Fighting disinformation with artificial intelligence: fundamentals, advances and challenges

Andrés Montoro-Montarroso; Javier Cantón-Correa; Paolo Rosso; Berta Chulvi; Ángel Panizo-Lledot; Javier Huertas-Tato; Blanca Calvo-Figueras; M. José Rementeria; Juan Gómez-Romero

Note: Este artículo se puede leer en español en:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87328

Recommended citation:

Montoro-Montarroso, Andrés; **Cantón-Correa, Javier**; **Rosso, Paolo**; **Chulvi, Berta**; **Panizo-Lledot, Ángel**; **Huertas-Tato, Javier**; **Calvo-Figueras, Blanca**; **Rementeria, M. José**; **Gómez-Romero, Juan** (2023). "Fighting disinformation with artificial intelligence: fundamentals, advances and challenges". *Profesional de la información*, v. 32, n. 3, e320322.
https://doi.org/10.3145/epi.2023.may.22

Manuscript received on 27th March 2023
Accepted on 17th May 2023

Andrés Montoro-Montarroso ✉
https://orcid.org/0000-0003-1893-3346
Universidad de Granada
Decsai
Citic-UGR
Periodista Rafael Gómez Montero, 2
18014 Granada, Spain
andres.montoro@ugr.es

Javier Cantón-Correa
https://orcid.org/0000-0002-8466-1679
Universidad Internacional de La Rioja
Fac. de Ciencias Sociales y Humanidades
Universidad de Granada
Decsai
Citic-UGR, Spain
javicanton@ugr.es

Paolo Rosso
https://orcid.org/0000-0002-8922-1242
Universitat Politècnica de València
Pattern Recognition and Human Language Technologies (*PRHLT*) *Research Center*
Camí de Vera, s/n
46022 Valencia, Spain
prosso@dsic.upv.es

Berta Chulvi
https://orcid.org/0000-0003-1169-0978
Universitat Politècnica de València
Pattern Recognition and Human Language Technologies (*PRHLT*) *Research Center*
Camí de Vera, s/n
46022 Valencia, Spain
berta.chulvi@upv.es

Ángel Panizo-Lledot
https://orcid.org/0000-0002-2195-3527
Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos
Alan Turing, s/n
28031 Madrid, Spain
angel.panizo@upm.es

Javier Huertas-Tato
https://orcid.org/0000-0003-4127-5505
Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingeniería de Sistemas Informáticos
Alan Turing, s/n
28031 Madrid, Spain
javier.huertas.tato@upm.es

Blanca Calvo-Figueras
https://orcid.org/0000-0001-6939-3576
Barcelona Supercomputing Center (*BSC*)
Language Technologies Unit
Plaça Eusebi Güell, 1-3
08034 Barcelona, Spain
blanca.calvo@bsc.es

M. José Rementeria
https://orcid.org/0000-0002-3140-1160
Barcelona Supercomputing Center (*BSC*)
Social and Media Impact Evaluation
Plaça Eusebi Güell, 1-3
08034 Barcelona, Spain
maria.rementeria@bsc.es

Juan Gómez-Romero
https://orcid.org/0000-0003-0439-3692

Universidad de Granada
Decsai
Citic-UGR
Periodista Rafael Gómez Montero, 2
18014 Granada, Spain
jgomez@decsai.ugr.es

Abstract

Internet and social media have revolutionised the way news is distributed and consumed. However, the constant flow of massive amounts of content has made it difficult to discern between truth and falsehood, especially in online platforms plagued with malicious actors who create and spread harmful stories. Debunking disinformation is costly, which has put artificial intelligence (AI) and, more specifically, machine learning (ML) in the spotlight as a solution to this problem. This work revises recent literature on AI and ML techniques to combat disinformation, ranging from automatic classification to feature extraction, as well as their role in creating realistic synthetic content. We conclude that ML advances have been mainly focused on automatic classification and scarcely adopted outside research labs due to their dependence on limited-scope datasets. Therefore, research efforts should be redirected towards developing AI-based systems that are reliable and trustworthy in supporting humans in early disinformation detection instead of fully automated solutions.

Keywords

Journalism; Disinformation; Computing; Artificial intelligence; AI; Machine learning; Fact-checking; Datasets; Natural language processing; NLP; Social network analysis; Deepfakes; Large language models.

Funding

This work was funded by the *European Commission*, project *Iberifier* (*Iberian Digital Media Research and Fact-Checking Hub*), under the call CEF-TC-2020–2 (*European Digital Media Observatory*), grant number 2020-EU-IA-0252.

1. Introduction

Amidst the prevailing post-truth era, people are overwhelmed with an enormous and uninterrupted flow of information, making it difficult to discern reliable material from content that seeks to mislead, whether intentionally (i.e., disinformation) or unintentionally (i.e., misinformation) (**Wardle**; **Derakhshan**, 2017). As a result, disinformation poses a significant and wide-ranging threat that can potentially transform any society's political, economic, and cultural fabric, thus eroding the fundamental principles of democratic nations.

While domain experts and fact-checkers may find it relatively easy to disprove hoaxes, more resources are necessary to drive and speed up their work and empower non-specialised citizens and organisations. Hence, the interest in developing technological tools for automatic information verification has grown, particularly in the ever-changing social media environment. Machine learning (ML), a subfield of Artificial Intelligence (AI), has significantly contributed to combating disinformation in recent years. Essentially, ML algorithms can be trained with data to automatically detect patterns indicative of disinformation and then apply these patterns to discern the likely truth or falsehood of unseen content. Deep Learning (DL), a subset of ML algorithms based on neural networks, has proved very useful in multiple domains (**LeCun**; **Bengio**; **Hinton**, 2015) and currently completely dominates the AI landscape. ML is also the predominant approach to fight disinformation (**Xu**; **Sheng**; **Wang**, 2023), but at the same time it can be used to generate synthetic content, increasing the impact of disinformation(**Masood** *et al.*, 2022).

ML is a very active, technical, and complex subject, making it difficult for non-specialists to understand and incorporate solutions arising in this field. At the same time, ML researchers must be aware of the multiple facets of a social problem like disinformation. Consequently, the research objective of this paper is to provide a brief and multidisciplinary guide to navigate the recent literature on AI to combat disinformation focusing on ML. This paper discusses the effectiveness of AI and ML techniques in detecting and counter-fighting disinformation and identifies the challenges and limitations of current approaches. We also suggest research directions for developing trustworthy AI-based systems that can assist humans in the early detection of disinformation.

Disinformation in social and digital media has prevalently spread through text. Therefore, when training ML algorithms, the primary characteristics considered are related to the syntax and content of the messages, including aspects such as syntactic, lexical, stylistic, and semantic features, which fall into the field of natural language processing (NLP). Furthermore, social network analysis (SNA) has researched the topology of disinformation networks. By

> Machine learning (ML), a subfield of Artificial Intelligence (AI), has significantly contributed to combating disinformation in recent years

analysing the network structure and identifying communities, it is possible to identify groups of users who are likely to generate and disseminate harmful content, whether in a coordinated or uncoordinated way. Accordingly, we centre our work on NLP and SNA as the areas of AI more often related to disinformation analysis.

Automated disinformation analysis has been addressed from multiple perspectives. Here we propose an organisation into three overlapping approaches:

- disinformation identification by automated classification;
- feature extraction to characterise disinformation; and
- providing support to fact-checking tasks.

This organization is consistent with the approaches of the revised research works and reflects the historical development of the area:

- Disinformation classification. Automated classification is the most straightforward way of disinformation analysis –given a labelled dataset, we can train an ML classification model to distinguish legit contents. However, this methodology has the drawback that trained models on one domain are hardly extensive to others.
- Feature-based disinformation identification. Feature extraction, in turn, focuses on finding characteristics of disinformation that can be used manually or automatically to detect content and communities of interest afterwards.
- Hybrid-based fact-checking. Detecting misleading content by specialized journalists has proved very effective for disinformation analysis but also bottleneck in the process. This limitation has led to the emergence of a third type of approach known as semi-automated fact-checking.

The remainder of this manuscript is accordingly divided into three parts. The first describes AI techniques and methods used to detect disinformative content. The second focuses on the AI methods proposed in the literature to combat disinformation, including the features used to train these models and how fact-checkers can take advantage of these technological advances. The third one describes the increasing use of AI to generate disinformative content automatically. Finally, we end the paper with a summary of the main findings and the most promising research lines for future work.

2. Background

ML is a powerful tool within AI that can help to address the growing problem of disinformation by automating the detection and analysis of untrustworthy content. This section provides the reader with a background on ML and an overview of the fundamentals of Natural Language Processing and Social Network Analysis. Readers familiar with AI and ML can skip this section; otherwise, more information can be found in the classical books by **Russell** and **Norvig** (2020) and **Bishop** (2006).

2.1. Machine Learning

Machine Learning is a field of AI that encompasses a range of methods, techniques, and tools for building intelligent systems by exploiting large volumes of data related to a specific problem. Specifically, ML falls under the pattern recognition paradigm, i.e., it identifies repeating characteristics in a data sample using statistical and computational processes. These patterns serve two primary functions: making predictions about future events (predictive analysis) and uncovering insights from the data (descriptive analysis). Depending on the learning mode and the process of obtaining patterns, there are three main families of ML techniques: Supervised, Unsupervised, and Reinforcement Learning. Based on artificial neural networks, Deep Learning mainly falls into Supervised Learning, but it can also be applied in Unsupervised and Reinforcement Learning setups. This subsection focuses on Supervised and Unsupervised techniques (including Deep Learning), the most representative ML techniques to combat disinformation.

Supervised Learning seeks to develop models from labelled training data that allows predicting the labels of unseen or future data. Supervised Learning can be classified into two basic categories, depending on the nature of the target variable: classification and regression. In classification, the target variable has a limited number of discrete values. Archetypical methods within this category are Decision Trees, Logistic Regression, Support Vector Machines, and the K-Nearest-Neighbour algorithm. In regression, the target variable is a real number. Some regression algorithms are Linear Regression, Polynomial Regression, Regression Splines, and Regression Trees. Supervised Learning methods are often combined to increase accuracy, yielding ensemble models such as Bagging, Boosting, and Random Forest.

Unsupervised Learning refers to techniques that deal with unlabelled or unstructured data. The most prevalent technique is clustering, utilised to identify hidden groups within a dataset for descriptive analysis. We have partitional clustering, where clusters are disjoint and typically encompass the entire item set (e.g., the *Dbscan* and k-means algorithms), and hierarchical clustering, where groups are organised into a hierarchy. Another notable technique within Unsupervised Learning is association rules, which aim to discover dependencies between a set of items in a database.

The current dominant trend in ML is Deep Learning (**Goodfellow**; **Bengio**; **Courville**, 2016), which was first applied in Supervised Learning setups and has since been extended to other paradigms. Deep Learning enhances traditional neural networks, which are computational models that, inspired by the form of neuronal synapses, can learn intricate decision boundaries from data. Because deep neural networks possess more intermediate layers and neurons in each layer, they can capture complex relationships in large datasets. Different types of algorithms fall within Deep Learning, such as

- Convolutional Neural Networks (CNNs), which are specialised neural networks that process data with a regular structure, like images;
- Recurrent Neural Networks (RNNs), which process sequential data allowing feedback loops in the networks and work well with time series; and
- Transformers, which learn to identify relevant sections of sequences by applying attention models and are very useful with textual data.

2.2. Natural Language Processing

Natural Language Processing (NLP) involves using computational linguistics techniques to analyse text in a specific language, whether written or spoken (**Manning**; **Schütze**, 1999). Before developing a ML model for natural language processing (e.g., a language model), it is crucial to tackling three critical challenges: text preprocessing, feature extraction, and representation.

1) Text preprocessing involves cleaning the text and eliminating unimportant elements so that only useful information remains. The fundamental steps of text preprocessing are tokenisation (the division of the raw text into units), stopword removal (elimination of common words not significant for the analysis) and stemming (heuristic-type rules for cutting off the ends of words or affix removal) or lemmatisation (transformation of words into their base form or lemma).

2) Feature extraction involves identifying and selecting basic features from raw text data suitable to the task. Some of the most widely used techniques for feature extraction are Part-Of-Speech tagging (POS) to identify lexical categories, Named-Entity Recognition (NER) for identifying entities within the text, and bag-of-words to represent linguistic units based on their frequency of occurrence.

Another more advanced feature extraction technique is Sentiment Analysis (SA), also called Opinion Mining, which aims to automatically grasp a text's sentiments, opinions, emotions, or attitudes (**Serrano-Guerrero** *et al.*, 2015). It can also include eliciting the author's psychological traits through specific-purpose annotated lexicons (**John**; **Srivastava**, 1999; **Pennebaker** *et al.*, 2015).

3) Representation involves creating a numerical encoding of the text so that other ML algorithms can use it. Many techniques exist, but word embeddings are the most widely used today. They are representations of text units in the form of numerical vectors that capture their semantics. *Word2Vec* (**Mikolov** *et al.*, 2013) and *GloVe* (**Pennington**; **Socher**; **Manning**, 2014) are the most used techniques for obtaining embeddings. There are also available public embeddings for common terms precalculated from massive text sources, like *Wikipedia*, that can be reused in other applications. Once a document is represented as numbers, ML techniques (and particularly Deep Learning methods) can be applied to solve a downstream task (e.g., text classification or text prediction).

In this regard, Transformer networks with attention mechanisms aim to overcome the limitations of previous methods by learning to hold on to essential parts of the input text (**Vaswani** *et al.*, 2017). Particularly, Large Language Models (LLMs) are neural network base systems specialised in predicting the next word in a sequence that can be used for text generation and translation between sequences. A particularly noteworthy LLM is the Generative Pre-trained Transformer (GPT) (**Brown** *et al.*, 2020). Its current incarnation *GPT-3* and *GPT-4* can generate natural language and perform a wide range of NLP tasks, such as text generation, machine translation and question-answering (**Zhu**; **Luo**, 2022). More recently, a variant of *GPT-3* named *ChatGPT* has been successfully trained through human interaction to engage in realistic conversations (**Megahed** *et al.*, 2023).

2.3. Social Network Analysis (SNA)

Social Network Analysis is the computational field that explores social entities' relationships, patterns, and structures to understand the system, position, and linkage between these actors (**Barabási**, 2016). SNA uses mathematical and computational methods to analyse data from social media through two different approaches (**Aggarwal**, 2011; **Camacho** *et al.*, 2020):

- structural analysis (topology of the network, communities, and important nodes); and
- content-based analysis (information about social media users, shared content).

Structural analysis focuses on studying the topology of a network by applying graph theory. Often-used structural metrics include local measures like centrality, degree, closeness or betweenness –used for identifying the importance of certain nodes (users) within the network, and global measures such as density, diameter, radius, or transitivity –used to study the global structure of the network. An essential problem in SNA is community detection, which aims to identify sets of more tightly connected nodes (**Bedi**; **Sharma**, 2016). The task of community detection is closely related to the clustering problem, so most techniques belong to this broad family of algorithms (**Fortunato**, 2010). Other approaches are based on the maximisation of modularity, a measure that balances the number of internal and external connections of a community. Some algorithms based on modularity are Newman's greedy method (**Newman**, 2004) and the Blondel method (**Blondel** *et al.*, 2008).

Content-based analysis examines both the content and the connections between nodes, for example, by incorporating text to provide additional context to the network (**Cambria**; **Wang**; **White**, 2014). Content analysis is commonly applied in the following ways:

- user profiling, which gathers extra information about the human actors –e.g. behaviour or physical features– in a network (**Harrigan** *et al.*, 2021);
- topic extraction, which identifies the main themes of discussion among a group of nodes (**Yin** *et al.*, 2012), or the interests of users through their social connections (**Wang** *et al.*, 2013);
- sentiment analysis, which examines the tone of the messages exchanges among the nodes (**Camacho** *et al.*, 2020).

3. Disinformation classification with Machine Learning

Supervised Learning is the most widely employed approach for the automatic identification of disinformation. Thereby, the identification of disinformation is usually modelled as a binary classification problem. Given a set of representative features of an information item I, the task is to predict whether I is truthful or not, i.e.:

$$f(I) = \begin{cases} 1, \text{if } I \text{ is a disinformation item} \\ 0, \text{if } I \text{ is not disinformative} \end{cases}$$

where f is the function we want to learn from the available data. The combination of the features to obtain f can be done manually or automatically. In the first case, Multi-Criteria Decision Making (MCDM) has been applied to define criteria and probability weights to calculate an information credibility score and rank the candidate solutions (**Pasi**; **De-Grandis**; **Viviani**, 2020). In the second case, DL has been applied to learn the features and the combination weights (**Amador**; **Molina-Solana**; **Gómez-Romero**, 2019; **Molina-Solana**; **Amador**; **Gómez-Romero**, 2018).

> The performance of Machine Learning depends directly on the quality of the data

Nevertheless, disinformation flows in shades of grey, not black and white, rendering a binary classification insufficient. In the literature, we can find more precise definitions of labels to capture the more subtle nuances of disinformation. For example, **Wang** (2017) proposed a manually labelled dataset with six fine-grained labels where the degree of truthfulness (pants-fire, false, barely true, half-true, mostly true, and true) of thousands of statements was evaluated. **Nakamura**, **Levy** and **Wang** (2020) used a labelling hierarchy of two, three, and six categories for each sample of their multimodal dataset enabling the implementation of classification models at different levels of granularity.

The performance of Supervised Learning depends directly on the quality of the labelled data, which usually represents situations, making it difficult to extend the models to other similar domains. This limitation is even more noticeable when applied to the automatic detection of disinformation since it is challenging to build datasets with enough quality to cover the nuances of disinformation in heterogeneous contexts (**Shu** *et al.*, 2017). Dataset construction involves

- data extraction, either through APIs provided by platform owners or web scraping methods, and
- annotation, which is a manual time-consuming and error-prone task with little automatic support (**Simko** *et al.*, 2021).

Annex includes datasets used in the literature for testing ML disinformation classification models.

As reported several times (**Guo** *et al.*, 2020; **Meel**; **Vishwakarma**, 2020; **Zhang**; **Ghorbani**, 2020), studies that work directly on automatic disinformation detection with Unsupervised Learning are scarce. Some works formulate automatic identification of disinformation as an anomaly detection problem on social networks, employing an autoencoder as an Unsupervised Learning method (**Li** *et al.*, 2021), another uses Bayesian statistics to compute the veracity of news and the credibility of their authors (**Yang** *et al.*, 2019). Nevertheless, most studies use Unsupervised Learning in a complementary way to Supervised Learning; that is, they use a Semi-supervised approach (**De-Souza** *et al.*, 2022; **Dong**; **Victor**; **Qian**, 2020; **Li**; **Lu** *et al.*, 2022; **Meel**; **Vishwakarma**, 2021; **Paka** *et al.*, 2021).

4. Feature-based automated disinformation detection

As explained, methods for disinformation detection need relevant features representative of the news items. Classically, they have been classified into content-based and context-based features (**Bondielli**; **Marcelloni**, 2019).

- Content-based features are relevant attributes extracted directly from the data item, usually a text stating or supporting the potential hoax and often associated with several images or videos that reinforce it.
- Context-based features refer to data or metadata surrounding the piece of information. This section focuses on various features that can be extracted and used to detect false information.

4.1. Natural language processing for stylistic characterisation of messages

Content-based methods use the linguistic features of false information, including syntactic and semantic characteristics (**Zhou** *et al.*, 2020) that can be obtained by applying NLP techniques (**Ruffo** *et al.*, 2023). Among syntactic features, we can find POS tags and relevant groups of words (bigrams, trigrams, or n-grams). Semantic features can be obtained through sentiment analysis, opinion mining, topic detection, or encodings with word embeddings.

A specific kind of linguistic feature is style-based features. The rationale behind methods based on them is that ML can capture the distinctive style that malicious actors use to increase the diffusion and acceptance of their content (**Zhou**; **Zafarani**, 2020). The style of news text has been formalised and measured in terms of the frequency of morphological

patterns (**Castelo** *et al.*, 2019; **Vogel**; **Meghana**, 2020), the presence of structural elements (**Bonet-Jover** *et al.*, 2021), the lexical variety and the use of punctuation symbols (**Azevedo** *et al.*, 2021), the complexity and level of readability of the text (**Castelo** *et al.*, 2019) and the emotional tone (**Giachanou**; **Rosso**; **Crestani**, 2019).

> The features employed for disinformation classification can be categorized into two groups: content-based features and context-based features

Regarding morphological patterns, in an early study, **Afroz, Brennan** and **Greenstadt**(2012) were able to identify false information by analysing the number of syllables and words, vocabulary, grammatical complexity, and POS tags. Misleading content spreaders were also found to use more informal language (**Giachanou** *et al.*, 2022), e.g., certain patterns in the use of personal pronouns and swear words (**Rashkin** *et al.*, 2017). Regarding the emotional tone of the discourse, **Del-Vicario** *et al.* (2016) showed that the emotional state of social media users is linked to their level of engagement in the community –more activity leads to more negative emotions and vice versa. Accordingly, the use of polarised language patterns is often seen as a sign of message engineering to increase impact by provoking negative emotions in the receiver, such as anger, disgust, or fear (**Giachanou**; **Rosso**; **Crestani**, 2021), and therefore, an indicator of low credibility (**Ghanem** *et al.*, 2021; **Stella**; **Ferrara**; **De-Domenico**, 2018).

Conversely, disinformers can learn style-based features to replicate the writing styles of trustworthy information sources and disguise their actions. This is particularly problematic if language models are used to generate disinformation, which is currently a trend and a challenge. For example, **Schuster** *et al.* (2020) showed that NLP models for disinformation identification based on stylistic features work well with human writing. Still, they tend to fail when confronted with synthetic text created by language models trained to replicate trusted media.

4.2. Contextual aspects of disinformation in social networks

Contextual features are extracted by considering the relevant data related to an information item, including metadata or other external elements. This information is primarily available in social networks, where context can be connected to the users, their posted messages, or the network (**Guo** *et al.*, 2020).

4.2.1. Features based on the context of the users

User-based features include the number of posts, number of followers, demographics, whether the account is verified, or the age of the account on the platform. A usual metric built from such profile data is user credibility, which can indicate the likelihood of sharing false information (**Shu**; **Wang**; **Liu**, 2019). Credibility can be obtained from network metadata to analyse whether there is a correlation between a user profile and the publication of false information (**Shu** *et al.*, 2019). Furthermore, user engagement (likes, retweets, and replies) with tweets written by verified users can also be used to assess credibility (**Yang** *et al*., 2019).

A very interesting type of social network user is bots. Bots are computer programs that carry out autonomous actions, including automatically generating false information and amplifying disinformation during the initial dissemination stages (**Shao**; **Ciampaglia** *et al.*, 2018). Bots tend to have particular profiles on social networks, e.g., they are usually recent accounts (**Davis** *et al.*, 2016) with lengthy usernames using weird characters (**Oehmichen** *et al.*, 2019). Their behaviour is also different from humans' (**Ruffo** et al., 2023); e.g., they retweet more, get fewer retweets, receive fewer replies and mentions, and publish fewer original tweets (**Ferrara** *et al.*, 2016). All these features can be obtained from the public profiles and the graph of retweets for automatic bot identification alone (**Des-Mesnards** *et al.*, 2022) or combined with message data (**Kudugunta**; **Ferrara**, 2018).

Disinformation is closely related to the user's personality and mental processes. Given that psychological characteristics regulate behaviour and interaction in the physical world, it is logical to assume that they also impact virtual communities. Psychological traits can influence how individuals interpret and engage with information, increasing the likelihood of spreading false information and toxic narratives. For example, inherently human cognitive biases such as limited reality perception and confirmation bias can increase the likelihood of perceiving fake news as real and thus encourage its dissemination (**Shu** *et al.*, 2017). Unlike disseminators of accurate information, disinformers have been found to be extroverted, less neurotic and present more stress in their tweets (**Shrestha**; **Spezzano**, 2022). In contrast, **Srinivas**, **Das** and **Pulabaigari** (2022) suggest that users who spread false political information are neurotic, conservative and have psychopathic traits. The difference in the conclusions of these works is mainly due to the way of detecting and measuring these psychological traits.

4.2.2. Features based on the context of the messages

Contextual user and message-based features are often not clearly distinguished (**Guo** *et al.*, 2020) and even merged (**Yang** *et al.*, 2019). Still, for clarity, we consider the context of the posted messages separately, which are different, more dynamic, and specific than the users' features (**Tacchini** *et al.*, 2017). Thus, metadata about posts in social networks has been mainly used to increase the effectiveness of another principal feature (**Della-Vedova** *et al.*, 2018). Likewise, multimedia resources associated with messages have been used to complement ML models, yielding multimodal disinformation analysis (**Hangloo**; **Arora**, 2022).

Multimodal analysis has been focused to date on images and addressed in three main forms: forensic –evaluates whether an image has been subjected to modification or manipulation (**Qi** *et al.*, 2019)–, contextual –the image and the text are consistent (**Kang**; **Hwang**; **Yu**, 2020; **Xiong** *et al.*, 2023)–, and hybrid –the image is processed to extract additional information to be used in (**Giachanou**; **Zhang**; **Rosso**, 2020; **Jing** *et al.*, 2023; **Khattar** *et al.*, 2019; **Li**; **Yao** *et al.*, 2022; **Singh** *et al.*, 2023; **Wang** *et al.*, 2018). For example, **Zhang**, **Giachanou** y **Rosso** (2022) combined textual, visual, and contextual information to build the "scene" depicted in the post, obtaining statistically significant differences in the appearance of specific places, weather, and seasons in false and truthful content.

4.2.3. Features based on the network structure

Network-based features refer to the static structure of the social network, such as central nodes and communities based on users' connections, and the more dynamic propagation of (dis)information, including critical actors, dissemination paths, and infiltration from one community to another (**Bondielli**; **Marcelloni**, 2019; **Zhou**; **Zafarani**, 2020).

Most works in the literature focus on detecting false information by modelling the information dissemination network, assuming that true and false information have different propagation patterns (**De-Souza** *et al.*, 2022; **Liu**; **Wu**, 2018; **Liu**; **Xu**, 2016; **Song** *et al.*, 2022). Other works have combined the analysis of propagation paths with spreaders' characteristics for disinformation classification (**Grinberg** *et al.*, 2019; **Shao**; **Ciampaglia** *et al.*, 2018; **Shao**; **Hui**; *et al.*, 2018). This approach is highly effective for stopping the propagation of disinformation, as it prioritises identifying (and removing) disinformative over the more costly analysis of individual publications. Specifically, the networking characteristics of users involved in disseminating false information have been investigated through initiatives such as the PAN challenges (**Buda**; **Bolonyai**, 2020; **Vogel**; **Meghana**, 2020). In addition, modern ML techniques have been recently applied to this topic, e.g., **Rath**, **Salecha** y **Srivastava** (2022) proposed a graph neural network model to identify nodes prone to disseminate false information using network topology and historical user activity data.

5. AI-supported fact-checking

Fact-checking is journalism focused on checking public assertions (**Graves**; **Nyhan**; **Reifler**, 2016). While verifying information is a foundational part of journalism, fact-checking emphasises the relevance of the checking process and the development of methods and tools to do so effectively and transparently. The first proposals to automate online fact-checking appeared more than 15 years ago (**Graves**, 2018), already highlighting that full automation is practically impossible because of the critical judgment, sensitivity, and experience required to make a decision that is not binary (**Arnold**, 2020). The fact-checking community acknowledges that the rapid dissemination of false information presents scalability issues –i.e., spreading a lie is way faster than debunking it (**Vosoughi**; **Roy**; **Aral**, 2018)– but this should not undermine the rigour of the fact-checking process.

Accordingly, the approaches in the literature tend to AI-supported fact-checking rather than automated fact-checking, which is often labelled as human-in-the-loop systems (**La-Barbera**; **Roitero**; **Mizzaro**, 2022; **Shabani** *et al.*, 2021; **Yang** *et al.*, 2021). AI can support fact-checking at different stages of the verification workflow (**Guo**; **Schlichtkrull**; **Vlachos** 2022; **Nakov**; **Corney** *et al.*, 2021):

(1) monitoring, recognition, and prioritisation of content susceptible to verification;
(2) evaluating whether claims are verifiable or not and topic prioritisation;
(3) searching for previous verifications that apply to the same case;
(4) retrieval of evidence for further investigation;
(5) semi-automated classification in categories (hoax, misleading content, false context, etc.);
(6) dissemination of the verifications; and
(7) speeding-up writing and documenting the fact-checks.

Proposals in the literature have primarily focused on stages 1-4. For stage 5, the contributions described in Section 3 could be applied, although they show limitations in their applicability to multiple domains, as already described.

Various methods have been proposed for the check-worthiness of claims (stages 1 and 2), either based on the ranking of claims by score prediction (**Kartal**; **Kutlu**, 2023; **Nakov**; **Da-San-Martino** *et al.*, 2021) or the classification of the claim using specific annotations (**Konstantinovskiy** *et al.*, 2021). Since automated systems can introduce biases in claim selection, research has pivoted towards tools like news alerts, speech recognition, and translation models to filter claims more effectively (**Rashkin** *et al.*, 2017).

Detecting previously fact-checked claims, including those verified in other languages or countries, has been addressed with NLP and information retrieval techniques (stages 3 and 4). In the first case, semantic textual similarity has been applied to match new claims with already-verified ones in English (**Thorne**; **Vlachos**, 2018) and Spanish (**Martín** *et al.*, 2022). In the second case, software tools with different levels of intelligence have been developed for evidence retrieval, including structured data extraction, speech recognition, reverse image search, video forensics, or natural language search (**Das** *et al.*, 2023).

> "The current trend leans towards fact-checking assisted by Artificial Intelligence rather than relying solely on fully automated fact-checking"

One remarkable tool covering different stages is *InVid*, a free platform that hosts tools to detect, authenticate and check the reliability and authenticity of images and videos.
https://www.invid-project.eu

The *vera.ai* project is expected to continue and expand AI-supported verification tools and services in Europe.
https://www.veraai.eu

6. The challenge of the automatic generation of disinformation

Large Language Models (LLMs), introduced in Section 2.2, are one of the most challenging technologies for the massive generation of textual disinformative content. For example, *GPT-3* and *ChatGPT* can produce synthetic text that can be exploited to spread disinformation in many ways (**Solaiman** *et al.*, 2019):

- to camouflage false content under the guise of real information;
- to create bots and web pages amplifying a disinformative discourse;
- to elude stylistic checkers, etc.

Additionally, since there is no control over the sources used to train LLMs, much of the content they learn and produce is false and biased (**Marcus**, 2022). Therefore, it is crucial to develop effective methods for detecting and mitigating the impact of LLM-generated disinformation; unfortunately, attempts to date are still ineffective (**Mitchell** *et al.*, 2023).

Disinformation is not limited to text format; images, videos, and audio can also be spread, often even more harmful than text. The term deepfake denotes very realistic content automatically generated or altered with DL techniques like Generative Adversarial Networks (GANs) (**Goodfellow** *et al.*, 2014), e.g., fake avatar generation, face and speech manipulation, and person and background substitution. Not surprisingly, deepfakes have been applied for disinformative purposes like damaging an individual's reputation or manipulating elections (**Greengard**, 2019; **Masood** *et al.*, 2022). Therefore, fighting deepfakes entails researching how they can be generated and detected (**Dagar**; **Vishwakarma**, 2022; **Mirsky**; **Lee**, 2022; **Saif**; **Tehseen**, 2022).

Face manipulation in images and videos, either partial or total, has been one of the most active areas of research to date and one that poses a significant challenge to fighting disinformation. Full face generation refers to the creation of a completely fake face (**Serengil**; **Ozpinar**, 2021) using architectures such as *ProGAN* (**Karras** *et al.*, 2018) or *StyleGAN* (**Karras**; **Laine**; **Aila**, 2019). Partial manipulation, in turn, refers to modifications like face swapping, attribute manipulation (hair, skin tone, eyes, etc.), face re-enactment, and lip-syncing (**Tolosana** *et al.*, 2020). Conversely, there is a wide range of ML-based techniques for detecting deepfakes. Particularly, convolutional neural networks (CNNs) with attention mechanisms have been recently used (**Dagar**; **Vishwakarma**, 2022; **Rana** *et al.*, 2022; **Tolosana** *et al.*, 2020), but their effectiveness lags behind the advances in deepfake generation and the possibility of manually refine the deepfakes in post-production.

7. Conclusions and future work

The speed and the amount of available data make it challenging to distinguish trustworthy information from disinformative content that is often disguised as legit and appeals to emotions and beliefs. Computational technologies have arisen as suitable tools to address disinformation but also have exceptional capabilities to exacerbate the problem through invention and falsification. In this manuscript, we have described the current trends in AI and ML applied to disinformation detection and characterisation, as well as the challenges posed by synthetic text and media generation.

Most of the reviewed proposals perform an a posteriori analysis of disinformative content once it has become impactful, focusing on different features that can be used in automatic classification. While the approaches assume that solutions in specific problems and domains can be extended to others, they strongly depend on the datasets used and the processes to create them. Therefore, there is a need for new, high-quality, and unbiased datasets, particularly in languages other than English. Furthermore, more efforts are required to transfer and evaluate trained models from one domain to another. Similarly, AI-based disinformation analysis tools are not widely available or lack the maturity that non-technological users need.

Early detection of disinformation is crucial to limit the impact of a phenomenon that otherwise is impossible to deter. Therefore, we identify two future research directions for fighting disinformation with AI and ML. The first one is the study of patterns of creation and propagation, including paths and ecosystems, to understand better and anticipate the spread of harmful propaganda and conspiracy theories. The second one is the application of intelligent technologies to amplify the scope of fact-checks and media literacy, similarly as disinformers engineer their messages to reach wider audiences.

These initiatives will require creating explainable AI methods to provide results and justify them and facilitating the interplay between technological tools and practitioners with deep domain knowledge, including fact-checkers, experts, and decision-makers. By addressing these challenges, we will progress towards AI-based systems that can detect and combat disinformation more effectively, ultimately contributing to a better-informed society.

> Early detection of disinformation is crucial to limit the impact of a phenomenon that otherwise is impossible to deter

8. References

Afroz, Sadia; **Brennan, Michael**; **Greenstadt, Rachel** (2012). "Detecting hoaxes, frauds, and deception in writing style online". In: *IEEE symposium on security and privacy*, pp. 461-475.
https://doi.org/10.1109/SP.2012.34

Aggarwal, Charu C. (2011). "An introduction to social network data analytics". In: Aggarwal, Charu C. (ed.). *Social network data analytics.* Springer.
https://doi.org/10.1007/978-1-4419-8462-3

Amador, Julio; **Molina-Solana, Miguel**; **Gómez-Romero, Juan** (2019). "Towards easy-to-implement misinformation automatic detection for online social media". In: *Proceedings of the conference for truth and trust online 2019*.
https://doi.org/10.36370/tto.2019.4

Arnold, Phoebe (2020). "The challenges of online fact checking". *Full fact*, 17 December.
https://fullfact.org/blog/2020/dec/the-challenges-of-online-fact-checking-how-technology-can-and-cant-help

Azevedo, Lucas; **D'Aquin, Mathieu**; **Davis, Brian**; **Zarrouk, Manel** (2021). "LUX (linguistic aspects under examination): discourse analysis for automatic fake news classification". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 41-56.
https://doi.org/10.18653/v1/2021.findings-acl.4

Barabási, Albert-László (2016). *Network science*. Cambridge University Press. ISBN: 978 1 107 07626 6
http://networksciencebook.com

Bedi, Punam; **Sharma, Chhavi** (2016). "Community detection in social networks". *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, v. 6, n. 3, pp. 115-135.
https://doi.org/10.1002/widm.1178

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Springer. ISBN: 978 0 387 31073 2
https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Blondel, Vincent D.; **Guillaume, Jean-Loup**; **Lambiotte, Renaud**; **Lefebvre, Etienne** (2008). "Fast unfolding of communities in large networks". *Journal of statistical mechanics: theory and experiment*, n. 10, pp. P10008.
https://doi.org/10.1088/1742-5468/2008/10/P10008

Bondielli, Alessandro; **Marcelloni, Francesco** (2019). "A survey on fake news and rumour detection techniques". *Information sciences*, v. 497, pp. 38-55.
https://doi.org/10.1016/j.ins.2019.05.035

Bonet-Jover, Alba; **Piad-Morffis, Alejandro**; **Saquete, Estela**; **Martínez-Barco, Patricio**; **García-Cumbreras, Miguel-Ángel** (2021). "Exploiting discourse structure of traditional digital media to enhance automatic fake news detection". *Expert systems with applications*, v. 169, 114340.
https://doi.org/10.1016/j.eswa.2020.114340

Brown, Tom B.; **Mann, Benjamin**; **Ryder, Nick**; **Subbiah, Melanie**; **Kaplan, Jared**; **Dhariwal, Prafulla**; **Neelakantan, Arvind**; **Shyam, Pranav**; **Sastry, Girish**; **Askell, Amanda**; **Agarwal, Sandhini**; **Herbert-Voss, Ariel**; **Krueger, Gretchen**; **Henighan, Tom**; **Child, Rewon**; **Ramesh, Aditya**; **Ziegler, Daniel M.**; **Wu, Jeffrey**; **Winter, Clemens**; **Hesse, Christopher**; **Chen, Mark**; **Sigler, Eric**; **Litwin, Mateusz**; **Gray, Scott**; **Chess, Benjamin**; **Clark, Jack**; **Berner, Christopher**; **McCandlish, Sam**; **Radford, Alec**; **Sutskever, Ilya**; **Amodei, Dario** (2020). "Language models are few-shot learners". *Advances in neural information processing systems*, v. 33, pp. 1877-1901.
https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Buda, Jakab; **Bolonyai, Flora** (2020). "An ensemble model using n-grams and statistical features to identify fake news spreaders on *Twitter*". In: *Working notes of CLEF 2020 - Conference and labs of the evaluation forum*, v. 2696.
https://ceur-ws.org/Vol-2696/paper_189.pdf

Camacho, David; **Panizo-Lledot, Ángel**; **Bello-Orgaz, Gema**; **González-Pardo, Antonio**; **Cambria, Erik** (2020). "The four dimensions of social network analysis: an overview of research methods, applications, and software tools". *Information fusion*, v. 63, pp. 88-120.
https://doi.org/10.1016/j.inffus.2020.05.009

Cambria, Erik; **Wang, Haixun**; **White, Bebo** (2014). "Guest editorial: big social data analysis". *Knowledge-based systems*, v. 69.
https://doi.org/10.1016/j.knosys.2014.07.002

Castelo, Sonia; **Almeida, Thais**; **Elghafari, Anas**; **Santos, Aécio**; **Pham, Kien**; **Nakamura, Eduardo**; **Freire, Juliana** (2019). "A topic-agnostic approach for identifying fake news pages". In: *Companion proceedings of the 2019 World Wide Web conference*, pp. 975-980.
https://doi.org/10.1145/3308560.3316739

Dagar, Deepak; **Vishwakarma, Dinesh K.** (2022). "A literature review and perspectives in deepfakes: generation, detection, and applications". *International journal of multimedia information retrieval*, v. 11, n. 3, pp. 219-289. *https://doi.org/10.1007/s13735-022-00241-w*

Das, Anubrata; **Liu, Houjiang**; **Kovatchev, Venelin**; **Lease, Matthew** (2023). "The state of human-centered NLP technology for fact-checking". *Information processing & management*, v. 60, n. 2, 103219. *https://doi.org/10.1016/j.ipm.2022.103219*

Davis, Clayton-Allen; **Varol, Onur**; **Ferrara, Emilio**; **Flammini, Alessandro**; **Menczer, Filippo** (2016). "BotOrNot: a system to evaluate social bots". In: *Proceedings of the 25th International conference companion on World Wide Web*, pp. 273-274. *https://doi.org/10.1145/2872518.2889302*

Della-Vedova, Marco L.; **Tacchini, Eugenio**; **Moret, Stefano**; **Ballarin, Gabriele**; **DiPierro, Massimo**; **De-Alfaro, Luca** (2018). "Automatic online fake news detection combining content and social signals". In: *22nd Conference of open innovations association (Fruct)*, pp. 272-279. *https://doi.org/10.23919/FRUCT.2018.8468301*

De-Souza, Mariana C.; **Nogueira, Bruno-Magalhães**; **Rossi, Rafael-Geraldeli**; **Marcacini, Ricardo-Marcondes**; **Dos-Santos, Brucce-Neves**; **Rezende, Solange-Oliveira** (2022). "A network-based positive and unlabeled learning approach for fake news detection". *Machine learning*, v. 111, n. 10, pp. 3549-3592. *https://doi.org/10.1007/s10994-021-06111-6*

Del-Vicario, Michela; **Vivaldo, Gianna**; **Bessi, Alessandro**; **Zollo, Fabiana**; **Scala, Antonio**; **Caldarelli, Guido**; **Quattrociocchi, Walter** (2016). "Echo chambers: emotional contagion and group polarization on facebook". *Scientific reports*, v. 6, 37825. *https://doi.org/10.1038/srep37825*

Des-Mesnards, Nicolas-Guenon; **Hunter, David-Scott**; **El-Hjouji, Zakaria**; **Zaman, Tauhid** (2022). "Detecting bots and assessing their impact in social networks". *Operations research*, v. 70, n. 1. *https://doi.org/10.1287/opre.2021.2118*

Dong, Xishuang; **Victor, Uboho**; **Qian, Lijun** (2020). "Two-path deep semisupervised learning for timely fake news detection". *IEEE transactions on computational social systems*, v. 7, n. 6, pp. 1386-1398. *https://doi.org/10.1109/TCSS.2020.3027639*

Ferrara, Emilio; **Varol, Onur**; **Davis, Clayton**; **Menczer, Filippo**; **Flammini, Alessandro** (2016). "The rise of social bots". *Communications of the ACM*, v. 59, n. 7, pp. 96-104. *https://doi.org/10.1145/2818717*

Fortunato, Santo (2010). "Community detection in graphs". *Physics reports*, v. 486, n. 3-5, pp. 75-174. *https://doi.org/10.1016/j.physrep.2009.11.002*

Ghanem, Bilal; **Ponzetto, Simone P.**; **Rosso, Paolo**; **Rangel, Francisco** (2021). "FakeFlow: fake news detection by modeling the flow of affective information". In: *Proceedings of the 16th Conference of the European chapter of the Association for Computational Linguistics*, pp. 679-689. *https://doi.org/10.18653/v1/2021.eacl-main.56*

Giachanou, Anastasia; **Ghanem, Bilal**; **Ríssola, Esteban A.**; **Rosso, Paolo**; **Crestani, Fabio**; **Oberski, Daniel** (2022). "The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers". *Data & knowledge engineering*, v. 138, 101960. *https://doi.org/10.1016/j.datak.2021.101960*

Giachanou, Anastasia; **Rosso, Paolo**; **Crestani, Fabio** (2019). "Leveraging emotional signals for credibility detection". In: *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pp. 877-880. *https://doi.org/10.1145/3331184.3331285*

Giachanou, Anastasia; **Rosso, Paolo**; **Crestani, Fabio** (2021). "The impact of emotional signals on credibility assessment". *Journal of the Association for Information Science and Technology*, v. 72, n. 9, pp. 1117-1132. *https://doi.org/10.1002/asi.24480*

Giachanou, Anastasia; **Zhang, Guobiao**; **Rosso, Paolo** (2020). "Multimodal multi-image fake news detection". In: *IEEE 7th International conference on data science and advanced analytics (DSAA)*, pp. 647-654. *https://doi.org/10.1109/DSAA49011.2020.00091*

Goodfellow, Ian; **Bengio, Yoshua**; **Courville, Aaron** (2016). *Deep learning*. MIT Press. ISBN: 978 0 262 035613

Goodfellow, Ian J.; **Pouget-Abadie, Jean**; **Mirza, Mehdi**; **Xu, Bing**; **Warde-Farley, David**; **Ozair, Sherjil**; **Courville, Aaron**; **Bengio, Yoshua** (2014). "Generative adversarial nets". *Advances in neural information processing systems*, v. 27. *https://papers.nips.cc/paper/5423-generative-adversarial-nets*

Graves, Lucas (2018). *Understanding the promise and limits of automated fact-checking*. Reuters Institute, University of Oxford.
https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf

Graves, Lucas; **Nyhan, Brendan**; **Reifler, Jason** (2016). "Understanding innovations in journalistic practice: a field experiment examining motivations for fact-checking". *Journal of communication*, v. 66, n. 1, pp. 102-138.
https://doi.org/10.1111/jcom.12198

Greengard, Samuel (2019). "Will deepfakes do deep damage?". *Communications of the ACM*, v. 63, n. 1, pp. 17-19.
https://doi.org/10.1145/3371409

Grinberg, Nir; **Joseph, Kenneth**; **Friedland, Lisa**; **Swire-Thompson, Briony**; **Lazer, David** (2019). "Fake news on *Twitter* during the 2016 U.S. presidential election". *Science*, v. 363, n. 6425, pp. 374-378.
https://doi.org/10.1126/science.aau2706

Guo, Bin; **Ding, Yasan**; **Yao, Lina**; **Liang, Yunji**; **Yu, Zhiwen** (2020). "The future of false information detection on social media: new perspectives and trends". *ACM computing surveys*, v. 53, n. 4.
https://doi.org/10.1145/3393880

Guo, Zhijiang; **Schlichtkrull, Michael**; **Vlachos, Andreas** (2022). "A survey on automated fact-checking". *Transactions of the Association for Computational Linguistics*, v. 10, pp. 178-206.
https://doi.org/10.1162/tacl_a_00454

Hangloo, Sakshini; **Arora, Bhavna** (2022). "Combating multimodal fake news on social media: methods, datasets, and future perspective". *Multimedia systems*, v. 28, n. 6, pp. 2391-2422.
https://doi.org/10.1007/s00530-022-00966-y

Harrigan, Paul; **Daly, Timothy M.**; **Coussement, Kristof**; **Lee, Julie A.**; **Soutar, Geoffrey N.**; **Evers, Uwana** (2021). "Identifying influencers on social media". *International journal of information management*, v. 56, 102246.
https://doi.org/10.1016/j.ijinfomgt.2020.102246

Jing, Jing; **Wu, Hongchen**; **Sun, Jie**; **Fang, Xiaochang**; **Zhang, Huaxiang** (2023). "Multimodal fake news detection via progressive fusion networks". *Information processing & management*, v. 60, n. 1, 103120.
https://doi.org/10.1016/j.ipm.2022.103120

John, Oliver P.; **Srivastava, Sanjay** (1999). "The big five trait taxonomy: history, measurement, and theoretical perspectives". In: Pervin, Lawrence A.; John, Oliver P. (eds.). *Handbook of personality: Theory and research*, pp. 102-138.
https://pages.uoregon.edu/sanjay/pubs/bigfive.pdf

Kang, SeongKu; **Hwang, Junyoung**; **Yu, Hwanjo** (2020). "Multi-modal component embedding for fake news detection". In: *14th international conference on ubiquitous information management and communication (Imcom)*.
https://doi.org/10.1109/IMCOM48794.2020.9001800

Karras, Tero; **Aila, Timo**; **Laine, Samuli**; **Lehtinen, Jaakko** (2018). "Progressive growing of GANs for improved quality, stability, and variation". In: *6th International conference on learning representations*.
https://research.nvidia.com/sites/default/files/pubs/2017-10_Progressive-Growing-of/karras2018iclr-paper.pdf

Karras, Tero; **Laine, Samuli**; **Aila, Timo** (2019). "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, pp. 4396-4405.
https://doi.org/10.1109/CVPR.2019.00453

Kartal, Yavuz-Selim; **Kutlu, Mucahid** (2023). "Re-think before you share: a comprehensive study on prioritizing check-worthy claims". *IEEE transactions on computational social systems*, v. 10, n. 1, pp. 362-375.
https://doi.org/10.1109/TCSS.2021.3138642

Khattar, Dhruv; **Goud, Jaipal-Singh**; **Gupta, Manish**; **Varma, Vasudeva** (2019). "MVAE: multimodal variational autoencoder for fake news detection". In: *The World Wide Web conference*, pp. 2915-2921.
https://doi.org/10.1145/3308558.3313552

Konstantinovskiy, Lev; **Price, Oliver**; **Babakar, Mevan**; **Zubiaga, Arkaitz** (2021). "Toward automated factchecking: developing an annotation schema and benchmark for consistent automated claim detection". *Digital threats: research and practice*, v. 2, n. 2.
https://doi.org/10.1145/3412869

Kudugunta, Sneha; **Ferrara, Emilio** (2018). "Deep neural networks for bot detection". *Information sciences*, v. 467, pp. 312-322.
https://doi.org/10.1016/j.ins.2018.08.019

La-Barbera, David; **Roitero, Kevin**; **Mizzaro, Stefano** (2022). "A hybrid human-in-the-loop framework for fact checking". In: *Proceedings of the 6th Workshop on natural language for artificial intelligence (NL4AI 2022)*, v. 3287.
https://ceur-ws.org/Vol-3287/paper4.pdf

LeCun, Yann; **Bengio, Yoshua**; **Hinton, Geoffrey** (2015). "Deep learning". *Nature*, v. 521, n. 7553, pp. 436-444.
https://doi.org/10.1038/nature14539

Li, Dun; **Guo, Haimei**; **Wang, Zhenfei**; **Zheng, Zhiyun** (2021). "Unsupervised fake news detection based on autoencoder". *IEEE access*, v. 9, pp. 29356-29365.
https://doi.org/10.1109/ACCESS.2021.3058809

Li, Shuo; **Yao, Tao**; **Li, Saifei**; **Yan, Lianshan** (2022). "Semantic-enhanced multimodal fusion network for fake news detection". *International journal of intelligent systems*, v. 37, n. 12, pp. 12235-12251.
https://doi.org/10.1002/int.23084

Li, Xin; **Lu, Peixin**; **Hu, Lianting**; **Wang, Xiao-Guang**; **Lu, Long** (2022). "A novel self-learning semi-supervised deep learning network to detect fake news on social media". *Multimedia tools and applications*, v. 81, n. 14, pp. 19341-19349.
https://doi.org/10.1007/s11042-021-11065-x

Liu, Yang; **Wu, Yi-Fang** (2018). "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks". *Proceedings of the AAAI conference on artificial intelligence*, v. 32, n. 1, pp. 354-361.
https://doi.org/10.1609/aaai.v32i1.11268

Liu, Yang; **Xu, Songhua** (2016). "Detecting rumors through modeling information propagation networks in a social media environment". *IEEE transactions on computational social systems*, v. 3, n. 2, pp. 46-62.
https://doi.org/10.1109/TCSS.2016.2612980

Manning, Christopher D.; **Schütze, Hinrich** (1999). *Foundations of statistical natural language processing*. MIT Press. ISBN: 978 0 262 133609

Marcus, Gary (2022). "AI platforms like *chatGPT* are easy to use but also potentially dangerous". *Scientific American*, 19 December.
https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous

Martín, Alejandro; **Huertas-Tato, Javier**; **Huertas-García, Álvaro**; **Villar-Rodríguez, Guillermo**; **Camacho, David** (2022). "FacTeR-check: semi-automated fact-checking through semantic similarity and natural language inference". *Knowledge-based systems*, v. 251, 109265.
https://doi.org/10.1016/j.knosys.2022.109265

Masood, Momina; **Nawaz, Mariam**; **Malik, Khalid M.**; **Javed, Ali**; **Irtaza, Aun**; **Malik, Hafiz** (2022). "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward". *Applied intelligence*, v. 54, pp. 3974-4026.
https://doi.org/10.1007/s10489-022-03766-z

Meel, Priyanka; **Vishwakarma, Dinesh K.** (2020). "Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities". *Expert systems with applications*, v. 153, 112986.
https://doi.org/10.1016/j.eswa.2019.112986

Meel, Priyanka; **Vishwakarma, Dinesh K.** (2021). "A temporal ensembling based semi-supervised convnet for the detection of fake news articles". *Expert systems with applications*, v. 177, 115002.
https://doi.org/10.1016/j.eswa.2021.115002

Megahed, Fadel M.; **Chen, Ying-Ju**; **Ferris, Joshua A.**; **Knoth, Sven**; **Jones-Farmer, L. Allison** (2023). "How generative AI models such as *chatGPT* can be (mis)used in SPC practice, education, and research? An exploratory study". *ArXiv*.
https://doi.org/10.48550/arXiv.2302.10916

Mikolov, Tomas; **Chen, Kai**; **Corrado, Greg**; **Dean, Jeffrey** (2013). "Efficient estimation of word representations in vector space". In: *1st International conference on learning representations (ICLR)*.
https://arxiv.org/abs/1301.3781

Mirsky, Yisroel; **Lee, Wenke** (2022). "The creation and detection of deepfakes". *ACM computing surveys*, v. 54, n. 1.
https://doi.org/10.1145/3425780

Mitchell, Eric; **Lee, Yoonho**; **Khazatsky, Alexander**; **Manning, Christopher D.**; **Finn, Chelsea** (2023). "DetectGPT: zero-shot machine-generated text detection using probability curvature". *ArXiv*.
https://doi.org/10.48550/arXiv.2301.11305

Molina-Solana, Miguel; **Amador, Julio**; **Gómez-Romero, Juan** (2018). "Deep learning for fake news classification". In: *I Workshop on deep learning*, pp. 1197-1201.
https://sci2s.ugr.es/caepia18/proceedings/docs/CAEPIA2018_paper_207.pdf

Nakamura, Kai; **Levy, Sharon**; **Wang, William Y.** (2020). "Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection". In: *Proceedings of the 12th International conference on language resources and evaluation*, pp. 6149-6157.
https://aclanthology.org/2020.lrec-1.755.pdf

Nakov, Preslav; **Corney, David**; **Hasanain, Maram**; **Alam, Firoj**; **Elsayed, Tamer**; **Barrón-Cedeño, Alberto**; **Papotti, Paolo**; **Shaar, Shaden**; **Da-San-Martino, Giovanni** (2021). "Automated fact-checking for assisting human fact-checkers". In: *Proceedings of the Thirtieth international joint conference on artificial intelligence (IJCAI)*, pp. 4551-4558.
https://doi.org/10.24963/ijcai.2021/619

Nakov, Preslav; **Da-San-Martino, Giovanni**; **Elsayed, Tamer**; **Barrón-Cedeño, Alberto**; **Míguez, Rubén**; **Shaar, Shaden**; **Alam, Firoj**; **Haouari, Fatima**; **Hasanain, Maram**; **Mansour, Watheq**; **Hamdan, Bayan**; **Ali, Zien-Sheikh**; **Babulkov, Nikolay**; **Nikolov, Alex**; **Shahi, Gautam-Kishore**; **Struß, Julia-Maria**; **Mandl, Thomas**; **Kutlu, Mucahid**; **Kartal, Yavuz-Selim** (2021). "Overview of the clef-2021 checkthat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news". In: *International conference of the cross-language evaluation forum for European languages*. *Experimental IR meets multilinguality, multimodality, and interaction*, pp. 264-291.
https://doi.org/10.1007/978-3-030-85251-1_19

Newman, Mark E. J. (2004). "Fast algorithm for detecting community structure in networks". *Physical review E*, v. 69, n. 6, 066133.
https://doi.org/10.1103/PhysRevE.69.066133

Oehmichen, Axel; **Hua, Kevin**; **Amador, Julio**; **Molina-Solana, Miguel**; **Gómez-Romero, Juan**; **Guo, Yi-ke** (2019). "Not all lies are equal. A study into the engineering of political misinformation in the 2016 US presidential election". *IEEE access*, v. 7, pp. 126305-126314.
https://doi.org/10.1109/ACCESS.2019.2938389

Paka, William-Scott; **Bansal, Rachit**; **Kaushik, Abhay**; **Sengupta, Shubhashis**; **Chakraborty, Tanmoy** (2021). "Cross-sean: a cross-stitch semi-supervised neural attention model for Covid-19 fake news detection". *Applied soft computing*, v. 107.
https://doi.org/10.1016/j.asoc.2021.107393

Pasi, Gabriella; **De-Grandis, Marco**; **Viviani, Marco** (2020). "Decision making over multiple criteria to assess news credibility in microblogging sites". In: *IEEE International conference on fuzzy systems (FUZZ-IEEE)*.
https://doi.org/10.1109/FUZZ48607.2020.9177751

Pennebaker, James W.; **Boyd, Ryan L.**; **Jordan, Kayla**; **Blackburn, Kate** (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.
https://repositories.lib.utexas.edu/handle/2152/31333

Pennington, Jeffrey; **Socher, Richard**; **Manning, Christopher** (2014). "GloVe: global vectors for word representation". In: *Proceedings of the 2014 Conference on empirical methods in natural language processing (Emnlp)*, pp. 1532-1543.
https://doi.org/10.3115/v1/D14-1162

Qi, Peng; **Cao, Juan**; **Yang, Tianyun**; **Guo, Junbo**; **Li, Jintao** (2019). "Exploiting multi-domain visual information for fake news detection". In: *IEEE International conference on data mining (ICDM)*, pp. 518-527.
https://doi.org/10.1109/ICDM.2019.00062

Rana, Md-Shohel; **Nobi, Mohammad-Nur**; **Murali, Beddhu**; **Sung, Andrew H.** (2022). "Deepfake detection: a systematic literature review". *IEEE access*, v. 10, pp. 25494-25513.
https://doi.org/10.1109/ACCESS.2022.3154404

Rashkin, Hannah; **Choi, Eunsol**; **Jang, Jin Y.**; **Volkova, Svitlana**; **Choi, Yejin** (2017). "Truth of varying shades: analyzing language in fake news and political fact-checking". In: *Proceedings of the 2017 Conference on empirical methods in natural language processing*, pp. 2931-2937.
https://doi.org/10.18653/v1/D17-1317

Rath, Bhavtosh; **Salecha, Aadesh**; **Srivastava, Jaideep** (2022). "Fake news spreader detection using trust-based strategies in social networks with bot filtration". *Social network analysis and mining*, v. 12, n. 66.
https://doi.org/10.1007/s13278-022-00890-z

Ruffo, Giancarlo; **Semeraro, Alfonso**; **Giachanou, Anastasia**; **Rosso, Paolo** (2023). "Studying fake news spreading, polarisation dynamics, and manipulation by bots: a tale of networks and language". *Computer science review*, v. 47, 100531.
https://doi.org/10.1016/j.cosrev.2022.100531

Russell, Stuart; **Norvig, Peter** (2020). *Artificial intelligence: a modern approach*. Pearson Series. ISBN: 978 0 134 610993

Saif, Shahela; **Tehseen, Samabia** (2022). "Deepfake videos: synthesis and detection techniques - a survey". *Journal of intelligent and fuzzy systems*, v. 42, n. 4, pp. 2989-3009.
https://doi.org/10.3233/JIFS-210625

Schuster, Tal; **Schuster, Roei**; **Shah, Darsh J.**; **Barzilay, Regina** (2020). "The limitations of stylometry for detecting machine-generated fake news". *Computational linguistics*, v. 46, n. 2, pp. 499-510.
https://doi.org/10.1162/coli_a_00380

Serengil, Sefik I.; **Ozpinar, Alper** (2021). "HyperExtended lightface: a facial attribute analysis framework". In: *International conference on engineering and emerging technologies (Iceet)*.
https://doi.org/10.1109/ICEET53442.2021.9659697

Serrano-Guerrero, Jesús; **Olivas, José A.**; **Romero, Francisco P.**; **Herrera-Viedma, Enrique** (2015). "Sentiment analysis: a review and comparative analysis of web services". *Information sciences*, v. 311, pp. 18-38.
https://doi.org/10.1016/j.ins.2015.03.040

Shabani, Shaban; **Charlesworth, Zarina**; **Sokhn, Maria**; **Schuldt, Heiko** (2021). "SAMS: human-in-the-loop approach to combat the sharing of digital misinformation". *CEUR workshop proceedings*, v. 2846.
https://ceur-ws.org/Vol-2846/paper27.pdf

Shao, Chengcheng; **Ciampaglia, Giovanni-Luca**; **Varol, Onur**; **Yang, Kai-Cheng**; **Flammini, Alessandro**; **Menczer, Filippo** (2018). "The spread of low-credibility content by social bots". *Nature communications*, v. 9, n. 1, pp. 4787.
https://doi.org/10.1038/s41467-018-06930-7

Shao, Chengcheng; **Hui, Pik-Mai**; **Wang, Lei**; **Jiang, Xinwen**; **Flammini, Alessandro**; **Menczer, Filippo**; **Ciampaglia, Giovanni-Luca** (2018). "Anatomy of an online misinformation network". *Plos one*, v. 13, n. 4, e0196087.
https://doi.org/10.1371/journal.pone.0196087

Shrestha, Anu; **Spezzano, Francesca** (2022). "Characterizing and predicting fake news spreaders in social networks". *International journal of data science and analytics*, v. 13, n. 4, pp. 385-398.
https://doi.org/10.1007/s41060-021-00291-z

Shu, Kai; **Sliva, Amy**; **Wang, Suhang**; **Tang, Jiliang**; **Liu, Huan** (2017). "Fake news detection on social media: a data mining perspective". *ACM SIGKDD explorations newsletter*, v. 19, n. 1, pp. 22-36.
https://doi.org/10.1145/3137597.3137600

Shu, Kai; **Wang, Suhang**; **Liu, Huan** (2019). "Beyond news contents: the role of social context for fake news detection". In: *Proceedings of the 12th ACM International conference on web search and data mining*, pp. 312-320.
https://doi.org/10.1145/3289600.3290994

Shu, Kai; **Zhou, Xinyi**; **Wang, Suhang**; **Zafarani, Reza**; **Liu, Huan** (2019). "The role of user profiles for fake news detection". In: *Proceedings of the 2019 IEEE/ACM International conference on advances in social networks analysis and mining*, pp. 436-439.
https://doi.org/10.1145/3341161.3342927

Simko, Jakub; **Racsko, Patrik**; **Tomlein, Matus**; **Hanakova, Martina**; **Moro, Robert**; **Bielikova, Maria** (2021). "A study of fake news reading and annotating in social media context". *New review of hypermedia and multimedia*, v. 27, n. 1-2, pp. 97-127.
https://doi.org/10.1080/13614568.2021.1889691

Singh, Prabhav; **Srivastava, Ridam**; **Rana, K. P. S.**; **Kumar, Vineet** (2023). "SEMI-fnd: stacked ensemble based multimodal inferencing framework for faster fake news detection". *Expert systems with applications*, v. 215, 119302.
https://doi.org/10.1016/j.eswa.2022.119302

Solaiman, Irene; **Brundage, Miles**; **Clark, Jack**; **Askell, Amanda**; **Herbert-Voss, Ariel**; **Wu, Jeff**; **Radford, Alec**; **Krueger, Gretchen**; **Kim, Jong-Wook**; **Kreps, Sarah**; **McCain, Miles**; **Newhouse, Alex**; **Blazakis, Jason**; **McGuffie, Kris**; **Wang, Jasmine** (2019). "Release strategies and the social impacts of language models". *ArXiv*.
https://doi.org/10.48550/arXiv.1908.09203

Song, Chenguang; **Teng, Yiyang**; **Zhu, Yangfu**; **Wei, Siqi**; **Wu, Bin** (2022). "Dynamic graph neural network for fake news detection". *Neurocomputing*, v. 505, pp. 362-374.
https://doi.org/10.1016/j.neucom.2022.07.057

Srinivas, P. Y. K. L.; **Das, Amitava**; **Pulabaigari, Viswanath** (2022). "Fake spreader is narcissist; real spreader is Machiavellian prediction of fake news diffusion using psycho-sociological facets". *Expert systems with applications*, v. 207, 117952.
https://doi.org/10.1016/j.eswa.2022.117952

Stella, Massimo; **Ferrara, Emilio**; **De-Domenico, Manlio** (2018). "Bots increase exposure to negative and inflammatory content in online social systems". *Proceedings of the National Scademy of Sciences*, v. 115, n. 49, pp. 12435-12440.
https://doi.org/10.1073/pnas.1803470115

Tacchini, Eugenio; **Ballarin, Gabriele**; **Della-Vedova, Marco L.**; **Moret, Stefano**; **De-Alfaro, Luca** (2017). "Some like it hoax: automated fake news detection in social networks". In: *CEUR Workshop proceedings*, v. 1960.
https://arxiv.org/abs/1704.07506

Thorne, James; **Vlachos, Andreas** (2018). "Automated fact checking: task formulations, methods and future directions". In: *Proceedings of the 27th International conference on computational linguistics*, pp. 3346-3359.
https://aclanthology.org/C18-1283

Tolosana, Rubén; **Vera-Rodríguez, Rubén**; **Fierrez, Julián**; **Morales, Aythami**; **Ortega-García, Javier** (2020). "Deepfakes and beyond: a survey of face manipulation and fake detection". *Information fusion*, v. 64, pp. 131-148.
https://doi.org/10.1016/j.inffus.2020.06.014

Vaswani, Ashish; **Shazeer, Noam**; **Parmar, Niki**; **Uszkoreit, Jakob**; **Jones, Llion**; **Gomez, Aidan N.**; **Kaiser, Łukasz**; **Polosukhin, Illia** (2017). "Attention is all you need". In: *31st Conference on neural information processing systems*.
https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf

Vogel, Inna; **Meghana, Meghana** (2020). "Fake news spreader detection on *Twitter* using character n-grams". In: *CEUR Workshop proceedings*, v. 2696.
https://ceur-ws.org/Vol-2696/paper_59.pdf

Vosoughi, Soroush; **Roy, Deb**; **Aral, Sinan** (2018). "The spread of true and false news online". *Science*, v. 359, n. 6380, pp. 1146-1151.
https://doi.org/10.1126/science.aap9559

Wang, Tingting; **Liu, Hongyan**; **He, Jun**; **Du, Xiaoyong** (2013). "Mining user interests from information sharing behaviors in social media". In: *Pacific-Asia conference on knowledge discovery and data mining*, pp. 85-98.
https://doi.org/10.1007/978-3-642-37456-2_8

Wang, William Y. (2017). "'Liar, liar pants on fire': a new benchmark dataset for fake news detection". In: *55th Annual meeting of the Association for Computational Linguistics*, v. 2, pp. 422-426.
https://doi.org/10.18653/v1/P17-2067

Wang, Yaqing; **Ma, Fenglong**; **Jin, Zhiwei**; **Yuan, Ye**; **Xun, Guangxu**; **Jha, Kishlay**; **Su, Lu**; **Gao, Jing** (2018). "EANN: event adversarial neural networks for multi-modal fake news detection". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 849-857.
https://doi.org/10.1145/3219819.3219903

Wardle, Claire; **Derakhshan, Hossein** (2017). *Information disorder: toward an interdisciplinary framework for research and policy making*. Council of Europe report.
https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Xiong, Shufeng; **Zhang, Guipei**; **Batra, Vishwash**; **Xi, Lei**; **Shi, Lei**; **Liu, Liangliang** (2023). "Trimoon: two-round inconsistency-based multi-modal fusion network for fake news detection". *Information fusion*, v. 93, pp. 150-158.
https://doi.org/10.1016/j.inffus.2022.12.016

Xu, Fan; **Sheng, Victor S.**; **Wang, Mingwen** (2023). "A unified perspective for disinformation detection and truth discovery in social sensing: a survey". *ACM computing surveys*, v. 55, n. 1.
https://doi.org/10.1145/3477138

Yang, Jing; **Vega-Oliveros, Didier**; **Seibt, Tais**; **Rocha, Anderson** (2021). "Scalable fact-checking with human-in-the-loop". In: *IEEE International workshop on information forensics and security (WIFS)*.
https://doi.org/10.1109/WIFS53200.2021.9648388

Yang, Shuo; **Shu, Kai**; **Wang, Suhang**; **Gu, Renjie**; **Wu, Fan**; **Liu, Huan** (2019). "Unsupervised fake news detection on social media: a generative approach". *Proceedings of the AAAI Conference on artificial intelligence*, v. 33, n. 1, pp. 5644-5651.
https://doi.org/10.1609/aaai.v33i01.33015644

Yin, Zhijun; **Cao, Liangliang**; **Gu, Quanquan**; **Han, Jiawei** (2012). "Latent community topic analysis". *ACM transactions on intelligent systems and technology*, v. 3, n. 4.
https://doi.org/10.1145/2337542.2337548

Zhang, Guobiao; **Giachanou, Anastasia**; **Rosso, Paolo** (2022). "SceneFND: multimodal fake news detection by modelling scene context information". *Journal of information science*, Online first.
https://doi.org/10.1177/01655515221087683

Zhang, Xichen; **Ghorbani, Ali A.** (2020). "An overview of online fake news: characterization, detection, and discussion". *Information processing and management*, v. 57, n. 2.
https://doi.org/10.1016/j.ipm.2019.03.004

Zhou, Xinyi; **Jain, Atishay**; **Phoha, Vir V.**; **Zafarani, Reza** (2020). "Fake news early detection". *Digital threats: research and practice*, v. 1, n. 2.
https://doi.org/10.1145/3377478

Zhou, Xinyi; **Zafarani, Reza** (2020). "A survey of fake news: fundamental theories, detection methods, and opportunities". *ACM computing surveys*, v. 53, n. 5.
https://doi.org/10.1145/3395046

Zhu, Q.; **Luo, J.** (2022). "Generative pre-trained transformer for design concept generation: an exploration". *Proceedings of the design society*, v. 2, pp. 1825-1834.
https://doi.org/10.1017/pds.2022.185

9. Annex

Table 1. Datasets created to train models for disinformation classification

Datasets	Application	Source	Size	Source information	Labels	Annotations	Feature coverage	Language	Publicly available	URL
CREDBANK	Credibility assessment	*Twitter*	> 60 million	Social media posts about 1,049 events	Tuple <degree (certainly, probably, uncertain), polarity (accurate, inaccurate, uncertain)>	Mechanical Turk	Content and context features	English	Yes	*https://compsocial.github.io/CREDBANK-data/*
PHEME	Rumour detection	*Twitter*	5,802	Social media posts about 1,049 events	Rumour (1,972), Non-rumour (3,830)	Expert annotation	Content features	English	Yes	*https://figshare.com/articles/dataset/PHEME_dataset_of_rumours_and_non-rumours/4010619*
LIAR	False information detection	PolitiFact.com	12,800	Political statements	pants-fire (1,047), false (2,507), barely- true (2,103), half-true (2,627), mostly-true (2,454), and true (2,053)	Expert annotation	Content and context features	English	Yes	*https://www.cs.ucsb.edu/~william/data/liar_dataset.zip*
FakeNewsNet	Study false information on social media	BuzzFeed.com and PolitiFact.com	422	News content	Fake (211), Real (211)	Expert annotation	Content and context features	English	Yes	*https://github.com/KaiDMML/FakeNewsNet*
MuMiN	Misinformation detection	*Twitter* and 115 fact-checking organisations	12,914 fact-checked claims and 21,565,018 tweets	Social media post and fact-checked claims	Misinformation, factual	Semi-automatic	Content and context features	Multilingual	Yes	*https://mumin-dataset.github.io/gettingstarted/*
MediaEval	Misinformation and conspiracies detection	*Twitter*	3,389	Social media posts	Promotes/Supports Conspiracy, Discusses Conspiracy and Non-Conspiracy	Expert annotation	Content and context features	English	Under request	*https://multimediaeval.github.io*
BuzzFeedNews dataset	False information detection	*Facebook*	2,282	Social media posts from 9 sources (3 right-wing bias, 3 left-wing bias and 3 credible)	Most true (1,669), No factual content (264), Mixture of true and false (245), Mostly false (104)	Expert annotation	Content and context features	English	Yes	*https://webis.de/data/buzzfeed-webis-fake-news-16.html*
BuzzFace dataset	False information detection and bots detection	*Facebook*	> 1.6 millions	Social media posts verified by *BuzzFeed* and comments and reactions about this posts	Only source data (*BuzzFeedNews* dataset) are labelled	Expert annotation	Content and context features	English	Yes	*https://github.com/gsantia/BuzzFace*
FacebookHoax	Hoax detection	*Facebook*	15,500	Social media posts 32 pages (14 conspiracy and 18 scientific)	Hoax (8,923), Non-Hoax (6,577)	Pages assumptions	Content and context features	English	Yes	*https://github.com/gabll/some-like-it-hoax*
FACTOID	False information spreaders detection	*Reddit*	4,150	3,354,450 social media posts authored by 4,150 users	Real news spreader (3,071), Fake news spreader (1,079)	Expert-based automatic annotation	Content and context features	English	Yes	*https://github.com/caisa-lab/FACTOID-dataset*
Spanish Fake News Corpus	False information detection	News media websites	971	News from 9 different topics	Fake (480), Real (491)	Expert annotation	Content features	Spanish	Yes	*https://github.com/jpposadas/FakeNewsCorpusSpanish*
Spanish Fake News Corpus 2.0	False information detection	News media websites and social networks	1,543	News and social media post from 12 different topics	Fake (766), Real (777)	Expert annotation	Content features	Spanish	Yes	*https://github.com/jpposadas/FakeNewsCorpusSpanish*
NLI19-SP	Misinformation detection	*Twitter*	46,919	Social media posts related with a pool of 61 hoaxes identified by fact-checker organisations	Contradiction (406), Entailment (2,521), Neutral (43,992)	Automatic annotation	Content and context features	Spanish and English	Under request	*https://aida.etsisi.upm.es/download/nli19-sp-dataset-fact-check*
PAN-AP-2020 corpus	False information spreaders detection	*Twitter*	500	Social media users from news posted on *Twitter*.	Real news spreader (250), fake news spreader (250)	Expert annotation	Content and context features	Spanish and English	Under request	*https://zenodo.org/record/4039435#.Y2z2fi8ryRs*