York, NY, USA, 71–89. https://doi.org/10.1145/3328526.3329582

[329] Yeqing Yan, Peng Zheng, and Yongjun Wang. 2024. Enhancing large language model capabilities for rumor detection with Knowledge-Powered Prompting. *Engineering Applications of Artificial Intelligence* 133 (2024), 108259. https://doi.org/10.1016/j.engappai.2024.108259

[330] Junting Ye and Steven Skiena. 2019. MediaRank: Computational Ranking of Online News Sources. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2469–2477. https://doi.org/10.1145/3292500.3330709

[331] Qingqing Yi, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2022. Chinese Sentence-level Event Factuality Identification with Recursive Neural Network. In *2022 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892209

[332] Shehel Yoosuf and Yin Yang. 2019. Fine-Grained Propaganda Detection with Fine-Tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov (Eds.). Association for Computational Linguistics, Hong Kong, China, 87–91. https://doi.org/10.18653/v1/D19-5011

[333] Himanshu Zade, Megan Woodruff, Erika Johnson, Mariah Stanley, Zhennan Zhou, Minh Tu Huynh, Alissa Elizabeth Acheson, Gary Hsieh, and Kate Starbird. 2023. Tweet Trajectory and AMPS-based Contextual Cues can Help Users Identify Misinformation. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 103 (apr 2023), 27 pages. https://doi.org/10.1145/3579536

[334] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 75–86. https://doi.org/10.18653/v1/S19-2010

[335] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

[336] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf

[337] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining

and Annotating Credibility Indicators in News Articles. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 603–612. https://doi.org/10.1145/3184558.3188731

[338] Canyu Zhang and Paul D. Clough. 2020. Investigating clickbait in Chinese social media: A study of WeChat. *Online Social Networks and Media* 19 (2020), 100095. https://doi.org/10.1016/j.osnem.2020.100095

[339] Heng Zhang, Peifeng Li, Zhong Qian, and Xiaoxu Zhu. 2023. Incorporating factuality inference to identify document-level event factuality. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13990–14002.

[340] Liang Zhao, Ting Hua, Chang-Tien Lu, and Ing-Ray Chen. 2016. A topic-focused trust model for Twitter. *Computer Communications* 76 (2016), 1–11. https://doi.org/10.1016/j.comcom.2015.08.001

[341] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *Comput. Surveys* 53, 5 (Oct. 2020), 1–40. https://doi.org/10.1145/3395046 arXiv:1812.00315

[342] Hafzullah İŞ and Taner TUNCER. 2018. Confidence Index Analysis of Twitter Users Timeline. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. 1–8. https://doi.org/10.1109/IDAP.2018.8620917