# An Experimental Comparison of the Most Popular Approaches to Fake News Detection

Pietro Dell'Oglio[a], Alessandro Bondielli[b], Francesco Marcelloni[a,*], Lucia C. Passaro[b]

[a]*Dipartimento di Ingegneria dell'Informazione, University of Pisa, Largo Lucio Lazzarino, 1, Pisa, Italy*
[b]*Dipartimento di Informatica, University of Pisa, Largo B. Pontecorvo, 3, Pisa, Italy*

**Abstract**

The use of social media platforms as a means of news dissemination and consumption has increased the proliferation of fake news. Fake news detection is becoming a pressing issue which has been faced in the literature with different approaches.

In this paper, we aim to investigate the generalization capabilities of the most popular approaches proposed in the literature by conducting a series of experiments on a collection of 10 different public fake news datasets. Specifically, we evaluate the performance of three types of classifiers based on: (i) traditional machine learning, (ii) deep learning, and (iii) Transformers, both fine-tuned and in a zero-shot setting. For consistency and broader applicability, we framed the problem as a binary classification task, enabling comparisons across a wider range of datasets. Our experiments are designed to evaluate the effectiveness and the generalization capacity of the models trained on specific datasets, with the aim of simulating real-world scenarios with domain shift and out-of-distribution data. This study makes a significant contribution to fake news detection research by offering a comprehensive cross-dataset evaluation of current methods and their generalization limits. While the results reaffirm the efficacy of fine-tuned models and highlight the

*Corresponding author
*Email addresses:* pietro.delloglio@ing.unipi.it (Pietro Dell'Oglio), alessandro.bondielli@unipi.it (Alessandro Bondielli), francesco.marcelloni@unipi.it (Francesco Marcelloni), lucia.passaro@unipi.it (Lucia C. Passaro)

potential of transfer learning capabilities in Large Language Models, they also expose significant limitations in generalization. Our findings underscore the urgent need for more robust and adaptable approaches to automatic fake news detection, capable of addressing the challenges posed by diverse contexts and the constantly evolving nature of disinformation techniques.

*Keywords:* Fake News Detection, Natural Language Processing, Classification, Machine learning, Deep learning, Benchmarking

## 1. Introduction

The growing presence of fake news on social media is becoming a major issue that needs attention. This trend is a consequence of the increasing importance of social media as a tool for news dissemination and consumption [1]. Journalists leverage these platforms to engage public opinion, especially with breaking news stories. Concurrently, users rely on verified social media accounts or their personal networks to stay up-to-date on breaking news and events. Social networks have proven to be remarkably useful, especially during crises, due to their inherent ability to disseminate news faster than traditional media outlets [2].

The distorted use of social media has become increasingly evident in recent years, as exemplified by the rise of the first "infodemic" during the COVID-19 pandemic. Scholars have described this period as a Post-Truth Era [3], where emotions and misinformation hold sway [4]. The Russo-Ukrainian conflict has exacerbated this trend, with disinformation emerging as a powerful strategic tool, as it has often been experimented in conflicts [5].

The lack of control and fact-checking mechanisms on social media platforms facilitates the spread of unverified and false information, which can substantially influence public opinion on crucial issues [6]. A notable example was the 2016 USA presidential election campaign, where the proliferation of fake news played a significant role [7].

Fake news on social media can take various forms, from rumors to clickbait articles, making effective detection and mitigation difficult through both manual and automated approaches. This has led to the establishment of numerous initiatives for independent fact-checking and fake news detection, and the topic has gained significant attention in the research community.

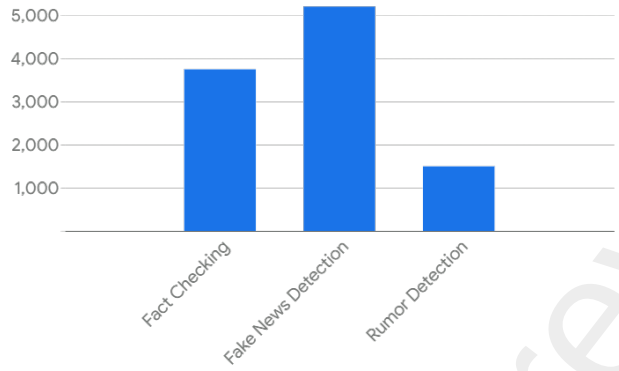



Figure 1: Distribution of papers retrieved from Scopus by searching from 01/01/2017 to 05/03/2025 for "Fact Checking", "Fake News Detection", and "Rumor Detection".

In particular, several competitions [8, 9] have been organized to evaluate computational fact-checking methodologies.

The literature on fake news detection, disinformation, and fact-checking is rapidly increasing despite the complexity and multi-faceted nature of the problem. Figure 1 shows the distribution of the papers retrieved from Scopus by searching for "Fake News Detection", "Rumor Detection" and "Fact Checking" within the title, abstract, and keyword fields of papers published from 01/01/2017 to 05/03/2025. We can observe that the amount of papers is considerable.

In a large chunk of the literature, the fake news detection problem is commonly formulated as a binary classification task aimed at determining whether news text is fake or not, making natural to approach the problem with supervised machine/deep learning strategies. Other lines of research explore alternative methodologies, such as data mining techniques, and leverage external resources to predict documents' credibility. Fact-checking has also played a significant role in these alternative approaches [10].

Despite these diverse methods, supervised learning of binary classifiers remain the most prevalent approach, although it necessitates the use of annotated datasets, which are frequently small in size and exhibit high heterogeneity in terms of topics [11]. This scenario presents a significant risk of developing classifiers that may not generalize well to new out-of-domain data [12].

While there have been attempts to create comprehensive datasets, building high-quality fake news datasets is still challenging [13].

In this paper, first we introduce the most commonly used families of state-of-the-art fake news detection approaches, based on traditional Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), hybrid, Prompting strategies and Large Language Models (LLMs). In the experimental analysis, we employed representative classifiers from each family of approaches. Specifically, we used Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes (NB) to represent traditional ML methods; Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and FakeBERT [14] (i.e., a CNN using pre-trained BERT embeddings) as representatives of DL methods; BERT [15], and DeBERTa [16] to represent fine-tuned Transformers; and LLaMa2 [17] and GPT-4o [18] as Transformers in a zero-shot setting.

Then, we discuss a set of experiments to evaluate the generalization ability of these approaches when managing fake news detection as a binary classification problem. We employed 10 datasets publicly available and already adopted in the literature.

More precisely, we carried out the following three groups of experiments:

**Dataset-specific** This group of experiments aims to find the optimal parameters for each $\langle algorithm, dataset \rangle$ pair. It allows us to assess the performance of each classifier on each specific dataset.

**Cross-Dataset** This group of experiments aims to evaluate the generalization capability of each classifier. Each classifier is trained on a specific dataset using the optimal parameters determined in the first group of experiments and then is tested on all the others.

**Mixed-Training/Single-Test** This group of experiments allows us to simulate real-world scenarios in which a unique model is used to classify cross-domain news. Each classifier is trained on a global training set composed of portions of instances taken from all the datasets. For the evaluation, we used two different test sets. The first test set is obtained by merging the portions of the datasets not used for training. The second test set is built by merging instances not used in the training in equal portions from all datasets.

Note that for all the groups of experiments, decoder-only Transformers (i.e., LLaMa2 and GPT-4o) are used in a zero-shot fashion without tuning. In particular, for the sake of comparison, we make inferences with these

models on exactly the same test data used for the individual experiments of the various groups.

Our findings reveal that while fine-tuned models demonstrate their efficacy, current approaches struggle to generalize across diverse datasets. This highlights the need for more robust solutions capable of adapting to the evolving landscape of misinformation.

The remainder of the paper is organized as follows. Section 2 provides an overview of the various fake news detection approaches proposed in the literature, organized by algorithm families. In Section 3, we present an overview of the publicly available datasets in English. Section 4 details our experimental setup, including the chosen datasets (Section 4.1), the selected techniques (Section 4.2), and parameter optimization (Section 4.3). In Section 5 we discuss the three sets of experiments with their results, namely Dataset-Specific Experiments (Section 5.1), Cross-Dataset Experiments (Section 5.2), and Mixed-Dataset Experiments (Section 5.3), and report the processing time in Section 5.4. In Section 6, we present a review of current surveys on fake news, highlighting their differences with the present work. Section 7 draws some conclusions and future directions.

## 2. Fake News Detection techniques

In recent years, the problem of detecting fake news has gained great popularity in the community of researchers dealing with NLP and Text Mining. However, this is still an open problem with great room for improvement.

The problem is often approached as a binary classification task, aimed at determining whether an article (or any other online text) is a fake news or not. For this reason, researchers have generally proposed approaches based on ML. In recent years, language models (LMs) have also been used, and some steps have been taken by exploiting prompting strategies and Large LMs (LLMs).

### *2.1. Terminology and definitions*

Analyzing the terminology associated with misinformation presents an opportunity to gain valuable insights into this phenomenon. The term "Fake News" has become an umbrella term commonly employed to describe the dissemination of inaccurate information in mainstream media, often conflating it with misinformation and false information. Various categorizations of false information have been proposed in the literature, typically based on

the source and nature of the data under examination. In Bondielli et al., 2019 [19], for instance, false information on the web is split into three subcategories, namely Fake News, Rumors, and others. Fake news is often used to encompass a wide range of false information that is spread via different media, including the web. According to the definition provided by Allcott and Gentzkow, Fake News refers to "a news article that is intentionally and verifiably false" [7]. Thus, the critical attributes of Fake News include intentionality and verifiability. Fake News articles deliberately spread false information that can be authenticated as such, ultimately leading to potential deception of the audience. Numerous recent studies have adopted this definition. Additionally, Fake News has been categorized based on various characteristics, including "serious fabrications", "large-scale hoaxes", and "humorous fakes" [20].

Rumors refer to unconfirmed information primarily disseminated by users on social media platforms. Also in this case there exist several definitions [6, 21].

In addition to Fake News and Rumors, there are other forms of false information. An example is "clickbait", which refers to the headlines of articles, social media posts, or videos designed to attract clicks solely to generate views, often the primary source of revenue online. Notably, the proliferation of clickbait has contributed to the dissemination of Fake News on the web [22]. Another form of false information online closely linked to phishing activities [10] involves "social spammers". These individuals or entities disseminate deceptive or misleading content across social media platforms, contributing to the spread of misinformation.

### *2.2. Traditional ML Approaches*

Early attempts to solve the fake news detection problem employed traditional ML techniques.

An essential distinction among these approaches lies in the features deemed pertinent for detecting fake news. They have been classified into *content features* and *context features* [10]. The former features relate to information directly extracted from the text. They are the most frequently employed features for depicting and identifying fake news because they provide specific signals for detection [23].

Regarding articles, examples of content features encompass the source of the article, the headline (a short text crafted to capture attention and engage the reader), stylistic and linguistic attributes, and audio/video-related

details. Depending on the nature of the content under consideration and the desired representations we aim to construct, a distinction can be made between linguistically-based features and visually-based features [10].

On the other hand, contextual features pertain to information concerning the environment and context in which a particular news item is found. These features are categorized based on three aspects of the context that one intends to represent [10]: the social media users (*user-based features*), the posts generated by users (*post-based features*), and the networks of users (*network-based features*).

The literature shows a predominant trend: most studies focus on applying only a specific category of features. Hybrid models incorporating multiple feature categories also exist. However, most of the fake news detection methods predominantly utilize content-based features [23].

The earliest approaches to fake news detection primarily employed supervised traditional ML algorithms. In particular, the most widely used techniques are LR, SVM, and NB [23].

LR is a statistical technique employed for binary classification tasks, where the goal is to predict the probability that an instance belongs to one of the two classes. In the context of fake news detection, LR has been employed to predict whether a given instance (i.e. a news article or a piece of text) is likely to be fake or real based on content features extracted from the text [24]. Although various content feature representations have been experimented with for fake news detection using LR [25, 26, 27], the most commonly adopted approach in most studies is the well-known Term Frequency-Inverse Document Frequency (TF-IDF) [28, 29, 30, 31, 32].

SVM aims to find the optimal hyperplane that separates data points into different classes with the maximum margin. SVM has been extensively used for binary classification [33] and, in particular, for fake news detection using TF-IDF features [27, 29, 31, 34, 35, 36].

NB is a probabilistic model based on the Bayes theorem, often used for classification tasks. Several approaches based on NB have been proposed for detecting fake news [37, 38] and, in particular, with the use of TF-IDF features [39, 40, 41].

Table 1 summarizes all the approaches described in this subsection.

In the literature, such approaches have been applied with various parameters for feature extraction, often adapting the process to the specific dataset. For the sake of comparison, as described in Section 4, we decided to use the same vectorization method (i.e., TF-IDF), which aligns with common

practices in the literature.

| Reference | Approach | Datasets |
|---|---|---|
| Rubin et al., 2016 [35] | SVM | Self-collected |
| Tacchini et al., 2017 [25] | LR | Self-collected |
| Ahmed et al., 2017 [28] | LR, SVM | Self-collected |
| Granik et al., 2017 [37] | NB | Self-collected |
| Ahmed et al., 2018 [29] | LR, SVM | Self-collected |
| Agudelo et al., 2018 [40] | NB | Dutta dataset |
| Della Vedova et al., 2018 [30] | LR | Self-collected |
| Agarwal et al., 2019 [36] | LR, SVM, NB | LIAR |
| Vicario et al., 2019 [27] | LR, SVM | Self-collected |
| Abdulrahman et al., 2020 [38] | LR, SVM, NB | Kaggle Dataset |
| Sansonetti et al., 2020 [34] | SVM | Self-collected |
| Jiang et al., 2021 [31] | LR, SVM | Isot, KDnugget |
| Rohera et al., 2022 [39] | NB | Dutta dataset |
| Adeyiga et al., 2023 [24] | LR, NB | Kaggle Dataset |
| Mohsen et al., 2024 [32] | LR, NB | Kaggle dataset, Fakes |

Table 1: Existing fake news detection approaches based on traditional ML.

### *2.3. DL Approaches*

Many modern methodologies employ deep neural network architectures, which can learn latent representations from input data, encompassing both contextual and content variations [42]. Nevertheless, deep neural networks often require more memory resources than traditional ML approaches. The challenge shifts towards effectively modeling the network to address the given task proficiently. In the following, we will discuss approaches which adopt both training from scratch and fine-tuning.

#### *2.3.1. Training from scratch*

Several DL-based approaches have been proposed in fake news detection [43, 44]. However, Recurrent Neural Networks (RNNs) and CNNs emerge as typical choices.

FNDNet employs a deep CNN architecture for fake news detection [45], whereas Bugueno et al., 2019 [46], proposed a model based on RNN for propagation tree classification. Bahad et al., 2019 [47], introduced an LSTM-RNN model with higher precision than the initial state-of-the-art CNN approach. Subsequently, Asghar et al., 2021 [48], proposed a BiLSTM combined with CNN. Although the BiLSTM-CNN outperformed other models, this approach is computationally expensive.

Several models were investigated by Girgis et al., 2018 [49], including CNN, LSTM, Vanilla, and Gated Recurrent Unit (GRU). The Vanilla model encountered issues with gradient vanishing, which were mitigated by the GRU model. Despite the superior performance of the GRU in their studies, it required longer training times. More recently, a study [50] on COVID-19 fake news compared GRU, LSTM, and BiLSTM, with BiLSTM outperforming the other two approaches.

Ensemble hybrid approaches have also been adopted and have obtained competitive performances [43, 51, 52]. Recently, efforts have also been made in the domain of Explainable ML. [53].

Table 2 summarizes all the approaches described in this subsection.

| **Ref** | **Approach** | **Datasets** |
|---|---|---|
| Ruchansky et al., 2017 [43] | CSI (based RNN-LSTM) | Self-collected, Weibo |
| Wang et al., 2017 [51] | CNN, BiLSTM, CNN-LSTM | LIAR (multi-class) |
| Zubiaga et al., 2018 [52] | LSTM | Self-collected |
| Girgis et al., 2018 [49] | BiLSTM, CNN, Vanilla, GRU, LSTM | LIAR (multi-class) |
| Shu et al., 2019 [53] | dEFEND | GossipCop, Politifact |
| Bugueno et al., 2019 [46] | RNN-LSTM | Twitter16 (rumors) |
| Bahad et al., 2019 [47] | BiLSTM-RNN | Kaggle dataset |
| Abdulrahman et al., 2020 [38] | CNN-LSTM, RNN-LSTM | Kaggle dataset |
| Kaliyar et al., 2020 [45] | CNN | Kaggle dataset |
| Jiang et al., 2021 [31] | CNN, LSTM, GRU | Isot, KDnugget |
| Asghar et al., 2021 [48] | BiLSTM-CNN | Pheme (rumors) |
| Rohera et al., 2022 [39] | LSTM | Dutta dataset |
| Chen et al., 2023 [50] | BiLSTM | FakeCovid Dataset |

Table 2: Existing fake news detection approaches based on Deep Neural Networks trained from scratch.

In the experimental comparison presented in section 4, we utilized CNN and BiLSTM architectures, drawing inspiration from prior works [50, 51], and incorporated word2vec embeddings [54] as in [50, 51, 55]. Additionally, as an example of a hybrid approach, we implemented FakeBERT [14], which is described in Section 2.3.2, where we discuss the Transformers.

*2.3.2. Fine-tuning and Pre-training*

Transformer architectures [56] such as BERT [15] and its transfer learning abilities have also been applied to the fake news domain [57, 58].

BERT has demonstrated state-of-the-art performance across various NLP tasks. Developed by Google, BERT incorporates pre-trained language representations and operates on a transformer-encoded architecture [15]. Notably, BERT excels in capturing contextual meaning within sentences or texts [59].

The data utilized in the BERT model are generic data gathered from Wikipedia and the Book Corpus. While these data contain a wide range of information, specific information on individual domains is still lacking. To overcome this problem, a study by Jwa et al., 2019 [60], incorporated news data in the pre-training phase to boost fake news identification skills. Compared to the state-of-the-art model stackLSTM, the proposed model exBAKE (BERT with extra unlabeled news corpora) outperformed by a 0.137 F1-score (F1).

Kaliyar et al., 2021 [14], introduced FakeBERT, a model that uses multiple parallel blocks of a one-dimensional deep CNN (1d-CNN) with varying kernel sizes and filters, along with BERT-based word embeddings. This combination allows FakeBERT to effectively handle text, improving its ability to address ambiguity.

Recently, Alghamdi et al., 2024 [61], introduced an approach based on BERT, to capture contextualised semantic knowledge, and a multichannel CNN (mCNN) integrated with stacked Bidirectional Gated Recurrent Units (sBiGRU) to jointly learn multi-aspect language representations. They also proposed a neural-based framework [62] that leverages enhanced Hierarchical Convolutional Attention Networks (eHCAN), which incorporates both style-based and sentiment-based features to enhance model performance.

DeBERTa (Decoding-enhanced BERT with disentangled attention) [16] is another Transformer architecture that has demonstrated excellent results in fake news detection, often outperforming BERT [63, 64, 65]. DeBERTa improves BERT by exploiting two novel techniques. The first is the disentangled attention mechanism, where each word is represented using two vectors that encode its content and position; the attention weights among words are computed using disentangled matrices on their contents and relative positions, respectively. The second is an enhanced mask decoder, which incorporates absolute positions to predict the masked tokens in the pre-training phase. Some studies have also been conducted recently exploiting LLMs with prompting [66]. These models exhibit a notable increase in scale compared to their predecessors. In some cases, they incorporate reinforcement learning techniques during training. Notable examples include OpenAI's GPT [18], Google's LLaMA [17], and Hugging Face's BLOOM [67].

LLMs in the fake news detection domain were investigated in two ways. They are utilized as generators of fake news [68, 69, 70] and as fake news detectors. In particular, the potential of LLMs in fake news detection was investigated in a study by Hu et al., 2023 [71]. First, they conducted an empirical

study and found that a sophisticated LLM such as GPT 3.5 could generally expose fake news and provide desirable multi-perspective rationales but still underperform the fine-tuned BERT. Then, they proposed that current LLMs may not substitute fine-tuned small LMs (SLMs) in fake news detection but can be a good advisor for SLMs by providing multi-perspective instructive rationales. To instantiate this proposal, they designed an adaptive rationale guidance network for fake news detection, in which SLMs selectively acquire insights on news analysis from the LLMs' rationales.

Table 3 summarizes all the approaches described in this subsection.

In our comparison, we used BERT and DeBERTa as examples of transformer architectures for fine-tuning, and LLaMa2 and GPT-4o as examples of LLMs exploited in a zero-shot setting. We also exploit FakeBERT as an example of complex CNN architectures using BERT embeddings.

| **Ref** | **Approach** | **Datasets** |
|---|---|---|
| Slovikovskaya et al., 2019 [57] | BERT, XLNet, RoBERTa | FNC-1 |
| Jwa et al., 2019 [60] | exBAKE | FNC-1 |
| Qazi et al., 2020 [58] | BERT | LIAR (multi-class) |
| Kaliyar et al., 2021 [14] | FakeBERT | Kaggle Dataset |
| Goel et al., 2021 [63] | DeBERTa | FakeNewsAMT, Celebrity |
| Shifath et al., 2021 [64] | DeBERTa | COVID-19 |
| Lekshmiamma et al., 2022 [65] | DeBERTa | CLEF2022-CheckThat! |
| Hu et al., 2023 [71] | GPT3.5turbo | GosspiCop |
| Alghamdi et al., 2024[62] | eHCAN | Politifact, GossipCop |
| Alghamdi et al., 2024 [61] | BERT-mCNN-sBiGRU | Politifact, GossipCop, Fakes, LIAR (binary) |

Table 3: Existing fake news detection approaches that use Transformer-based architectures.

## 3. Available Datasets

A major challenge in fake news detection is the collection and creation of suitable datasets. This task is multifaceted and complex under many lights. First, data collection itself is costly and demanding, as it requires gathering a considerable amount of online information (articles, tweets, etc.) and annotating it with accurate truthfulness evaluations. Ideally, such a dataset or corpus should contain items labeled as fake or real news, or associated with a score indicating truthfulness. Various strategies can be employed for this purpose. One approach is expert-oriented fact-checking, which involves

experts assessing information truthfulness. Websites like BuzzFeedNews[1] or Snopes[2] utilize this strategy, acting as archives of misinformation and hoaxes [19]. However, these websites are typically limited to specific domains and require domain-specific expertise. Furthermore, this expert-oriented approach cannot algorithmically produce usable datasets, thus making difficult to obtain datasets that generalize across different domains [72]. However, a key advantage is the ability to trace a news back to its original source, whether it is a video, social media post, or news article. Another data annotation strategy involves crowdsourcing platforms for data evaluation. It was recently exploited for creating several fake news datasets [72, 73].

A number of publicly available datasets have been proposed for fake news detection and surrounding problems. Compared to other fields and/or problems in NLP, providing publicly available data may be harder for a few reasons. First, identifying relevant data and devising effective strategies to collect them is a complex and time consuming task. Second, the lack of standardized definitions for false information further complicates data labeling. Third, the majority of data resides on social media platforms, which are often restrictive in granting data access due to privacy concerns.

For what concerns fake news in particular, however, some publicly available resources exist. Table 4 shows, to the best of our knowledge, all the datasets in English that were strictly developed to solve the fake news detection task. We report both the domain of the news and the source from which they were collected.

[1] `https://www.buzzfeednews.com/`
[2] `https://www.snopes.com/`

| Dataset | Domain | Source |
|---|---|---|
| FakeNewsNet [10] | Society,Politics | Various, Twitter |
| LIAR [51] | Politics | Facebook, Twitter |
| LIAR-PLUS [74] | Politics | Facebook, Twitter |
| Isot [28] | Multiple | Various |
| BS Detector [75] | Multiple | Various |
| Horne [76] | Politics | Various |
| Dutta dataset [77] | Politics,Society | Various |
| Facebook Hoax [25] | Science | Facebook |
| FNC-1 dataset [78] | Politics, Society, Technologies | Various |
| CLEF2022-CheckThat! [65] | Multiple | Various |
| Fake vs Satire* [79] | Politics | Various |
| Celebrity [72] | Celebrities | Various |
| Fake News AMT [72] | Multiple | Various |
| GROVER [80] | Multiple | Grover |
| CREDBANK [81] | Multiple | Tweet |
| KDNugget [31] | Multiple | Various |
| HWB [82] | Health | Various |
| Infodemic [5] | Covid-19 | Tweet, Various |
| Covid-19 dataset [83] | Covid-19 | Tweet |
| FakeCovid [84] | Covid-19 | Various |
| Notre Dame Fire (NDF) [73] | Notre Dame Fire | Twitter |
| Fakes [85] | War in Syria | Various |
| Cidii (RIDI) [86] | Islamic Issues | Various, Facebook |

Table 4: A non-exhaustive list of the fake news available datasets for fake news detection.

FakeNewsNet [10] is a dataset available on GitHub and contains 1627 articles that represent a sample of news published on Facebook during periods closely related to the 2016 American presidential elections. The articles were published between September 19 and September 23, and between September 26 and September 27. Each post also includes the link that connects it to the referenced article, which was verified by five BuzzFeed journalists. The dataset was subsequently enriched with additional metadata [87].

LIAR is a dataset which was collected from the PolitiFact[3] website through its API [51]. It comprises short statements from different contexts, including newspaper articles or radio interventions. This dataset is structured in such a way that the labels intended to identify the truthfulness of the texts include multiple classes, ranging from "false" to "true", with intermediate classes such as "mostly true", "half true", and "barely true". In addition to the label, each annotation includes a justification for the assigned label, the author, and the context from which it was taken. The dataset was then enriched, and the LIAR-PLUS dataset was produced [74].

[3] `https://www.politifact.com/`

Isot is a compilation of thousands of fake news and truthful articles obtained from different legitimate news sites and sites flagged as unreliable by Politifact.com [28].

BS Detector[4] is a dataset developed by Kaggle and is unique as it is constructed using a web crawler called "BS detector"[5], which is designed to verify the truthfulness of news. In this case, the labels for the dataset instances are determined by the outputs of an automated system, the BS detector extension, rather than human annotators, as in BuzzFeedNews and LIAR. Therefore, it is impossible to define the Kaggle dataset as a gold standard, for this reason [19].

Other interesting resources are: the two Horne datasets [76], a collection of political real news, satire news and fake news from Buzzfeed and other sources; the Dutta dataset [77], the Facebook hoaxes dataset [25]; the dataset produced during the FNC-1 (Fake News Detection Challenge) [78] and the one produced during the CLEF2022-CheckThat! challenge.[6]

An issue that arises in designing datasets for the task of fake news detection is the existence, on the web, of articles that fall under the broad category of false information but are not necessarily fake news as they belong to the realm of satire. In Rubin et al., 2016 [35], efforts were made to differentiate satire from real news, resulting in a corpus of 240 articles encompassing both satirical and real news across four domains (civic, science, business, and soft news). Instead, in Golbeck et al., 2018 [79], the differentiation between satirical news and fake news has been made. In general, there does not seem to be a standard on whether satire is considered fake or real news.

Two unusual datasets are Celebrity and FakeNewsAMT [72]. They are designed and constructed using different approaches, and then used to develop computational models to tackle fake news detection tasks. Celebrity is a dataset where fake and real news are collected online. The authors focused their research on the domain of famous public figures, as they are often subject to rumors, hoaxes, and fake news. The sources include, among others, Entertainment Weekly, People Magazine, and RadarOnline [72]. FakeNewsAMT is a dataset of news that refers to six domains (sports, business, entertainment, politics, technology, and education). The real news articles

[4] https://www.kaggle.com/datasets/mrisdal/fake-news

[5] https://github.com/bs-detector/bs-detector

[6] https://www.kdnuggets.com/2011/02/free-public-datasets.html

were collected using reliable sources, while their fake counterparts were created using Amazon Mechanical Turk, a crowdsourcing platform. Users were asked to transform the real news into fake news and emulate, if possible, the journalistic style [72].

In Grover the fake news are again generated starting from the real ones by using the homonym generative model [80].

CREDBANK[7] contains approximately 60 million tweets published over a hundred days starting from October 2015. These tweets are grouped into events, and each event is annotated using the Amazon Mechanical Turk crowdsourcing platform by thirty annotators [81].

KDNugget [31] contains 3171 real news articles and 3164 fake news articles. The real news is extracted from online newspapers (i.e. the New York Times, WSJ, Bloomberg, NPR, and Guardian) and was published in 2015 and 2016. The fake news articles were randomly selected from datasets available on Kaggle.

Several datasets were also produced about a specific topic that, in some cases, coincides with breaking news, such as HWB [82] regarding health; Infodemic [5] and Covid [83] regarding the COVID-19 pandemic; NDF [73] about the Notre Dame Fire; Fakes [85] about Islamic Issues; and Cidii [86] regarding the war in Syria.

## 4. Experimental Setup

Many current approaches to fake news detection treat the matter as a classification problem. The majority of them exploit supervised learning to learn models from individual real and fake news. The context and the source from which the fake news is propagated are usually not considered. Although there have been several efforts to construct extensive datasets, as reported in Section 3, the available datasets are often limited, leading to challenges with supervised approaches.

In our experiments, we used 10 of the most representative datasets shown in Table 4. Further, we exploited ten methods representative of the different families of approaches outlined in Section 2. These methods were selected based on their prevalence and established effectiveness within the fake news detection literature.

[7] http://compsocial.github.io/CREDBANK-data/

In particular, for the Traditional ML models, we exploited an LR inspired by the one used in Jiang et al., 2021 [31], and an SVM and an NB, both inspired by Rohera et al., 2022 [39]. For the DL Techniques, we exploited a CNN and a BiLSTM, inspired respectively by the works of Wang et al., 2017 [51], and Chen et al., 2023 [50]. We also exploited a BERT-based deep convolutional approach, named FakeBERT [14]. As for LLMs, we used BERT [15] and DeBERTa [16] classifiers with fine-tuning. LLaMa2 [17] and GPT-4o [18] were instead used in a zero-shot fashion. In the following, we refer to FakeBERT (which has been re-implemented as no official implementation have been released) as CNN-BERT for the sake of clarity. It is in fact a CNN architecture that uses pre-trained BERT word embeddings, but it is not a Transformer. Section 4.2 explains the selection of methods used for our experiments. We also give some implementation detail.

### *4.1. Description of the datasets*

Table 5 reports the domain, the type of news, the number of instances belonging to the Fake Class and the Real Class, respectively, for the 10 datasets used in the experiments. We chose these datasets among the publicly available ones in the English language because they allow us to explore different domains and evaluate the cross-domain generalization capability of the models. The datasets can be categorised into two families: *generic datasets* which contain data concerning generic topics such as politics or a mixture of topics, and *narrow datasets* which contain data pertaining to a specific topic (e.g., a world event such as the fire of the Notre Dame Cathedral for Passaro et al., 2022 [73]). The *generic datasets* category includes FakeVsSatire, Horne, Isot, LIAR-PLUS, Politifact and Celebrity, while the *narrow datasets* category includes Cidii, Fakes, Infodemic and NDF. All the instances in the datasets include a textual entry representing an article, a social media post, or a short statement, and an entry with a binary label (i.e., fake or real). Some exceptions and exceptional cases are handled as follows.

Table 5: Datasets used in the experiments. Generic datasets are marked with †; Narrow datasets are marked with ‡ in the domain column.

| Dataset | Domain | Type | N. of Fake News | N. of Real News |
|---|---|---|---|---|
| Celebrity | Vips† | Articles | 250 | 250 |
| Cidii | Islamic Issues‡ | Articles and posts | 300 | 422 |
| FakeVsSatire | Politics† | Articles | 283 | 203 |
| Fakes | War in Syria‡ | Articles | 378 | 426 |
| Horne | Politics† | Articles | 123 | 203 |
| Infodemic | Covid-19‡ | Tweets | 5031 | 5526 |
| Isot | Multiple† | Articles | 22855 | 21416 |
| LIAR-PLUS | Politics† | Short statements | 5654 | 7130 |
| NDF | Notre Dame Fire‡ | Articles and tweets | 216 | 338 |
| Politifact | Politics† | Articles | 183 | 321 |

The LIAR-PLUS dataset is a six-class dataset. We have binarized it, including the classes "True", "Mostly-True", and "Half-True" in the Real class and the "False", "Pants-Fire", and "Barely-True" in the Fake class, as it is commonly done in several other works [47].

The NDF dataset is a binary-class dataset. It is annotated manually and with two different crowdsourcing experiments, producing three labelling variants, named Manually Labelling, Out-of-Context Labelling (OOC), and In-Context Labelling (IC). The IC variant has been made by humans during a crowdsourcing experiment, with the benefit of contextual information to guide their decisions. Annotators in the OOC experiments did not have additional information. For our research, we utilized the IC variant.

The FakeVsSatire dataset presents a unique challenge. It includes satirical information in its real class. Satirical newspaper articles often share features with fake news, making this dataset particularly challenging.

The FakeNewsNet dataset is a comprehensive collection of fake and real articles from Politifact and Gossipcop. However, at the time we downloaded the texts from the links in the dataset, only data from Politifact were available.

All the datasets that were not already available as training, validation, and test sets have been split by using an approach with stratification, following the subdivision made by the authors of the Infodemic dataset (0.6 for the training set and 0.2 for the validation and test sets, respectively). We used the same training, validation, and test sets for all comparison approaches.

To further support the research community and promote transparency,

we provide a dedicated GitHub repository[8], which includes the code for the models used and the references and pointers to the publicly available datasets. We aim to facilitate future research in fake news detection by providing access to our code and facilitating the retrieval of available datasets. Researchers can utilize our repository as a foundation for their experiments, adapt our code to new datasets, and explore novel approaches to the task.

### *4.2. Implementation Details*

We provide below an overview of the approaches utilized in our experiments to evaluate the performance of various ML and DL methods for fake news detection. The selected approaches, summarized in Table 6, represent a mix of traditional ML classifiers and state-of-the-art DL techniques, including Trasformers.

Table 6: Approaches used for the experimentation.

| **Approach** | **Type** | **Features/Method** |
|---|---|---|
| LR | Traditional ML | Tf-Idf (Bow) |
| SVM | Traditional ML | Tf-Idf (Bow) |
| NB | Traditional ML | Tf-Idf (Bow) |
| CNN | DL (Training from scratch) | Word2Vec (Embeddings) |
| BiLSTM | DL (Training from scratch) | Word2Vec (Embeddings) |
| CNN-BERT | DL (Training from scratch) | BERT (Embeddings) |
| BERT | Encoder-only Transformer | BERT (Fine tuninig) |
| DeBERTa | Encoder-only Transformer | DeBERTa (Fine tuninig) |
| LLaMa2 | Decoder-only Transformer | Zero-shot Prompting |
| GPT-4o | Decoder-only Transformer | Zero-shot Prompting |

We acknowledge that the "DL" family of approaches in our analysis encompasses a wide range of methods, differing significantly in architecture complexity and training/running costs. We differentiate between "Traditional ML" and "DL" purely on a technical basis, i.e., based on whether or not they use Deep Neural Networks, regardless of their internal workings.

It is worth noting that the choices made in our experimental setup were guided by our main objective: to compare various methods in a rigorous and fair manner while also ensuring their adaptation to the task of fake news detection. We have applied these choices consistently across all groups of experiments described in Sections 5.1, 5.2, and 5.3.

[8] `https://github.com/PietroDellOglio-nlp/Most-Popular-Approaches-to-Fake-News-Detection`

For traditional ML methods (LR, SVM, and NB in our paper) and DL methods (CNN, BiLSTM, and CNN-BERT in our paper), we acknowledge that the adaptation to the fake news detection task often depends heavily on feature engineering specific to the training dataset. To address this, we opted for a generic feature extraction process that could be applied to all datasets. While this approach may not fully exploit dataset-specific features, it ensures a fair comparison and avoids inadvertently favoring or disadvantaging any particular dataset.

Transformer-based encoders (i.e., BERT and DeBERTa) achieve task specialization through fine-tuning, and we relied on this well-established process to adapt them to fake news detection. For decoder-only models (i.e., LLaMa2 and GPT-4o), we used manually designed prompts to align their outputs with the task requirements. While prompt engineering has its limitations, We aimed to fully leverage the capabilities of these models in a zero-shot setting.

By adopting this approach, we aimed to strike a balance between fairness, rigor, and adaptability.

We used the `scikit-learn` library[9] to implement the LR, SVM, and NB classifiers. In particular, we used the LogisticRegression module for LR, the SVC module for SVM, and the MultinomialNB for NB. MultinomialNB is a simplified variant of NB, particularly well-suited for textual data. It works well for a high-dimensional dataset and is extremely fast, with few tunable parameters. The feature representation method used for these three classifiers is the well-known TF-IDF.

For CNN and BiLSTM implementations we used TensorFlow Keras.[10] We used pre-trained 300-dimensional word2vec embeddings from Google News [54] to warm-start the text embeddings. The implementation of the CNN consists of an Embedding layer followed by a Dropout, a single 1D-convolutional layer with a Global Max Pooling, and a single dense hidden layer. The output layer has one unit and uses a sigmoid activation function. The implementation of the BiLSTM consists of an Embedding layer followed by two LSTMs, one that processes the sequence forward and one that processes it backward. The outputs of both directions are then concatenated and passed to a single dense hidden layer. The output layer has one unit and exploits a sigmoid

[9] `https://scikit-learn.org`
[10] `https://keras.io/`

activation function.

Both the CNN and BiLSTM implementations were optimised by using the Adam optimizer,[11] as it is frequently done in the literature [44, 46, 47].

CNN-BERT is an example of a more complex DL model, and its implementation is inspired by the FakeBERT described in Kaliyar et al., 2021 [14].

BERT and deBERTta were fine-tuned using the Trainer module from Huggingface.[12] As for the BERT model, we selected the `bert-base-cased` version. We opted for it over other variants because it provides a good balance between performance and computational cost as demonstrated by previous research. We decided to employ the cased version because the presence of capital lettering in specific context of the texts is relevant to detect fake news [88]. As for the DeBERTa model, we exploited `deberta-base`.

Finally, LLaMa2 and GPT-4o were utilized in a zero-shot fashion, leveraging their ability to perform text classification tasks without requiring fine-tuning on domain-specific data.

For LLaMa2, we selected the `meta-llama/Llama-2-7b-chat-hf` variant, balancing computational cost with performance suitability for our experiments. For GPT-4o, we used the OpenAI APIs to access its capabilities efficiently.

Below, we provide the prompt employed for classifying each piece of news as either "Fake" or "Real":

```
You have to act as a disinformation detector.
You will be provided with a piece of text (a tweet
or a news article).
Your task is to label the text as 'Fake' or 'Real'.
Answer with a single token.

Here is the text:
{text}
Your answer:
```

To process the output and obtain the specific tokens of "Fake" and "Real" on each model call we used the Outlines library[13] that provides ways to con-

[11] https://keras.io/api/optimizers/adam/
[12] https://huggingface.co/
[13] https://pypi.org/project/outlines/

trol the generation of LMs to make their output more predictable, reducing the completion to a choice between a fixed number of tokens.

### *4.3. Parameter Optimization*

Before carrying out the experiments, except for GPT-4o and LLaMa which were used only on inference, we searched for the best parameterization of each approach on each dataset by exploiting Optuna,[14] an open source hyperparameter optimization framework to automate hyperparameter search. Table 7 details the parameters tuned for each algorithm and the range of values considered during optimization.

For CNN, BiLSTM, CNN-BERT, and BERT, we set a default of 50 epochs and early stopping at 2 epochs to stop training if the validation loss does not decrease for two consecutive epochs. We used F1 computed on the validation set to select the best model for each dataset. Due to computational constraints, we set the batch size to 8. The other parameters were left to their default configuration.

Table 8 shows the best hyperparameters obtained by each approach on each dataset.

## 5. Experimental results

We designed three groups of experiments both to assess the ability of the selected methods to solve the task on different datasets, and to evaluate their generalization capability on unseen data. The first group of experiments is described in Section 5.1. It aims to find the optimal parameters for each (algorithm, dataset) pair, and allows us to assess the performance of each classifier on each specific dataset. The second group of experiments is described in Section 5.2, and aims to evaluate the generalization capability of each classifier. The last group of experiments is described in Section 5.3, and allows us to simulate real-world scenarios in which a unique model is used to classify cross-domain news.

### *5.1. Dataset-Specific Experiments*

In the first set of experiments, we optimize each approach on each dataset. Our goal is to allow each approach to perform at its highest potential for the

[14] https://optuna.org/

Table 7: Values for parameters optimization.

| Model | **Parameter** | **Tested Range** |
|---|---|---|
| LR | Penalty | L2; None |
| | C | 0.01; 0.1; 1.0; 10.0; 100.0 |
| | Solver | lbfgs; liblinear; sag; saga |
| SVM | Penalty | L1; L2 |
| | C | 0.01; 0.1; 1.0; 10.0; 100.0 |
| | Loss | Hinge; Squared Hinge |
| | Dual | True; False |
| NB | Alpha | 0.01; 0.1; 1.0; 10.0; 100.0 |
| | fit_prior | True; False |
| CNN | Filter size | 3; 4; 5 |
| | Number of Filters | 16; 32; 64; 96; 128 |
| | Dropout | 0.2; 0.4; 0.6; 0.8 |
| | Hidden Units | 8; 16; 32; 64 |
| | Learning Rate | 1e-5; 1e-4; 1e-3; 1e-2 |
| BiLSTM | Number of Units | 16; 32; 64; 96; 128 |
| | Dropout | 0.2; 0.4; 0.6; 0.8 |
| | Hidden Units | 8; 16; 32; 64 |
| | Learning Rate | 1e-5; 1e-4; 1e-3; 1e-2 |
| CNN-BERT | CNN Filters | 64; 96; 128 |
| | Kernel Size | 3; 4; 5 |
| | Dense Units | 16; 32; 64 |
| | Learning Rate | 1e-5; 1e-4; 1e-3; 1e-2 |
| BERT | Learning Rate | 4e-5; 2e-5; 3e-2 |
| | Weight Decay | 0.001; 0.01; 0.1; None |
| DeBERTa | Learning Rate | 4e-5; 2e-5; 3e-2 |
| | Weight Decay | 0.001; 0.01; 0.1; None |

Table 8: Best values after parameters optimization.

| Model | Parameters | Celebrity | Cidli | FakeVsSatire | Fakes | Horne | Infodemic | Isot | LIAR | NDF | Politifact |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LR | Penalty | l2 | none | none | l2 | none | l2 | none | none | none | none |
| | C | 10.0 | 0.01 | 0.01 | 100.0 | 0.01 | 100.0 | 100.0 | 1.0 | 0.01 | 0.01 |
| | Solver | lbfgs | lbfgs | lbfgs | lbfgs | lbfgs | lbfgs | lbfgs | lbfgs | lbfgs | lbfgs |
| SVM | Penalty | l1 | l1 | l1 | l1 | l1 | l1 | l1 | l1 | l1 | l1 |
| | C | 1.0 | 100.0 | 10.0 | 100.0 | 100.0 | 1.0 | 1.0 | 0.1 | 100.0 | 10.0 |
| | Loss | sh | sh | sh | sh | sh | sh | sh | sh | sh | sh |
| | Dual | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| NB | Alpha | 1.0 | 0.1 | 0.01 | 100.0 | 0.1 | 0.1 | 0.01 | 1.0 | 0.1 | 0.01 |
| | fit-prior | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| CNN | F.size | 3 | 3 | 4 | 5 | 5 | 4 | 4 | 4 | 4 | 5 |
| | Filters | 96 | 96 | 96 | 64 | 96 | 128 | 96 | 128 | 128 | 32 |
| | Dropout | 0.6 | 0.6 | 0.2 | 0.4 | 0.2 | 0.4 | 0.1 | 0.2 | 0.4 | 0.2 |
| | H.Units | 8 | 8 | 8 | 8 | 32 | 64 | 64 | 16 | 32 | 16 |
| | LR | 1e-05 | 1e-05 | 0.001 | 1e-05 | 0.001 | 1e-05 | 0.001 | 1e-06 | 0.0001 | 0.0001 |
| | Epochs | 11 | 50 | 16 | 13 | 6 | 7 | 4 | 50 | 9 | 26 |
| BiLSTM | Units | 32 | 96 | 128 | 96 | 16 | 16 | 96 | 128 | 96 | 16 |
| | Dropout | 0.2 | 0.1 | 0.8 | 0.6 | 0.2 | 0.8 | 0.1 | 0.6 | 0.6 | 0.4 |
| | H.Units | 8 | 16 | 32 | 16 | 8 | 8 | 64 | 64 | 16 | 16 |
| | LR | 1e-06 | 0.0001 | 1e-06 | 1e-05 | 1e-05 | 0.001 | 0.001 | 0.001 | 0.001 | 0.0001 |
| | Epochs | 3 | 5 | 35 | 20 | 50 | 3 | 5 | 6 | 4 | 50 |
| CNN-BERT | Filters | 96 | 128 | 64 | 128 | 128 | 128 | 128 | 64 | 128 | 128 |
| | K.Size | 4 | 5 | 5 | 3 | 5 | 5 | 3 | 3 | 5 | 5 |
| | D.Units | 64 | 16 | 16 | 32 | 16 | 16 | 64 | 32 | 32 | 32 |
| | LR | 0.001 | 1e-06 | 1e-06 | 1e-05 | 0.001 | 1e-05 | 0.0001 | 1e-05 | 1e-05 | 0.001 |
| | Epochs | 5 | 29 | 32 | 32 | 8 | 7 | 7 | 7 | 30 | 5 |
| BERT | LR | 2e-5 | 2e-5 | 2e-5 | 2e-5 | 2e-5 | 4e-5 | 2e-5 | 2e-5 | 2e-5 | 4e-5 |
| | W.Decay | 0.1 | 0.1 | 0.01 | 0.001 | 0.1 | 0.1 | 0 | 0.001 | 0.01 | 0.1 |
| | Epochs | 3 | 5 | 3 | 3 | 5 | 5 | 5 | 5 | 2 | 2 |
| DeBERTa | LR | 2e-5 | 2e-5 | 2e-5 | 2e-5 | 2e-5 | 2e-5 | 2e-5 | 2e-5 | 2e-5 | 4e-5 |
| | W.Decay | 0.01 | 0 | 0.001 | 0 | 0.001 | 0.01 | 0 | 0.001 | 0.01 | 0 |
| | Epochs | 4 | 5 | 4 | 5 | 4 | 4 | 4 | 4 | 5 | 5 |

dataset. We optimize each model on each dataset's training set, and evaluate its performances on the corresponding test set. GPT-4o and LLaMa2 were used in a zero-shot fashion on the same test sets.

Table 9 shows the F1 achieved by the best models on each dataset.

Table 9: The F1 obtained by each model on each dataset. Best F1 per dataset are in bold.

| Dataset | LR | SVM | NB | CNN | BiLSTM | BERT | DeBERTa | CNN-BERT | LLaMa2 | GPT-4o |
|---|---|---|---|---|---|---|---|---|---|---|
| Celebrity | .60 | .67 | .47 | .54 | .50 | .76 | **.82** | .50 | .50 | **.82** |
| Cidii | .97 | .86 | .86 | .58 | .59 | **.98** | .97 | .59 | .59 | .94 |
| FakeVsSatire | .70 | .69 | .68 | .56 | .51 | .72 | **.78** | .58 | .43 | .58 |
| Fakes | .43 | .44 | .42 | .53 | **.56** | .44 | .47 | .53 | .53 | .51 |
| Horne | .76 | .71 | .68 | .56 | .62 | .77 | **.95** | .70 | .62 | .74 |
| Infodemic | .93 | .93 | .88 | .55 | .70 | **.97** | **.97** | .68 | .51 | .85 |
| Isot | .99 | .99 | .96 | .96 | .97 | **1.0** | **1.0** | .99 | .48 | .76 |
| LIAR-PLUS | .56 | .57 | .56 | .57 | .57 | .61 | .63 | .58 | .57 | **.68** |
| NDF | .86 | .82 | .79 | .56 | .63 | .84 | **.87** | .59 | .60 | .86 |
| Politifact | .77 | .75 | .71 | .64 | .65 | .84 | **.86** | .64 | .35 | .79 |

We observe that DeBERTa outperforms the other models, achieving the highest F1, excelling on most datasets, independently of sizes and domains. The second best performing model is BERT, confirming that pre-trained models fine-tuned with datasets labeled for solving the fake news detection task are most suitable than models trained from scratch. This observation aligns with previous research, such as the comparative analysis in Alghamdi et al., 2022 [89], which evaluated various ML, DL, and Transformer-based models on datasets like LIAR, PolitiFact, GossipCop, and a COVID-19 dataset. Their study, similarly to ours, found that `bert-base` exhibited strong performance, particularly on the COVID-19 dataset, surpassing other models in accuracy. Notable are also the performance obtained by GPT-4o, which in some cases outperforms BERT (Celebrity, NDF, LIAR-PLUS), and DeBERTa only in one case. This is in-line with the research made by Hu et al., 2023 [71]. They made a comparison between GPT-3.5 in zero-shot and few-shot settings with BERT fine-tuned, demonstrating the superiority of BERT in classifying fake and real news using the Gossipcop dataset. However, the performance obtained by GPT-4o is an interesting outcome because one of the issues in fake news detection is the lack of training data. GPT-4o can access data on the web, and most of this data is, in fact, freely

consumable. The majority of the datasets used for the experiment consist of freely available texts; if they are not, it is easy to obtain related pieces of information (NDF, Covid). The performance of GPT-4o on Isot is particularly interesting: it is by far the lowest performing model, and all other trained models achieved near perfect results on the dataset. This may reflect the fact that the dataset could include features that are easily captured and associated with the correct labels via training, but are less evident when using a model in zero-shot settings. In contrast, LLaMa 2, despite its general capabilities, achieves significantly lower results without tuning.

Traditional ML models like LR and SVM perform surprisingly well, often outperforming DL models, particularly on smaller datasets (NDF, Politifact). DL models (CNN, BiLSTM, CNN-BERT) indeed show mixed results, excelling on some datasets but underperforming on others, especially smaller ones. In particular, CNN-BERT and CNN perform similarly on average, but CNN-BERT outperforms CNN on large datasets such as Infodemic and Isot. CNN-BERT has, in fact, a more complex architecture than the simple CNN. Furthermore, a critical issue encountered with these models is their tendency to overfit the training data relatively quickly. This can be observed in the limited number of epochs required before the loss stopped decreasing. This behavior suggests that these models might be overly sensitive to the specific characteristics of the training data, hindering their ability to generalize to unseen examples. This consideration is even more evident when the models were trained on one dataset and tested on different datasets (see Section 5.2). In general, it seems that the parameter search performed has not been sufficient for these DL models, which seem to require more extensive training data and architecture adjustments to improve performance. In the original paper, CNN-BERT was tested on a large dataset downloaded from Kaggle that we could not identify, obtaining an F1 of 0.98. The result, however, is comparable to the one obtained by CNN-BERT on Isot.

In any case, we observe that performances of the models vary significantly across datasets. For example, the Fakes dataset and LIAR-PLUS appear to be challenging for all the models, while Isot is quite simple to solve. It is important to note that the LIAR dataset is originally a 6-class dataset that we transformed into a 2-class dataset. Other papers [61, 89] used this dataset in the same manner, confirming the lower score compared to other datasets such as Politifact and Gossipcop. The low score on the Fakes dataset across all approaches is well known in the literature [32, 61], and even the BERT-mCNN-sBiGRU state-of-the-art hybrid model described in Alghamdi et al.,

2024 [61] cannot reach competitive results if compared to other datasets. It achieves an F1 of 0.71 on Fakes against 0.92 on Politifact and 0.77 on LIAR. This might be due to the difficulty of resolving the data as it is a very focused dataset on a complex event, and model biases could also hamper it.

The simplicity of Isot should not be surprising, as several other works obtained similar results, such as Hakak et al., 2021 [90]. The performances on the FakeVsSatire dataset are also significant. The dataset is challenging, and even BERT and DeBERTa present difficulties. Other datasets, such as Infodemic and Cidii, show variable results across the models. Politifact is widely used and obtains results in line with the current literature [61, 62, 89]

The F1, Precision (P) and Recall (R) achieved by each model on each dataset for each class are shown in Tables 10–13.

Table 10: The F1, P and R obtained by the ML models on each dataset for each class.

| **Dataset** | **LR** | | | | | | **SVM** | | | | | | **NB** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Celebrity | .59 | .62 | .60 | .60 | .58 | .59 | .66 | .70 | .67 | .68 | .64 | .65 | .48 | .76 | .58 | .42 | .18 | .25 |
| Cidii | 1.0 | .93 | .96 | .95 | 1.0 | .97 | .82 | .85 | .83 | .89 | .87 | .88 | .88 | .78 | .83 | .85 | .92 | .89 |
| FakeVsSatire | .87 | .34 | .49 | .67 | .96 | .79 | .68 | .48 | .57 | .69 | .84 | .76 | .81 | .31 | .45 | .65 | .94 | .77 |
| Fakes | .45 | .47 | .46 | .39 | .38 | .38 | .47 | .41 | .44 | .42 | .48 | .45 | .46 | .50 | .48 | .38 | .34 | .36 |
| Horne | .71 | 1.0 | .83 | 1.0 | .36 | .52 | .70 | .92 | .80 | .75 | .36 | .48 | .66 | 1.0 | .79 | 1.0 | .16 | .27 |
| Infodemic | .94 | .92 | .93 | .92 | .94 | .93 | .93 | .94 | .93 | .93 | .92 | .92 | .89 | .88 | .88 | .87 | .88 | .87 |
| Isot | .98 | .98 | .98 | .98 | .98 | .98 | .99 | .99 | .99 | .99 | .99 | .99 | .94 | .97 | .96 | .97 | .94 | .96 |
| LIAR-PLUS | .60 | .60 | .60 | .49 | .49 | .49 | .61 | .61 | .61 | .50 | .49 | .50 | .60 | .62 | .61 | .49 | .47 | .48 |
| NDF | .86 | .92 | .89 | .86 | .76 | .81 | .81 | .92 | .86 | .85 | .67 | .75 | .89 | .75 | .81 | .68 | .86 | .76 |
| Politifact | .72 | 1.0 | .83 | 1.0 | .28 | .44 | .78 | .91 | .84 | .75 | .52 | .61 | .69 | .94 | .80 | .71 | .23 | .35 |

Table 11: The F1, P and R obtained by the CNN and BiLSTM models on each dataset for each class.

| **Dataset** | **CNN** | | | | | | **BiLSTM** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Celebrity | .52 | .88 | .65 | .62 | .20 | .30 | .50 | .16 | .24 | .50 | .84 | .62 |
| Cidii | .44 | .06 | .11 | .58 | .94 | .72 | .50 | .05 | .09 | .58 | .96 | .73 |
| FakeVsSatire | .00 | .00 | .00 | .57 | .96 | .71 | .39 | .31 | .35 | .56 | .64 | .60 |
| Fakes | .52 | 1.0 | .69 | .00 | .00 | .00 | .55 | .77 | .65 | .55 | .31 | .40 |
| Horne | .00 | .00 | .00 | .57 | .96 | .71 | .62 | 1.0 | .76 | .00 | .00 | .00 |
| Infodemic | .57 | .55 | .56 | .52 | .54 | .53 | .66 | .85 | .74 | .76 | .53 | .62 |
| Isot | .98 | .93 | .96 | .94 | .98 | .96 | .95 | .98 | .97 | .98 | .95 | .97 |
| LIAR-PLUS | .57 | .90 | .70 | .51 | .12 | .20 | .59 | .76 | .66 | .51 | .33 | .40 |
| NDF | .61 | .73 | .67 | .40 | .27 | .32 | .75 | .58 | .66 | .51 | .69 | .59 |
| Politifact | .64 | .98 | .78 | .00 | .00 | .00 | .65 | 1.0 | .78 | .00 | .00 | .00 |

Table 12: The F1, P and R obtained by the BERT, CNN-BERT and DeBERTa models on each dataset for each class.

| **Dataset** | **BERT** | | | | | | **CNN-BERT** | | | | | | **DeBERTa** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Celebrity | .75 | .78 | .76 | .77 | .74 | .75 | .50 | .96 | .65 | .50 | .04 | .07 | .80 | .86 | .83 | .85 | .78 | .81 |
| Cidii | 1.0 | .95 | .97 | .96 | 1.0 | .98 | .00 | .00 | .00 | .58 | 1.0 | .73 | .97 | .98 | .97 | .97 | .95 | .96 |
| FakeVsSatire | .79 | .46 | .58 | .70 | .91 | .79 | .00 | .00 | .00 | .58 | 1.0 | .73 | .69 | .83 | .76 | .86 | .74 | .79 |
| Fakes | .37 | .48 | .42 | .52 | .40 | .45 | .52 | 1.0 | .69 | .00 | .00 | .00 | .49 | .27 | .35 | .46 | .68 | .55 |
| Horne | .71 | .95 | .81 | .81 | .36 | .50 | .81 | .65 | .72 | .57 | .76 | .65 | .97 | .95 | .96 | .92 | .96 | .94 |
| Infodemic | .97 | .98 | .97 | .97 | .97 | .97 | .66 | .76 | .71 | .69 | .57 | .62 | .97 | .99 | .98 | .98 | .96 | .97 |
| Isot | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | .99 | .98 | .99 | .98 | .99 | .99 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| LIAR-PLUS | .65 | .67 | .66 | .57 | .54 | .55 | .58 | .91 | .71 | .57 | .14 | .23 | .67 | .69 | .68 | .59 | .56 | .57 |
| NDF | .82 | .92 | .87 | .85 | .69 | .76 | .60 | .92 | .73 | .28 | .04 | .08 | .95 | .84 | .89 | .78 | .93 | .85 |
| Politifact | .87 | .88 | .87 | .78 | .76 | .77 | .66 | .88 | .76 | .47 | .19 | .27 | .94 | .84 | .89 | .75 | .89 | .81 |

Table 13: The F1, P and R obtained by the LLaMa2 and GPT-4o models on each dataset for each class.

| Dataset | LLaMa2 | | | | | | GPT-4o | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Class 0 | | | Class 1 | | | Class 0 | | | Class 1 | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Celebrity | .50 | .98 | .66 | .50 | .02 | .04 | .93 | .70 | .80 | .76 | .95 | .84 |
| Cidii | .59 | 1.0 | .74 | .00 | .00 | .00 | .94 | .96 | .95 | .94 | .92 | .93 |
| FakeVsSatire | .42 | 1.0 | .59 | 1.0 | .02 | .03 | .50 | .01 | .03 | .58 | .99 | .73 |
| Fakes | .53 | .99 | .69 | .50 | .01 | .03 | .53 | .97 | .68 | .39 | .02 | .04 |
| Horne | .62 | 1.0 | .77 | .00 | .00 | .00 | .95 | .63 | .76 | .61 | .94 | .74 |
| Infodemic | .52 | .96 | .67 | .36 | .02 | .04 | .77 | .99 | .87 | .98 | .68 | .81 |
| Isot | .48 | .98 | .65 | .29 | .01 | .01 | .66 | 1.0 | .79 | 1.0 | .52 | .68 |
| LIAR-PLUS | .57 | .98 | .72 | .59 | .03 | .06 | .70 | .69 | .69 | .61 | .63 | .62 |
| NDF | .61 | .97 | .75 | .33 | .02 | .04 | .82 | .97 | .89 | .93 | .66 | .77 |
| Politifact | .50 | .10 | .17 | .32 | .81 | .46 | .79 | .94 | .86 | .84 | .57 | .68 |

We computed the Friedman test to verify if there exist statistical significant differences between the models. The Null hypothesis (H0) states that there is no significant difference in performance among the models, while the alternative hypothesis (H1) states that at least one model performs significantly differently than the others. For the computation of the statistic and the p-value, we used the SciPy method. [15]

We obtained a Friedman Test Statistic of 46.37 and a p-value of 5.1e-07. Since the p-value is less than alpha (0.05), the null hypothesis is rejected, indicating a significant difference. The post-hoc Nemenyi test is then performed to identify on which specific pairs of models there is statistically significant difference. Figure 2 shows the p-values computed by the post-hoc Nemenyi Test. Each cell represents the p-value for the pairwise comparison between the models. A p-value below 0.05 indicates a statistically significant difference in performance between the two models in the corresponding row and column. We observe that DeBERTa and BERT show statistically significant differences in performance compared to NB, CNN, BiLSTM, and LLaMa2 across the datasets.

### *5.2. Cross-Dataset Experiments*

In the second set of experiments we performed a cross-dataset evaluation to assess how well models trained on one dataset can generalize to the other

[15] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.friedmanchisquare.html`

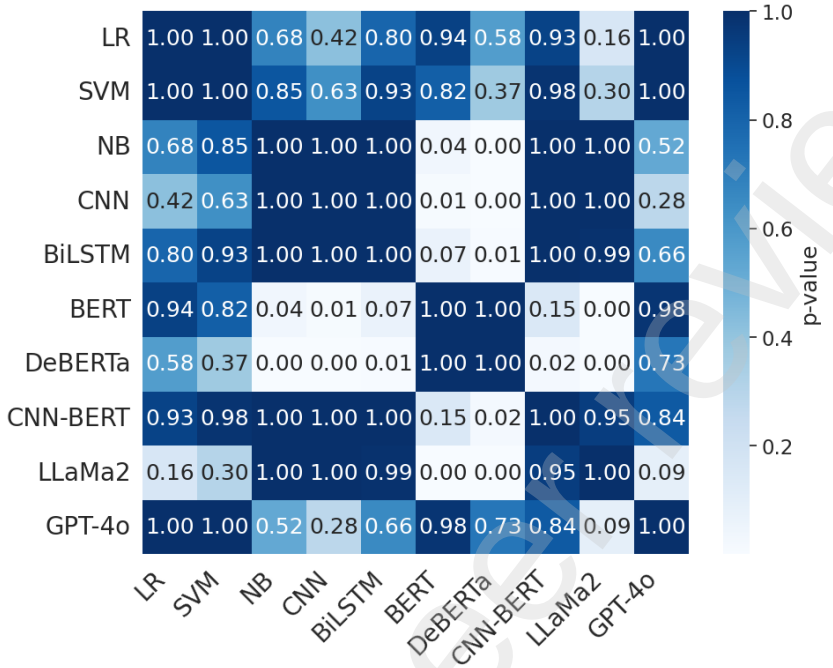



Figure 2: Heatmap of the p-values computed by the post-hoc Nemenyi Test (Pairwise Model Comparisons)

datasets. Generalization is crucial in the context of automatic fake news detection, as fake news constantly evolve and mimic the style and language of real news, making them increasingly difficult to detect. They are often based on highly plausible events, requiring in-depth contextual knowledge and a strong ability to understand the text. A model that generalizes well should adapt to these changes and maintain its effectiveness even as the landscape of fake news shifts.

Our goal is to evaluate the ability of the models considered in this paper to achieve good classification performances on datasets different from that used in the training phase. For each model and for each dataset, we used the best parametrization that emerged after the parameter optimization described in the previous sub-section. We trained the model on one dataset and then we tested it on the remaining datasets. In this experiment, we utilize all available data points from a dataset for training, disregarding the original train, test, and validation splits. We repeated the process for each model and each dataset, except for LLaMa2 and GPT-4o, which were only tested

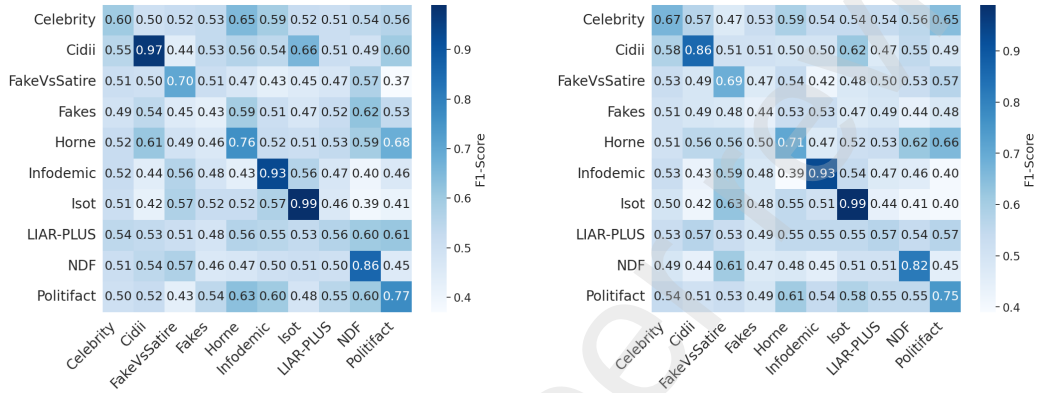
on the datasets without performing the training phase.

Figures 3–5 present the performance obtained by using the traditional ML approaches, the DL approaches trained from scratch and the Transformer-based approaches, respectively. Each sub-figure is a heatmap showing the F1 scores of a single model across the datasets. As a reference, the highest F1 obtained by the models during the experiment described in Section 5.1 are reported in the diagonal of the heatmap. By analyzing the heatmaps, we observe generally lower performance when testing on other datasets, even for models that performed best in the first experiment. In some cases, however, DeBERTa excels compared to other models. For instance, DeBERTa demonstrates strong generalization when trained on the Isot dataset and evaluated on NDF and Celebrity, when trained on Horne and evaluated on Celebrity and Politifact, or conversely, when trained on Politifact and evaluated on Celebrity and Horne. This suggests that the DeBERTa architecture may enhance generalization across diverse domains, though this advantage is not consistently observed across all datasets.

Figure 6 shows the F1 obtained by GPT-4o and LLaMa2 on all the datasets. We remind readers that we are using a zero-shot setting, and therefore, there is no training phase involved. As expected, these F1 are similar to the ones reported in the experiment in Section 5.1. The figure confirms that GPT-4o generally achieves high scores across the majority of datasets, often exceeding 0.80. It has the lowest scores on Fakes and FakeVsSatire datasets, which are particularly challenging for all the models. LLaMa2 confirms its inability to solve the fake news detection task in a zero-shot setting, performing worst on all the datasets.

The performance of fake news detection models clearly depends on the specific characteristics of the datasets used for training and evaluation. However, it is interesting to observe how all the models obtained very high scores in the first experiment described in Section 5.1 when trained and tested on Isot. In the second experiment, when trained on Isot and tested on the remaining datasets their performances decrease considerably. Combining this observation with the fact that GPT-4o and LLaMa2 perform the worst on the Isot dataset compared to the other approaches, we can deduce that the Isot dataset possesses unique characteristics.

To further explore performance variations with respect to the characteristics of the training dataset, we conducted additional experiments aimed at disentangling the impact of the dataset domain. As described in Section 4.1, we are using two different categories of datasets in our experiments: generic

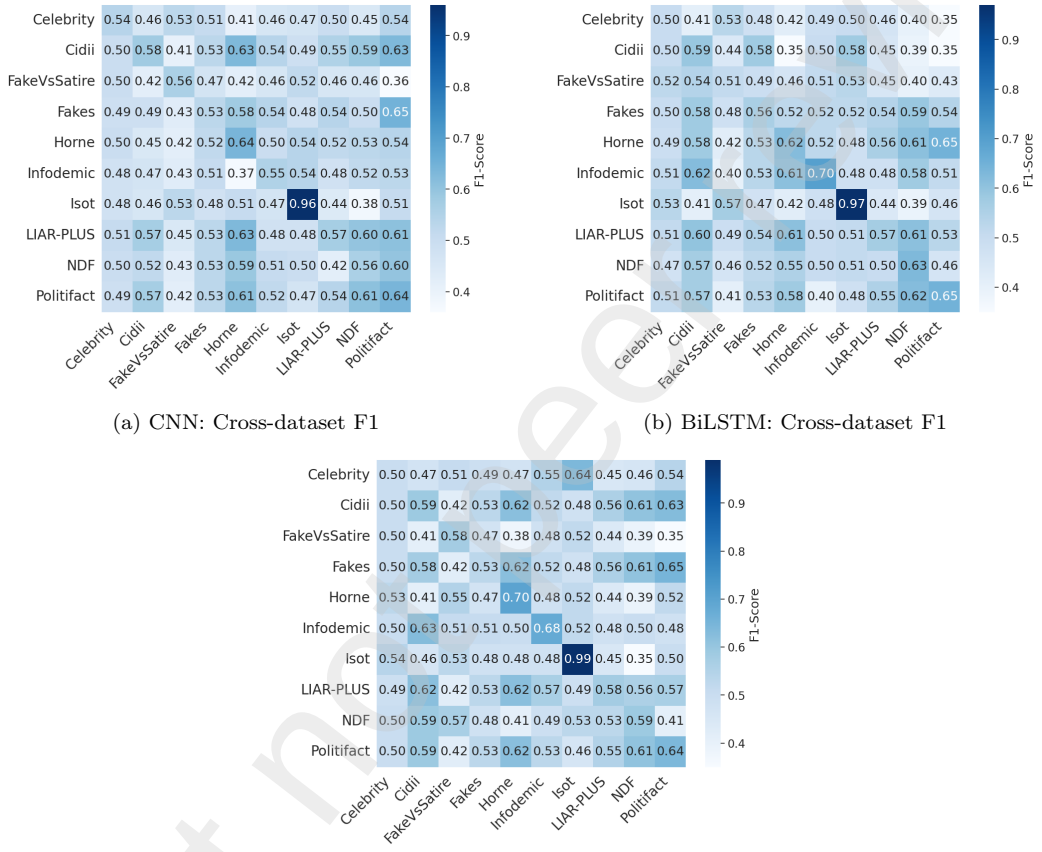



(a) LR: Cross-dataset F1

(b) SVM: Cross-dataset F1

(c) NB: Cross-dataset F1

Figure 3: Cross-dataset F1 for traditional ML approaches. As a reference, we have reported in the diagonal of the heatmap the highest F1 obtained by the models during the experiment described in Section 5.1.

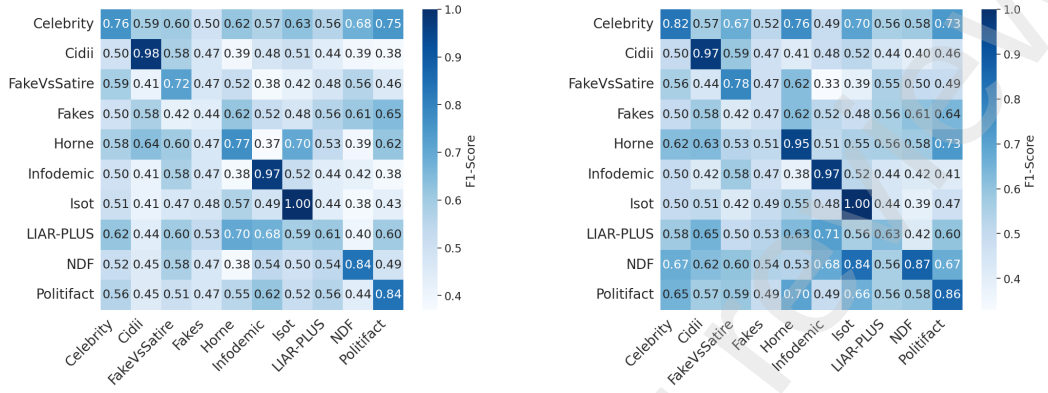



(a) CNN: Cross-dataset F1

(b) BiLSTM: Cross-dataset F1

(c) CNN-BERT: Cross-dataset F1

Figure 4: Cross-dataset F1 for DL approaches trained from scratch. As a reference, we have reported in the diagonal of the heatmap the highest F1 obtained by the models during the experiment described in Section 5.1.




(a) BERT: Cross-dataset F1

(b) DeBERTa: Cross-dataset F1

Figure 5: Cross-dataset F1 for transformer-based models. As a reference, we have reported in the diagonal of the heatmap the highest F1 obtained by the models during the experiment described in Section 5.1.


(a) LLaMa2

(b) GPT-4o

Figure 6: LLaMa2 and GPT-4o: F1 obtained in a zero-shot setting for each dataset.

Table 14: Average F1 achieved from the models on the generic† (i.e Horne, Isot, FakeVsSatire, Politifact, LIAR-PLUS, Celebrity) and narrow‡ (i.e. Cidii, NDF, Fakes and Infodemic) datasets, when trained on the dataset in the column. Best F1 in bold.

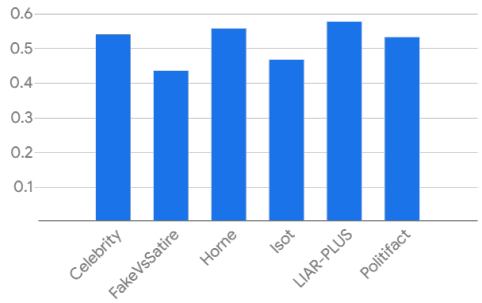
| Model | Test | Celebrity† | Cidii‡ | FakeVsSatire† | Fakes‡ | Horne† | Infodemic‡ | Isot† | LIAR† | NDF‡ | Politifact† |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LR | Generic | .55 | .45 | .45 | .51 | .54 | **.50** | .49 | .55 | .50 | .52 |
| | Narrow | .53 | .47 | **.50** | .53 | .49 | .47 | .48 | .52 | .47 | **.55** |
| SVM | Generic | .55 | .48 | **.52** | .49 | .55 | .48 | .50 | .54 | .50 | .56 |
| | Narrow | .51 | .47 | .49 | .50 | .49 | **.50** | .49 | .50 | .46 | .51 |
| NB | Generic | .53 | .47 | .48 | .52 | .52 | .48 | **.52** | .55 | .48 | .51 |
| | Narrow | .53 | .50 | **.50** | .51 | .50 | .47 | **.50** | .51 | .47 | .52 |
| CNN | Generic | .48 | .47 | .45 | .52 | .50 | .47 | .49 | .53 | .52 | .50 |
| | Narrow | .47 | .44 | .49 | .49 | .50 | **.50** | .49 | .52 | .48 | .51 |
| BiLSTM | Generic | .45 | .49 | .47 | .52 | .51 | **.50** | .48 | .53 | .49 | .50 |
| | Narrow | .48 | .49 | .49 | .43 | .52 | .48 | .48 | .51 | .48 | .50 |
| CNN-BERT | Generic | .52 | .46 | .43 | **.53** | .51 | .49 | .44 | .52 | .49 | .51 |
| | Narrow | .52 | .44 | .43 | .51 | .51 | .46 | .44 | .52 | .45 | .51 |
| BERT | Generic | .63 | **.53** | .49 | **.53** | **.60** | .46 | .48 | **.62** | .50 | .53 |
| | Narrow | **.54** | **.55** | .49 | .51 | .39 | .49 | .48 | .54 | .51 | .52 |
| DeBERTa | Generic | **.68** | .48 | **.52** | **.53** | .59 | .47 | .47 | .57 | **.64** | **.63** |
| | Narrow | **.54** | .45 | .43 | **.57** | **.55** | .43 | .46 | **.57** | **.61** | .53 |

datasets with open-domain data (e.g., Horne, Isot, FakeVsSatire, Politifact, LIAR-PLUS, and Celebrity) and narrow datasets focused on a specific time span, topic, or domain (e.g., Cidii, NDF, Fakes, and Infodemic). We train the model on one of the generic (narrow) datasets and test it on all the remaining datasets.

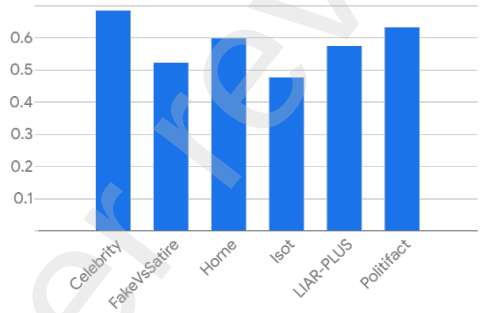
Table 14 shows the average F1 achieved by the models on the generic and narrow datasets, based on the dataset indicated in the column they were trained on. We exclude GPT-4o and LLaMa2, as they were not fine-tuned further.

In Figure 7 we also show the results using bar plots for the the best-performing model (i.e., DeBERTa) to provide better visualization. Specifically, each sub-figure shows the F1 for each of the four combinations: Generic-to-Narrow (7a); Generic-to-Generic (7b); Narrow-to-Narrow (7c); Narrow-to-Generic (7d).

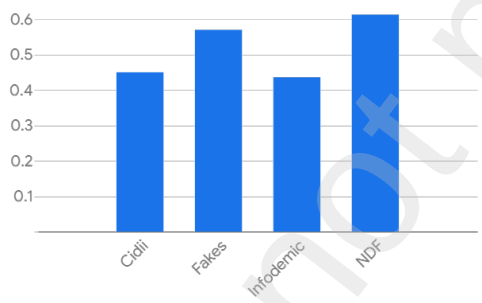
As for the models trained on generic datasets, we observe better perfor-



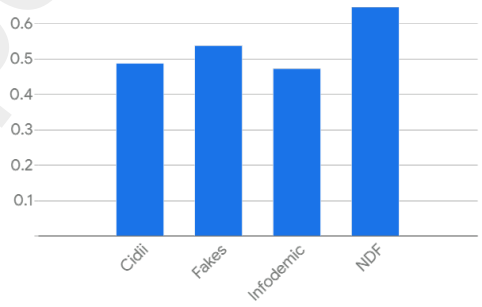

(a) Average F1 achieved on narrow datasets by DeBERTa trained on the generic dataset identified by the label on the X axis.




(b) Average F1 achieved on generic datasets by DeBERTa trained on the generic dataset identified by the label on the X axis.




(c) Average F1 achieved on narrow datasets by DeBERTa trained on the narrow dataset identified by the label on the X axis..




(d) Average F1 achieved on generic datasets by DeBERTa trained on the narrow dataset identified by the label on the X axis.

Figure 7: Average F1 achieved on the generic (narrow) datasets by DeBERTa trained on the narrow (generic) dataset identified by the label on the X axis.

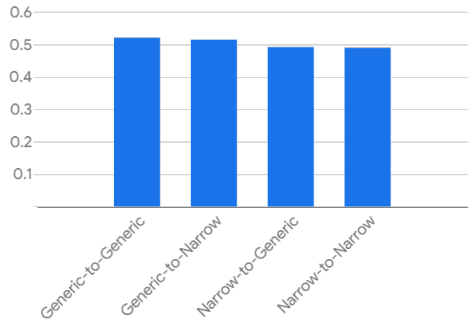
Table 15: Average F1 obtained by the all the approaches in the four scenarios. Best F1 in bold.

| Model | Test | Generic | Narrow |
|---|---|---|---|
| LR | Generic | .52 | .49 |
| | Narrow | .51 | .49 |
| SVM | Generic | .54 | .49 |
| | Narrow | .50 | .48 |
| NB | Generic | .52 | .49 |
| | Narrow | .51 | .49 |
| CNN | Generic | .49 | .49 |
| | Narrow | .50 | .48 |
| BiLSTM | Generic | .49 | .50 |
| | Narrow | .50 | .47 |
| CNN-BERT | Generic | .49 | .47 |
| | Narrow | .49 | .49 |
| BERT | Generic | .56 | .51 |
| | Narrow | .49 | **.51** |
| DeBERTa | Generic | **.58** | **.53** |
| | Narrow | **.51** | **.51** |

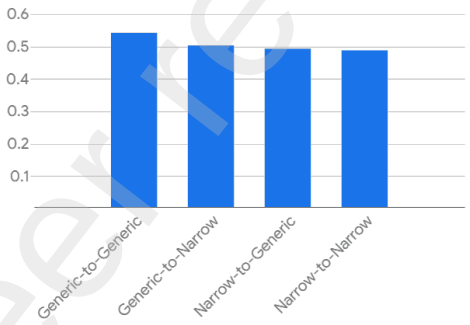
mances when also tested on generic datasets, with a particular improvement in the case of the DeBERTa trained on Horne, Politifact, and Celebrity. As for the models trained on narrow datasets, differences are less pronounced with the exception of NDF. However, the scenario in which DeBERTa is tested on generic datasets generally performs better than when it is tested on narrow datasets, with the unique exception when it is trained on Fakes.

Table 15 presents the average F1 in the four scenarios (Generic-to-Generic, Generic-to-Narrow, Narrow-to-Generic, and Narrow-to-Narrow) for all the approaches. To clarify further, let us consider the Generic-to-Generic scenario. In this case, we used one generic dataset for training and the other generic datasets for testing. The reported F1 represent the average of the F1 obtained when each generic dataset was used as the training set. Figures 8–10 show the same average F1 in a plot visualization for the traditional ML, DL, and Transformer-based approaches, respectively.

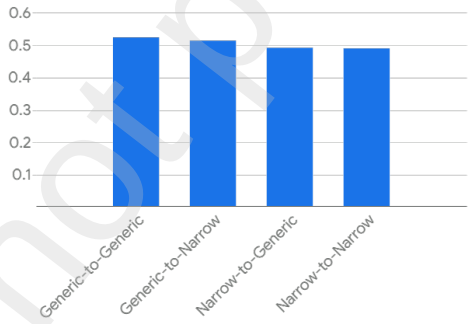
We observe that, in general, training BERT and DeBERTa on generic datasets and testing on generic datasets tend to yield better average F1 compared to other scenarios. The "Generic-to-Generic" scenario outperforms the "Narrow-to-Narrow" scenario, suggesting that training on a diverse set of fake news examples leads to a model that can better generalize across different




(a) Average F1 achieved by LR in each scenario.

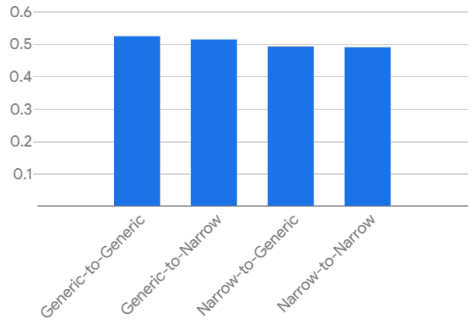



(b) Average F1 achieved by SVM in each scenario.

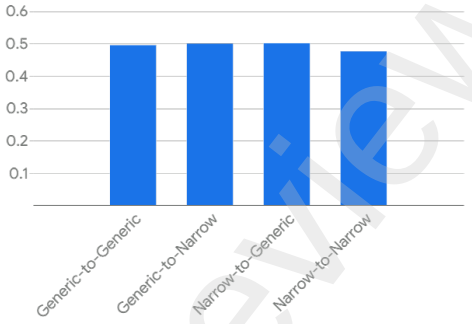



(c) Average F1 achieved by NB in each scenario.

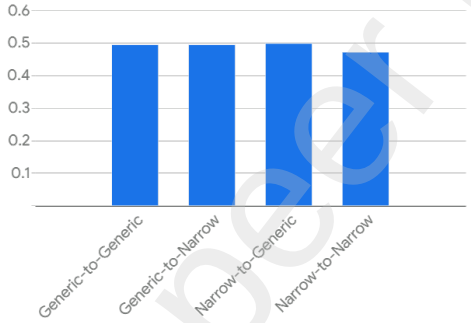
Figure 8: Average F1 obtained by the traditional ML approaches in the four scenarios.




(a) Average F1 achieved by CNN in each scenario.

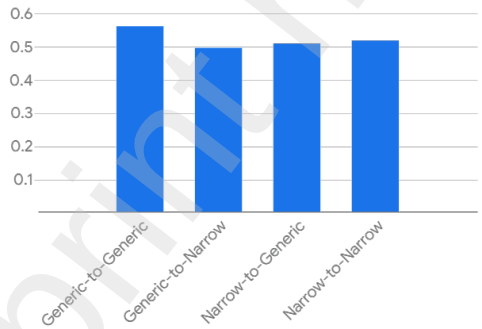



(b) Average F1 achieved by BiLSTM in each scenario.

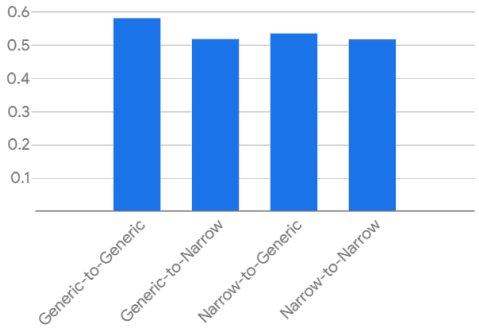



(c) Average F1 achieved by CNN-BERT in each scenario.

Figure 9: Average F1 obtained by the DL models approaches in the four scenarios.




(a) Average F1 achieved by BERT in each scenario.




(b) Average F1 achieved by DeBERTa in each scenario.

Figure 10: Average F1 obtained by the transformer-based architectures in the four scenarios.

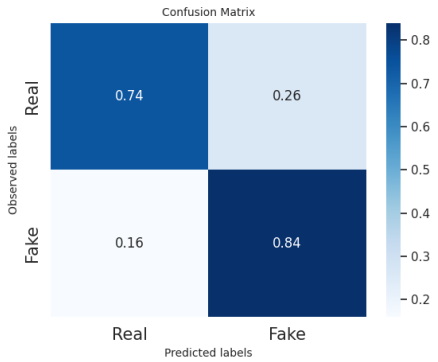
types of generic fake news. The "Generic-to-Generic" scenario also performs better than the "Generic-to-Narrow" scenario, implying that a model trained on diverse generic data might be less effective when applied to narrow types of fake news.

In general, we can observe a rather poor generalization capability in all scenarios, including the "Generic-to-Generic" one. If we analyze the errors in more detail, we observe an interesting aspect. Figure 11 shows four confusion matrices as examples of the "Generic-to-Generic" scenario. The matrices present the results obtained by DeBERTa, respectively, on Horne trained on Celebrity (Fig. 11a), on Politifact trained on Horne (Fig. 11b), on Isot trained on NDF (Fig. 11c) and on Horne trained on Politifact (Fig. 11d). They are the four best results obtained by DeBERTa, the best-performing model if we exclude GPT-4o. While these cases demonstrate promising generalization capabilities (particularly Fig. 11a, Fig. 11c, and Fig. 11d), they are not representative of the overall performance of DeBERTa. The majority of scenarios, including the relatively high-performing (Fig. 11b), exhibit a significant number of false negatives, i.e. the model classifies a fake news as a real one. This behaviour is particularly relevant for fake news detection in a real world scenario, and arguably more dangerous than the opposite in terms of impact on society.

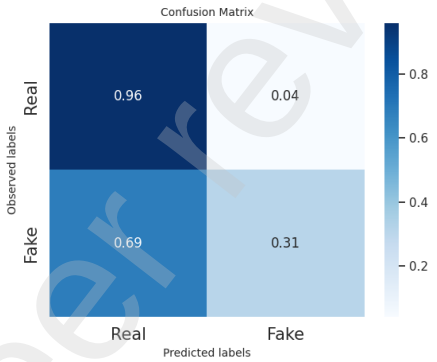
It is possible that a classification-based approach that involves training on specific data to learn the distinction between real and fake news may be sub-optimal, as the domain and the timing of the news, as well as its relationship with other news, may play a crucial role that is overlooked when leveraging simple text-based classification.

### *5.3. Mixed-Dataset Experiments*

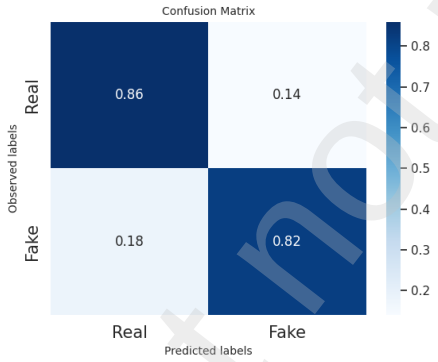
A possible strategy for mitigating the issues highlighted in Section 5.2 may be to leverage a training dataset that includes as much variety as possible in its contents and real/fake news. To address this, in the third experiment we create a training set, named *global training set*, comprising instances extracted from all the datasets. For each dataset, we randomly selected the same number of instances for the Fake and Real classes. This number was set to 100 and was imposed by the Horne dataset size, the smallest dataset. Thus, the global training set is a balanced set comprising 2000 instances, divided into 1000 fake and 1000 real. The parameters used for training are the ones used in the best-performing dataset during the parameter optimization made during the first experiment. We tested the models on each dataset,



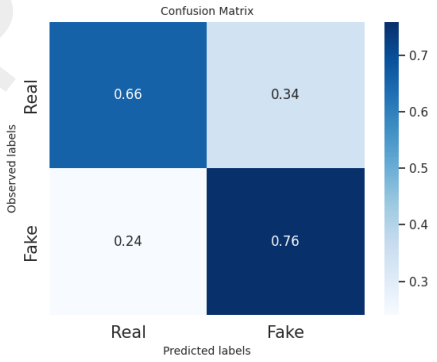

(a) Confusion Matrix obtained on Horne (test set) by DeBERTa trained on Celebrity (training set)




(b) Confusion Matrix obtained on Politifact (test set) by DeBERTa trained on Horne (training set)




(c) Confusion Matrix obtained on Isot (test set) by DeBERTa trained on NDF (training set)




(d) Confusion Matrix obtained on Horne (test set) by DeBERTa trained on Politifact (training set)

Figure 11: Examples of confusion matrices using DeBERTa.

considering only the portion not used for the training phase. Table 16 reports the number of instances for the two classes used in the test phase.

Table 16: Distribution of the instances in the subsets of the datasets used in the test phase.

| **Dataset** | **Fake class** | **Real class** |
|---|---|---|
| Celebrity [72] | 150 | 150 |
| CIDII [86] | 200 | 322 |
| FakeVsSatire [79] | 183 | 103 |
| Fakes [85] | 278 | 326 |
| Horne [76] | 23 | 103 |
| Infodemic [5] | 4933 | 5426 |
| Isot [28] | 22755 | 21316 |
| LIAR-PLUS [74] | 5556 | 7030 |
| NDF [73] | 116 | 238 |
| Politifact [10] | 83 | 221 |

We also created another test set, denoted as *global test set*, by taking equal portions of instances (20 for the fake class and 20 for the real class) from all instances of each dataset not used in the training phase. We also tested all the models with this new test set.

Tables 17–20 show the P, R, and F1 achieved by each model on each dataset for each class. Table 21 shows the F1 achieved by the models.

Table 17: P, R, and F1 obtained by the ML models on each dataset for each class.

| **Dataset** | **LR** | | | | | | **SVM** | | | | | | **NB** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Celebrity | .57 | .61 | .59 | .58 | .54 | .56 | .68 | .59 | .63 | .64 | .73 | .68 | .52 | .74 | .61 | .56 | .34 | .42 |
| Cidii | .87 | .87 | .87 | .79 | .80 | .80 | .88 | .77 | .82 | .69 | .83 | .75 | .77 | .84 | .80 | .70 | .6 | .64 |
| FakeVsSatire | .59 | .68 | .63 | .80 | .73 | .77 | .61 | .61 | .61 | .78 | .78 | .78 | .46 | .81 | .59 | .82 | .47 | .60 |
| Fakes | .53 | .50 | .51 | .45 | .48 | .46 | .56 | .55 | .56 | .48 | .49 | .49 | .53 | .59 | .56 | .45 | .39 | .42 |
| Horne | .86 | .77 | .81 | .30 | .43 | .35 | .86 | .66 | .75 | .26 | .52 | .34 | .84 | .86 | .85 | .33 | .30 | .31 |
| Infodemic | .74 | .81 | .78 | .77 | .70 | .73 | .76 | .79 | .78 | .76 | .73 | .74 | .65 | .74 | .69 | .66 | .55 | .60 |
| Isot | .62 | .83 | .71 | .77 | .53 | .63 | .68 | .84 | .75 | .81 | .64 | .71 | .52 | .83 | .64 | .64 | .28 | .39 |
| LIAR-PLUS | .57 | .61 | .59 | .46 | .43 | .45 | .57 | .68 | .62 | .48 | .36 | .41 | .57 | .62 | .59 | .46 | .41 | .43 |
| NDF | .86 | .86 | .86 | .72 | .71 | .71 | .85 | .85 | .85 | .70 | .70 | .70 | .80 | .70 | .75 | .51 | .65 | .57 |
| Politifact | .8 | .79 | .79 | .46 | .46 | .46 | .87 | .77 | .82 | .53 | .69 | .60 | .78 | .84 | .81 | .49 | .39 | .44 |
| Global test | .59 | .71 | .64 | .63 | .50 | .56 | .59 | .61 | .60 | .60 | .59 | .59 | .56 | .75 | .64 | .62 | .41 | .49 |

Table 18: P, R, and F1 obtained by the CNN and BiLSTM models on each dataset for each class.

| **Dataset** | **CNN** | | | | | | **BiLSTM** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Celebrity | .52 | .77 | .62 | .57 | .30 | .40 | .49 | .82 | .61 | .47 | .16 | .23 |
| Cidii | .64 | .85 | .73 | .51 | .25 | .33 | .65 | .87 | .75 | .57 | .26 | .36 |
| FakeVsSatire | .37 | .87 | .52 | .73 | .19 | .31 | .34 | .77 | .47 | .56 | .16 | .25 |
| Fakes | .54 | .89 | .67 | .50 | .12 | .20 | .55 | .66 | .60 | .48 | .36 | .41 |
| Horne | .82 | .73 | .77 | .20 | .30 | .24 | .79 | .82 | .80 | .05 | .04 | .04 |
| Infodemic | .52 | .70 | .59 | .47 | .29 | .36 | .54 | .78 | .64 | .52 | .26 | .35 |
| Isot | .50 | .84 | .63 | .61 | .23 | .33 | .48 | .84 | .61 | .54 | .17 | .26 |
| LIAR-PLUS | .57 | .77 | .65 | .47 | .26 | .34 | .56 | .80 | .66 | .47 | .22 | .30 |
| NDF | .69 | .78 | .73 | .39 | .29 | .33 | .67 | .72 | .70 | .33 | .28 | .30 |
| Politifact | .72 | .81 | .76 | .24 | .15 | .18 | .72 | .84 | .77 | .27 | .15 | .19 |
| Global test | .49 | .79 | .60 | .46 | .18 | .26 | .51 | .80 | .62 | .54 | .23 | .32 |

Table 19: P, R, and F1 obtained by the BERT, CNN-BERT and DeBERTa models on each dataset for each class.

| **Dataset** | **BERT** | | | | | | **CNN-BERT** | | | | | | **DeBERTa** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Celebrity | .85 | .66 | .74 | .72 | .88 | .79 | .52 | .69 | .59 | .54 | .36 | .43 | .83 | .83 | .83 | .83 | .83 | .83 |
| Cidii | .98 | .84 | .90 | .79 | .97 | .87 | .62 | .61 | .62 | .39 | .40 | .39 | .98 | .92 | .95 | .89 | .97 | .93 |
| FakeVsSatire | .50 | .80 | .61 | .83 | .54 | .66 | .36 | .65 | .46 | .65 | .36 | .46 | .68 | .75 | .71 | .85 | .80 | .82 |
| Fakes | .54 | .73 | .62 | .47 | .28 | .35 | .55 | .69 | .61 | .49 | .34 | .40 | .62 | .17 | .26 | .47 | .88 | .62 |
| Horne | .92 | .70 | .80 | .36 | .73 | .48 | .83 | .76 | .79 | .22 | .30 | .25 | .92 | .91 | .92 | .62 | .65 | .64 |
| Infodemic | .89 | .85 | .87 | .84 | .89 | .86 | .57 | .55 | .56 | .53 | .55 | .54 | .88 | .91 | .90 | .90 | .87 | .88 |
| Isot | .90 | .99 | .95 | .99 | .90 | .94 | .55 | .69 | .61 | .62 | .48 | .54 | .96 | 1.0 | .98 | 1.0 | .96 | .98 |
| LIAR-PLUS | .64 | .51 | .56 | .50 | .63 | .56 | .55 | .46 | .50 | .43 | .51 | .47 | .66 | .53 | .59 | .52 | .65 | .58 |
| NDF | .93 | .74 | .83 | .63 | .89 | .74 | .67 | .59 | .63 | .33 | .42 | .37 | .94 | .89 | .91 | .79 | .89 | .84 |
| Politifact | .90 | .86 | .88 | .68 | .75 | .72 | .76 | .73 | .74 | .35 | .38 | .36 | .87 | .94 | .90 | .78 | .61 | .69 |
| Global test | .68 | .71 | .69 | .69 | .67 | .68 | .51 | .59 | .55 | .52 | .45 | .48 | .73 | .72 | .73 | .72 | .73 | .73 |

Table 20: P, R, and F1 obtained by the LLaMa2 and GPT-4o models on each dataset for each class.

| **Dataset** | **LLaMa2** | | | | | | **GPT-4o** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Class 0** | | | **Class 1** | | | **Class 0** | | | **Class 1** | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Celebrity | .50 | .98 | .66 | .50 | .02 | .04 | .76 | .94 | .84 | .92 | .71 | .80 |
| Cidii | .61 | .97 | .75 | .17 | .01 | .02 | .95 | .96 | .95 | .93 | .92 | .92 |
| FakeVsSatire | .36 | .99 | .53 | .75 | .02 | .03 | .50 | .01 | .02 | .64 | .99 | .78 |
| Fakes | .54 | 1.0 | .70 | .67 | .01 | .01 | .54 | .97 | .69 | .38 | .02 | .04 |
| Horne | .82 | .99 | .89 | .00 | .00 | .00 | .97 | .64 | .77 | .36 | .91 | .52 |
| Infodemic | .52 | .97 | .68 | .47 | .03 | .06 | .77 | .99 | .87 | .98 | .68 | .81 |
| Isot | .48 | .98 | .65 | .29 | .01 | .01 | .66 | 1.0 | .79 | 1.0 | .52 | .68 |
| LIAR-PLUS | .56 | .98 | .71 | .46 | .02 | .05 | .70 | .69 | .69 | .61 | .63 | .62 |
| NDF | .67 | .97 | .79 | .30 | .03 | .05 | .86 | .97 | .91 | .91 | .68 | .78 |
| Politifact | .70 | .12 | .20 | .27 | .87 | .41 | .83 | .94 | .88 | .76 | .49 | .60 |
| Global test | .50 | .90 | .64 | .47 | .10 | .16 | .64 | .72 | .68 | .68 | .60 | .64 |

The performance of models varies significantly across the datasets, highlighting again the challenge of generalizing fake news detection across different domains. However, we can observe a trend similar to the one in the first experiment, with lower scores and some exceptions. LR and SVM showed an increase in performance when tested on Horne. NB, CNN, and BiLSTM perform better when trained on the global training set and tested on Horne, compared to the scores of the first experiment. CNN, BiLSTM, and CNN-BERT also obtained a slight boost in performance when tested on Cidii. In the case of BERT, when tested on Celebrity, the F1 increased from 0.76 to 0.77.

In general, DeBERTa is the top performer, immediately followed by BERT and GPT-4o. LR and SVM consistently show good performance, often rivaling or surpassing NB on several datasets. CNN and BiLSTM models exhibit fluctuating performances and are under-performing compared to BERT, LR, and SVM in most cases. This behavior confirms the problems observed in the first experiment, highlighting that these models require larger training sets, architecture adjustments, and further fine-tuning to determine their optimal parameters. Even CNN-BERT present fluctuating performances, which align with those obtained by CNN.

The global test set obtained relatively low F1. This and the fluctuating results obtained during the second experiment confirm that some datasets are easier for certain models, making generalization challenging when treating fake news detection as a binary classification task.

Table 21: F1 for each model trained on the global training set (except for LLaMa2 and GPT-4o) and tested on the remaining instances of the datasets not used in the global training set, and on the global test set (best F1 in bold).

| Dataset | LR | SVM | NB | CNN | BiLSTM | BERT | DeBERTa | CNN-BERT | LLaMa2 | GPT-4o |
|---|---|---|---|---|---|---|---|---|---|---|
| Celebrity | .58 | .66 | .54 | .54 | .49 | .77 | **.83** | .53 | .50 | .82 |
| Cidii | .85 | .80 | .75 | .62 | .64 | .89 | **.94** | .53 | .60 | **.94** |
| FakeVsSatire | .72 | .72 | .60 | .44 | .38 | .69 | **.78** | .36 | .37 | .64 |
| Fakes | .49 | .53 | .51 | **.54** | .53 | .52 | .50 | **.54** | **.54** | .53 |
| Horne | .71 | .64 | .76 | .66 | .68 | .71 | **.87** | .68 | .81 | .69 |
| Infodemic | .76 | .77 | .66 | .51 | .54 | .87 | **.89** | .55 | .52 | .84 |
| Isot | .68 | .74 | .55 | .53 | .50 | .95 | **.98** | .59 | .48 | .75 |
| LIAR-PLUS | .53 | .55 | .53 | .55 | .55 | .57 | .58 | .49 | .56 | **.66** |
| NDF | .82 | .81 | .69 | .62 | .58 | .80 | **.89** | .54 | .66 | .87 |
| Politifact | .71 | .75 | .72 | .63 | .65 | .84 | **.85** | .64 | .32 | .82 |
| Global test | .61 | .60 | .58 | .49 | .52 | .69 | **.73** | .52 | .49 | .66 |

### *5.4. Processing Time*

Table 22 shows the training time (in seconds) for each method on each dataset, including vectorization. It is important to note that for each dataset, the *training* variant corresponds to models trained exclusively on the training set of the respective dataset, as defined in the Dataset-Specific experiments discussed in Section 5.1), whereas the *entire* variant represents models trained on the entire dataset, i.e., utilizing all available data points of the dataset for training as discussed in Section 5.2. The last row summarizes the results for models trained on the global training set as explained in Section 5.3.

As expected, the more complex models require longer training times. In particular, we observe that the most performing model DeBERTa requires the longest training times.

Table 22: Training time (in seconds) for each method for the different learning sets used in the three experiments.

| Dataset | LR | SVM | NB | CNN | BiLSTM | BERT | DeBERTa | CNN-BERT |
|---|---|---|---|---|---|---|---|---|
| Celebrity (training) | 7.63 | 5.91 | 5.53 | 192.02 | 196.97 | 189.07 | 237.05 | 50.01 |
| Celebrity (entire) | 7.11 | 8.51 | 5.46 | 194.41 | 197.93 | 210.51 | 290.84 | 42.62 |
| Cidii (training) | 6.04 | 5.50 | 6.30 | 212.69 | 241.61 | 247.42 | 406.38 | 98.67 |
| Cidii (entire) | 5.81 | 5.59 | 6.44 | 220.64 | 227.22 | 372.50 | 532.82 | 120.75 |
| FakeVsSatire (training) | 5.85 | 6.19 | 4.73 | 197.89 | 320.77 | 112.99 | 226.94 | 165.54 |
| FakeVsSatire (entire) | 6.01 | 6.43 | 5.38 | 199.13 | 362.28 | 159.71 | 287.80 | 223.96 |
| Fakes (training) | 7.53 | 6.10 | 5.48 | 192.00 | 298.61 | 167.53 | 465.00 | 139.27 |
| Fakes (entire) | 7.54 | 8.72 | 6.23 | 195.44 | 344.57 | 229.62 | 598.94 | 164.26 |
| Horne (training) | 5.86 | 6.12 | 5.18 | 189.63 | 304.99 | 125.53 | 158.07 | 54.14 |
| Horne (entire) | 5.57 | 7.83 | 5.39 | 189.23 | 336.09 | 167.11 | 195.21 | 60.40 |
| Infodemic (training) | 7.22 | 6.99 | 6.01 | 211.85 | 322.86 | 2976.10 | 3959.74 | 153.17 |
| Infodemic (entire) | 7.58 | 9.68 | 6.49 | 228.51 | 401.97 | 4975.64 | 5548.36 | 233.33 |
| Isot (training) | 30.67 | 31.20 | 25.82 | 280.17 | 1476.91 | 15001.79 | 19485.40 | 770.47 |
| Isot (entire) | 40.58 | 47.25 | 29.45 | 313.20 | 2042.75 | 20772.43 | 24366.53 | 977.61 |
| LIAR-PLUS (training) | 6.59 | 7.07 | 5.95 | 465.26 | 710.13 | 4485.00 | 5994.13 | 190.46 |
| LIAR-PLUS (entire) | 8.37 | 10.39 | 8.02 | 536.26 | 835.91 | 5342.10 | 6687.75 | 241.62 |
| NDF (training) | 5.41 | 5.03 | 4.50 | 183.68 | 214.27 | 107.22 | 438.73 | 79.97 |
| NDF (entire) | 6.55 | 6.57 | 5.41 | 214.70 | 219.96 | 130.25 | 391.41 | 97.51 |
| Politifact (training) | 6.74 | 6.15 | 5.08 | 208.39 | 353.23 | 387.96 | 291.04 | 47.67 |
| Politifact (entire) | 7.34 | 6.55 | 6.17 | 212.52 | 390.18 | 150.15 | 365.89 | 47.75 |
| *Global training set* | 6.68 | 9.02 | 8.34 | 193.21 | 290.87 | 859.98 | 1093.43 | 112.82 |

## 6. Related Work

Significant research efforts have been dedicated to fake news detection by researchers worldwide over time. In recent years, we have witnessed a surge of innovative approaches, which have shown promising results in detecting fake news, as already discussed in Section 2.

A number of surveys have been published to comprehensively examine the approaches and existing datasets. These surveys, summarized in Table 23, provide a comprehensive overview of the research landscape, highlighting the reliance on ML and DL techniques in most approaches. Very few surveys also include LMs in their review, and often, they limit themselves to citing BERT, focusing on more traditional approaches. Some of these surveys also include an empirical analysis similar to the one we have performed, but less complete and detailed.

We have compared ten different models with ten datasets in three different experiments. Furthermore, to the best of our knowledge, while some efforts

have been made to apply LLMs and prompting in fake news detection, as discussed in Section 2.3, and although some researchers have explored more variations of the approaches we used, no one has conducted a comparison as comprehensive as ours, particularly in terms of the number and variety of datasets involved and the types of experiments.

| Ref | ML | DL | (L)LMs | Datasets review | Empirical Analysis |
|---|---|---|---|---|---|
| Shu et al., 2017 [10] | ✓ | ✓ | x | ✓ | x |
| Kumar et al., 2018 [91] | ✓ | x | x | ✓ | x |
| Oshikawa et al., 2018 [92] | ✓ | ✓ | x | ✓ | x |
| Parikh et al., 2018 [93] | ✓ | x | x | ✓ | x |
| Stahl et al., 2018 [94] | ✓ | ✓ | x | x | x |
| Zubiaga et al., 2018 [6] | ✓ | ✓ | x | ✓ | x |
| Ahmed et al., 2018 [29] | ✓ | x | x | ✓ | ✓ |
| Agarwal et al., 2019 [36] | ✓ | x | x | ✓ | ✓ |
| Bondielli et al., 2019 [19] | ✓ | ✓ | x | ✓ | x |
| Dutta et al., 2019 [77] | ✓ | ✓ | x | ✓ | ✓ |
| Elhadad et al., 2019 [95] | ✓ | x | x | ✓ | x |
| Habib et al., 2019 [96] | ✓ | ✓ | x | x | x |
| Katsaros et al., 2019 [75] | ✓ | ✓ | x | ✓ | ✓ |
| Pierri et al., 2019 [97] | ✓ | ✓ | x | ✓ | x |
| Sharma et al., 2019 [98] | ✓ | ✓ | x | ✓ | x |
| Meel et al., 2020 [99] | ✓ | ✓ | x | ✓ | x |
| Zhang et al., 2020 [100] | ✓ | ✓ | x | ✓ | x |
| Zhou et al., 2020 [101] | ✓ | ✓ | x | ✓ | x |
| Collins et al., 2021 [102] | ✓ | ✓ | x | x | x |
| D'Ulizia et al., 2021 [13] | x | x | x | ✓ | x |
| Kumar et al., 2021 [103] | ✓ | ✓ | ✓ | x | x |
| Mridha et al., 2021 [104] | ✓ | ✓ | ✓ | x | x |
| Varma et al., 2021 [105] | ✓ | ✓ | ✓ | x | x |
| Jiang et al., 2021 [31] | ✓ | ✓ | x | ✓ | ✓ |
| Lahby et al., 2022 [106] | ✓ | ✓ | x | x | x |
| Li et al., 2022 [107] | x | ✓ | x | x | x |
| Rohera et al., 2022 [39] | ✓ | ✓ | x | ✓ | ✓ |
| Alghamdi et al., 2022 [89] | ✓ | ✓ | ✓ | x | ✓ |
| Kondamudi et al., 2023 [23] | ✓ | ✓ | ✓ | ✓ | x |
| Chen et al., 2023 [50] | x | ✓ | x | x | ✓ |
| Rastogi et al., 2023 [108] | ✓ | ✓ | ✓ | ✓ | x |
| Alghamdi et al., 2023 [109] | ✓ | ✓ | ✓ | x | ✓ |
| Liu et al., 2024 [110] | ✓ | ✓ | ✓ | ✓ | x |
| Farhangian et al., 2024 [111] | ✓ | ✓ | ✓ | ✓ | ✓ |
| Our paper | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 23: A comparison of existing surveys on fake news detection.

One of the earliest surveys on fake news detection was conducted by Shu et al., 2017 [10]. They comprehensively review methods for detecting fake news on social media platforms, focusing on traditional ML approaches and DL techniques. Additionally, they incorporate a fake news characterization based on psychology and social theories, evaluation metrics, and a list of four

fake news detection datasets. Various aspects of false information, including the actors involved in its dissemination, are discussed in Kumar et al., 2018 [91]. Similarly, Parikh et al., 2018 [93], concentrate solely on traditional ML approaches in their analysis of detection methods.

Authors in Oshikawa et al., 2018 [92], systematically review and compare task formulations, datasets, and NLP solutions developed to address the fake news detection problem. They also highlight the distinction between fake news detection and other related tasks. The research presented in Stahl et al., 2018 [94], analyzes the origins and mechanisms behind the existence of fake news. The authors show a dissertation on Linguistic Cue and Network Analysis approaches and propose a three-part method utilizing NB Classifiers, SVM, and Semantic Analysis for detecting fake news on social media.

Zubiaga et al., 2018 [6], provide a characterization of fake news and an overview of the detection methods and resolution techniques, focusing on rumors. They also review six datasets for rumor detection.

In their survey, Bondielli et al., 2019 [19], discuss the various definitions of fake news and rumors in the literature. They highlight the challenges in collecting relevant data for performing fake news and rumors detection, presenting multiple approaches and several publicly available datasets. Additionally, they report a comprehensive analysis of the various techniques used for rumor and fake news detection, although they do not conduct an empirical study.

Several publications provide systematic surveys on the process of fake news detection on social media [95, 97]. Additionally, a taxonomy focusing on the negative impact of online fake news was provided by Meel et al., 2020 [99]. In recent years, several other literature surveys have also been presented [23, 96, 98, 100, 103, 106, 107]. One of the most comprehensive works is the survey conducted by Rastogi et al., 2023 [108].

Another interesting study aims to review and evaluate fake news detection methods from four perspectives: the false knowledge conveyed by the fake news, its writing style, its propagation patterns, and the credibility of its source [101]. Instead, Collins et al., 2021 [102], focus on hybrid-ML techniques, whereas Mridha et al., 2021 [104], provide a comprehensive overview of DL-based methods. Finally, Varma et al., 2021 [105], systematically review fake news detection techniques, distinguishing between those developed before and after the Covid-19 pandemic.

One of the challenges in the fake news detection task is the need for

datasets. A survey systematically reviews twenty-seven popular datasets for fake news detection [13]. The authors provide insights into the characteristics of each dataset and conduct a comparative analysis among them.

In general, a few works also include an empirical study of one or more approaches. We summarized these empirical studies in Table 24.

| **Ref** | **ML models** | **DL models** | **(L)LMs** | **Datasets** | **Cross-dataset analysis** |
|---|---|---|---|---|---|
| Ahmed et al., 2018 [29] | 6 | x | x | 1 | x |
| Agarwal et al., 2019 [36] | 5 | x | x | 1 | x |
| Dutta et al., 2019 [77] | 2 | x | x | 1 | x |
| Katsaros et al., 2019 [75] | 6 | 2 | x | 3 | x |
| Jiang et al., 2021 [31] | 5 | 3 | x | 2 | x |
| Rohera et al., 2022 [39] | 3 | 1 | x | 1 | x |
| Alghamdi et al., 2022 [89] | 7 | 8 | 2 | 4 | x |
| Chen et al., 2023 [50] | x | 3 | x | 1 | x |
| Alghamdi et al., 2023 [109] | 5 | 5 | 10 | 1 | x |
| Farhangian et al., 2024 [111] | 8 | 3 | 9 | 4 | x |
| **Our paper** | **3** | **3** | 4 | **10** | ✓ |

Table 24: Existing surveys on fake news detection that include an experimental analysis of the datasets. We report the papers and the number of tested models and used datasets.

In Agarwal et al., 2019 [36], the authors conduct an analysis by training five classifiers on the LIAR dataset to evaluate their performance on this dataset of labeled news statements. They employ NB, LR, SVM, Stochastic Gradient Descent, and Random Forest (RF) classifiers. Similarly, Dutta et al., 2019 [77], perform an experimental analysis of two traditional approaches, namely NB and LR, using a dataset named "fake_or_real_news" available on Kaggle. In Katsaros et al., 2019 [75], a comprehensive performance evaluation of eight ML algorithms for fake news detection and classification is presented. This evaluation includes both traditional ML and DL approaches. Regarding the datasets, they utilize the Liar dataset, the Signal Media News Articles Dataset, and a Kaggle dataset named "Getting Real about Fake News" (the BS Detector dataset). On the other hand, Rohera et al., 2022 [39], conduct a comparison between several traditional ML models such as Passive Aggressive Algorithm, RF, and NB, along with a DL model, and an LSTM. They utilize "fake_or_real_news" as a unique dataset. One of the most complete analysis has been conducted in Jiang et al., 2021 [31]. Although it is not a survey in strict sense, authors made a large literature review. They evaluated the performances of five ML models and three DL models on two datasets, the

Isot and KDnugget datasets, and proposed their model.

More recently, the fake news problem has also been investigated in the context of emotion-based studies. A survey systematically reviewing these approaches is presented in [110], which contains a list of available datasets for several languages. In Alghamdi et al., 2022 [89], Alghamdi et al., 2023 [109], and Farhangian et al., 2024 [111], authors have conducted extensive comparisons of various approaches to fake news detection. These studies explored diverse ML models, incorporating different features and augmenting Transformer-based architectures (e.g., BERT) with layers like CNNs, BiLSTMs, and GRUs. While studies provide valuable insights into model performance on specific datasets, our research differentiates itself by focusing on cross-dataset and cross-domain generalization. By evaluating models across a diverse range of datasets varying in size and topic, we assessed models robustness and adaptability to new, unseen data.

## 7. Conclusions

In this paper, we have considered fake news detection as a binary classification task. First, we have presented an overview of current state-of-the-art approaches and available datasets. Then, we focused on ten approaches, which are representative of the different families proposed in the literature, and ten datasets. Finally, we performed three different experiments to evaluate the generalization ability of these approaches.

Our experiments highlighted the effectiveness of the fine-tuned Transformers (BERT and DeBERTa) which outperformed the other approaches. Notably, GPT-4o, even in a zero-shot setting, demonstrated surprisingly promising results, highlighting the potential of LLMs in this task. However, our cross-dataset evaluation has also revealed a critical challenge: all models, excluding GPT-4o, exhibited low and variable performance in generalization across different datasets. This finding underscores a fundamental issue with traditional approaches, which assume specific features can consistently distinguish between fake and real news regardless of the context. Our results suggest that this assumption is not entirely valid, as models trained on one dataset often struggle to maintain their performance when applied to different datasets. It is worth highlighting the need to explore novel paradigms in approaching fake news detection. Rather than relying solely on generic features and isolated news, the focus should be shifted to a more realistic

scenario where the context in which a news appears is considered a crucial factor.

Our future research will focus on context-aware methodologies, specifically analyzing the information divergence between a fake news, the factual events from which it originates, and the surrounding contextual elements. We expect that a similar approach could overcome the limitations on the generalization of the current detection systems. We believe that this approach is essential for developing more effective and robust fake news detection systems that can better adapt to the rapidly evolving dynamics of information dissemination. Moreover, it is widely acknowledged that communication is inherently multi-modal. While we have initiated an exploration of the interplay between textual and visual inputs in the context of the Italian language [112, 113], a thorough and comparative study on this topic remains an open avenue for our future research. Indeed, despite the complexities of modelling multimodal data and the current limitations of Vision-LMs, a cross-dataset evaluation incorporating both textual and visual information could yield valuable insights and contribute to the development of more robust and comprehensive fake news detection systems.

This study has limitations that present opportunities for future research. Expanding the scope of our analysis to include a more diverse range of datasets and methodologies could significantly enhance the generalizability of our findings. Additionally, a more detailed examination of new architectures, coupled with the exploration of alternative adaptation methods, could provide deeper insights into their respective strengths and limitations.

In conclusion, while our experiments have reaffirmed the efficacy of fine-tuned models and the potential of GPT-4o, they have also highlighted the limitations of current approaches in generalizing across diverse datasets. We believe it may be helpful to explore beyond traditional feature-based methods. By incorporating contextual information, we can develop more robust and adaptable systems better equipped to tackle the complex and ever-changing nature of fake news.

**Acknowledgements**

This work has been partly funded by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" under the NextGeneration EU programme, and the Italian Ministry of University and Research (MUR)

in the framework of the PRIN 2022JLB83Z "Psychologically-tailored approaches to Debunk Fake News detected automatically by an innovative artificial intelligence approach", the FoReLab and CrossLab projects (Departments of Excellence).

## References

[1] N. Newman, W. Dutton, G. Blank, Social media in the changing ecology of news: The fourth and fifth estates in britain, International journal of internet science 7 (1) (2013).

[2] S. Vieweg, Microblogged contributions to the emergency arena: Discovery, interpretation and implications, Computer Supported Collaborative Work (2010) 515–516.

[3] S. Lewandowsky, U. K. Ecker, J. Cook, Beyond misinformation: Understanding and coping with the "post-truth" era, Journal of applied research in memory and cognition 6 (4) (2017) 353–369.

[4] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, arXiv preprint arXiv:2103.12541 (2021).

[5] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1, Springer, 2021, pp. 21–29.

[6] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, Detection and resolution of rumours in social media: A survey, ACM Computing Surveys (CSUR) 51 (2) (2018) 1–36.

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

[8] A. Barrón-Cedeno, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan,

A. Nikolov, et al., 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 11, Springer, 2020, pp. 215–236.

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

[10] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD explorations newsletter 19 (1) (2017) 22–36.

[11] G. Pennycook, D. G. Rand, The psychology of fake news, Trends in Cognitive Sciences 25 (5) (2021) 388–402. doi:https://doi.org/10.1016/j.tics.2021.02.007.
URL https://www.sciencedirect.com/science/article/pii/S1364661321000516

[12] J. Su, C. Cardie, P. Nakov, Adapting fake news detection to the era of large language models, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1473–1490. doi:10.18653/v1/2024.findings-naacl.95.
URL https://aclanthology.org/2024.findings-naacl.95

[13] A. D'Ulizia, M. C. Caschera, F. Ferri, P. Grifoni, Fake news detection: a survey of evaluation datasets, PeerJ Computer Science 7 (2021) e518.

[14] R. K. Kaliyar, A. Goswami, P. Narang, Fakebert: Fake news detection in social media with a bert-based deep learning approach, Multimedia tools and applications 80 (8) (2021) 11765–11788.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[16] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[18] OpenAI, Gpt-4 technical report (2023). arXiv:2303.08774.

[19] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Information Sciences 497 (2019) 38–55.

[20] V. L. Rubin, Y. Chen, N. K. Conroy, Deception detection for news: three types of fakes, Proceedings of the Association for Information Science and Technology 52 (1) (2015) 1–4.

[21] N. DiFonzo, P. Bordia, Rumor, gossip and urban legends, Diogenes 54 (1) (2007) 19–35.

[22] C. Silverman, Lies, damn lies and viral content (2015).

[23] M. R. Kondamudi, S. R. Sahoo, L. Chouhan, N. Yadav, A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches, Journal of King Saud University-Computer and Information Sciences 35 (6) (2023) 101571.

[24] J. A. Adeyiga, P. G. Toriola, T. E. Abioye, A. E. Oluwatosin, et al., Fake news detection using a logistic regression model and natural language processing techniques (2023).

[25] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, L. De Alfaro, Some like it hoax: Automated fake news detection in social networks, arXiv preprint arXiv:1704.07506 (2017).

[26] J. M. G. Ogdol, B.-L. T. Samar, C. Catarroja, Binary logistic regression based classifier for fake news, J. High. Educ. Res. Discip (2018).

[27] M. D. Vicario, W. Quattrociocchi, A. Scala, F. Zollo, Polarization and fake news: Early warning of potential misinformation targets, ACM Transactions on the Web (TWEB) 13 (2) (2019) 1–22.

[28] H. Ahmed, I. Traore, S. Saad, Detection of online fake news using n-gram analysis and machine learning techniques, in: Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments:

First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1, Springer, 2017, pp. 127–138.

[29] H. Ahmed, I. Traore, S. Saad, Detecting opinion spams and fake news using text classification, Security and Privacy 1 (1) (2018) e9.

[30] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, L. De Alfaro, Automatic online fake news detection combining content and social signals, in: 2018 22nd conference of open innovations association (FRUCT), IEEE, 2018, pp. 272–279.

[31] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, A. Ali, A novel stacking approach for accurate detection of fake news, IEEE Access 9 (2021) 22626–22639.

[32] F. Mohsen, B. Chaushi, H. Abdelhaq, D. Karastoyanova, K. Wang, Automated detection of misinformation: A hybrid approach for fake news detection, Future Internet 16 (10) (2024) 352.

[33] S. R. Sain, The nature of statistical learning theory (1996).

[34] G. Sansonetti, F. Gasparetti, G. D'aniello, A. Micarelli, Unreliable users detection in social media: Deep learning techniques for automatic detection, IEEE Access 8 (2020) 213154–213167.

[35] V. L. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the second workshop on computational approaches to deception detection, 2016, pp. 7–17.

[36] V. Agarwal, H. P. Sultana, S. Malhotra, A. Sarkar, Analysis of classifiers for fake news detection, Procedia Computer Science 165 (2019) 377–383.

[37] M. Granik, V. Mesyura, Fake news detection using naive bayes classifier, in: 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON), IEEE, 2017, pp. 900–903.

[38] A. Abdulrahman, M. Baykara, Fake news detection using machine learning and deep learning algorithms, in: 2020 International Conference on Advanced Science and Engineering (ICOASE), IEEE, 2020, pp. 18–23.

[39] D. Rohera, H. Shethna, K. Patel, U. Thakker, S. Tanwar, R. Gupta, W.-C. Hong, R. Sharma, A taxonomy of fake news classification techniques: Survey and implementation aspects, IEEE Access 10 (2022) 30367–30394.

[40] G. E. R. Agudelo, O. J. S. Parra, J. B. Velandia, Raising a model for fake news detection using machine learning in python, in: Challenges and Opportunities in the Digital Era: 17th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2018, Kuwait City, Kuwait, October 30–November 1, 2018, Proceedings 17, Springer, 2018, pp. 596–604.

[41] E. M. Mahir, S. Akhter, M. R. Huq, et al., Detecting fake news using machine learning and deep learning algorithms, in: 2019 7th international conference on smart computing & communications (ICSCC), IEEE, 2019, pp. 1–5.

[42] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks (2016).

[43] N. Ruchansky, S. Seo, Y. Liu, Csi: A hybrid deep model for fake news detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 797–806.

[44] Q. Liao, H. Chai, H. Han, X. Zhang, X. Wang, W. Xia, Y. Ding, An integrated multi-task model for fake news detection, IEEE Transactions on Knowledge and Data Engineering 34 (11) (2021) 5154–5165.

[45] R. K. Kaliyar, A. Goswami, P. Narang, S. Sinha, Fndnet–a deep convolutional neural network for fake news detection, Cognitive Systems Research 61 (2020) 32–44.

[46] M. Bugueño, G. Sepulveda, M. Mendoza, An empirical analysis of rumor detection on microblogs with recurrent neural networks, in: Social Computing and Social Media. Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part I 21, Springer, 2019, pp. 293–310.

[47] P. Bahad, P. Saxena, R. Kamal, Fake news detection using bi-directional lstm-recurrent neural network, Procedia Computer Science 165 (2019) 74–82.

[48] M. Z. Asghar, A. Habib, A. Habib, A. Khan, R. Ali, A. Khattak, Exploring deep neural networks for rumor detection, Journal of Ambient Intelligence and Humanized Computing 12 (2021) 4315–4333.

[49] S. Girgis, E. Amer, M. Gadallah, Deep learning algorithms for detecting fake news in online text, in: 2018 13th international conference on computer engineering and systems (ICCES), IEEE, 2018, pp. 93–97.

[50] M.-Y. Chen, Y.-W. Lai, J.-W. Lian, Using deep learning models to detect fake news about covid-19, ACM Transactions on Internet Technology 23 (2) (2023) 1–23.

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

[52] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, K. Bontcheva, T. Cohn, I. Augenstein, Discourse-aware rumour stance classification in social media using sequential classifiers, Information Processing & Management 54 (2) (2018) 273–290.

[53] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, defend: Explainable fake news detection, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 395–405.

[54] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[55] A. Roy, K. Basak, A. Ekbal, P. Bhattacharyya, A deep ensemble framework for fake news detection and classification, arXiv preprint arXiv:1811.04670 (2018).

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

[57] V. Slovikovskaya, Transfer learning from transformers to fake news challenge stance detection (fnc-1) task, arXiv preprint arXiv:1910.14353 (2019).

[58] M. Qazi, M. U. Khan, M. Ali, Detection of fake news using transformer model, in: 2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET), IEEE, 2020, pp. 1–6.

[59] S. Kula, M. Choraś, R. Kozik, Application of the bert-based architecture in fake news detection, in: 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12, Springer, 2021, pp. 239–249.

[60] H. Jwa, D. Oh, K. Park, J. M. Kang, H. Lim, exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert), Applied Sciences 9 (19) (2019) 4062.

[61] J. Alghamdi, Y. Lin, S. Luo, The power of context: A novel hybrid context-aware fake news detection approach, Information 15 (3) (2024) 122.

[62] J. Alghamdi, Y. Lin, S. Luo, Enhancing hierarchical attention networks with cnn and stylistic features for fake news detection, Expert Systems with Applications 257 (2024) 125024.

[63] P. Goel, S. Singhal, S. Aggarwal, M. Jain, Multi domain fake news analysis using transfer learning, in: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2021, pp. 1230–1237.

[64] S. Shifath, M. F. Khan, M. S. Islam, A transformer based approach for fighting covid-19 fake news, arXiv preprint arXiv:2101.12027 (2021).

[65] H. Ramakrishna, I. Lekshmi Ammal, A. K. Madasamy, Nitk-it_nlp at checkthat!-2022: Window based approach for fake news detection using transformers., in: CLEF (Working Notes), 2022, pp. 649–655.

[66] C. Whitehouse, T. Weyde, P. Madhyastha, N. Komninos, Evaluation of fake news detection with knowledge-enhanced language models, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 16, 2022, pp. 1425–1429.

[67] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[68] B. Jiang, Z. Tan, A. Nirmal, H. Liu, Disinformation detection: An evolving challenge in the age of llms, in: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM), SIAM, 2024, pp. 427–435.

[69] Y. Sun, J. He, L. Cui, S. Lei, C.-T. Lu, Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges, arXiv preprint arXiv:2403.18249 (2024).

[70] J. Su, T. Y. Zhuo, J. Mansurov, D. Wang, P. Nakov, Fake news detectors are biased against texts generated by large language models, arXiv preprint arXiv:2309.08674 (2023).

[71] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, Bad actor, good advisor: Exploring the role of large language models in fake news detection, arXiv preprint arXiv:2309.12247 (2023).

[72] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, arXiv preprint arXiv:1708.07104 (2017).

[73] L. C. Passaro, A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, In-context annotation of topic-oriented datasets of fake news: A case study on the notre-dame fire event, Information Sciences 615 (2022) 657–677.

[74] T. Alhindi, S. Petridis, S. Muresan, Where is your evidence: Improving fact-checking by justification modeling, in: Proceedings of the first workshop on fact extraction and verification (FEVER), 2018, pp. 85–90.

[75] D. Katsaros, G. Stavropoulos, D. Papakostas, Which machine learning paradigm for fake news detection?, in: IEEE/WIC/ACM International Conference on Web Intelligence, 2019, pp. 383–387.

[76] B. Horne, S. Adali, This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, in: Proceedings of the international AAAI conference on web and social media, Vol. 11, 2017, pp. 759–766.

[77] P. S. Dutta, M. Das, S. Biswas, M. Bora, S. S. Saikia, Fake news prediction: a survey, International Journal of Scientific Engineering and Science 3 (3) (2019) 1–3.

[78] B. Riedel, I. Augenstein, G. P. Spithourakis, S. Riedel, A simple but tough-to-beat baseline for the fake news challenge stance detection task, arXiv preprint arXiv:1707.03264 (2017).

[79] J. Golbeck, M. Mauriello, B. Auxier, K. H. Bhanushali, C. Bonk, M. A. Bouzaghrane, C. Buntain, R. Chanduka, P. Cheakalos, J. B. Everett, et al., Fake news vs satire: A dataset and analysis, in: Proceedings of the 10th ACM conference on web science, 2018, pp. 17–21.

[80] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, Advances in neural information processing systems 32 (2019).

[81] T. Mitra, E. Gilbert, Credbank: A large-scale social media corpus with associated credibility annotations, in: Proceedings of the international AAAI conference on web and social media, Vol. 9, 2015, pp. 258–267.

[82] K. Anoop, P. Deepak, V. Lajish, Emotion cognizance improves fake news identification, arXiv preprint arXiv:1906.10365 1 (2019).

[83] X. Li, Y. Xia, X. Long, Z. Li, S. Li, Exploring text-transformers in aaai 2021 shared task: Covid-19 fake news detection in english, arXiv preprint arXiv:2101.02359 (2021).

[84] G. K. Shahi, D. Nandini, Fakecovid–a multilingual cross-domain fact check news dataset for covid-19, arXiv preprint arXiv:2006.11343 (2020).

[85] F. K. A. Salem, R. Al Feel, S. Elbassuoni, M. Jaber, M. Farah, Fakes: A fake news dataset around the syrian war, in: Proceedings of the international AAAI conference on web and social media, Vol. 13, 2019, pp. 573–582.

[86] S. K. Hamed, M. J. Ab Aziz, M. R. Yaakub, Disinformation detection about islamic issues on social media using deep learning techniques, Malaysian Journal of Computer Science 36 (3) (2023) 242–270.

[87] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A stylometric inquiry into hyperpartisan and fake news, arXiv preprint arXiv:1702.05638 (2017).

[88] A. Affelt, How to spot fake news, in: All That's Not Fit to Print, Emerald Publishing Limited, 2019, pp. 57–84.

[89] J. Alghamdi, Y. Lin, S. Luo, A comparative study of machine learning and deep learning techniques for fake news detection, Information 13 (12) (2022) 576.

[90] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, W. Z. Khan, An ensemble machine learning approach through effective feature extraction to classify fake news, Future Generation Computer Systems 117 (2021) 47–58.

[91] S. Kumar, N. Shah, False information on web and social media: A survey, arXiv preprint arXiv:1804.08559 (2018).

[92] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, arXiv preprint arXiv:1811.00770 (2018).

[93] S. B. Parikh, P. K. Atrey, Media-rich fake news detection: A survey, in: 2018 IEEE conference on multimedia information processing and retrieval (MIPR), IEEE, 2018, pp. 436–441.

[94] K. Stahl, Fake news detection in social media, California State University Stanislaus 6 (2018) 4–15.

[95] M. K. Elhadad, K. F. Li, F. Gebali, Fake news detection on social media: a systematic survey, in: 2019 IEEE Pacific Rim conference on communications, computers and signal processing (PACRIM), IEEE, 2019, pp. 1–8.

[96] A. Habib, M. Z. Asghar, A. Khan, A. Habib, A. Khan, False information detection in online content and its role in decision making: a systematic literature review, Social Network Analysis and Mining 9 (2019) 1–20.

[97] F. Pierri, S. Ceri, False news on social media: a data-driven survey, ACM Sigmod Record 48 (2) (2019) 18–27.

[98] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, Combating fake news: A survey on identification and mitigation techniques, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (3) (2019) 1–42.

[99] P. Meel, D. K. Vishwakarma, 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 153 (2020) 112986.

[100] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management 57 (2) (2020) 102025.

[101] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys (CSUR) 53 (5) (2020) 1–40.

[102] B. Collins, D. T. Hoang, N. T. Nguyen, D. Hwang, Trends in combating fake news on social media–a survey, Journal of Information and Telecommunication 5 (2) (2021) 247–266.

[103] S. Kumar, S. Kumar, P. Yadav, M. Bagri, A survey on analysis of fake news detection techniques, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, 2021, pp. 894–899.

[104] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, M. S. Rahman, A comprehensive review on fake news detection with deep learning, IEEE Access 9 (2021) 156151–156170.

[105] R. Varma, Y. Verma, P. Vijayvargiya, P. P. Churi, A systematic survey on deep learning and machine learning approaches of fake news detection in the pre-and post-covid-19 pandemic, International Journal of Intelligent Computing and Cybernetics 14 (4) (2021) 617–646.

[106] M. Lahby, S. Aqil, W. M. Yafooz, Y. Abakarim, Online fake news detection using machine learning techniques: A systematic mapping

study, Combating Fake News with Computational Intelligence Techniques (2022) 3–37.

[107] J. Li, M. Lei, A brief survey for fake news detection via deep learning models, Procedia Computer Science 214 (2022) 1339–1344.

[108] S. Rastogi, D. Bansal, A review on fake news detection 3t's: Typology, time of detection, taxonomies, International Journal of Information Security 22 (1) (2023) 177–212.

[109] J. Alghamdi, Y. Lin, S. Luo, Towards covid-19 fake news detection using transformer-based models, Knowledge-Based Systems 274 (2023) 110642.

[110] Z. Liu, T. Zhang, K. Yang, P. Thompson, Z. Yu, S. Ananiadou, Emotion detection for misinformation: A review, Information Fusion (2024) 102300.

[111] F. Farhangian, R. M. Cruz, G. D. Cavalcanti, Fake news detection: Taxonomy and comparative study, Information Fusion 103 (2024) 102140.

[112] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at EVALITA 2023: Overview of the multimodal fake news detection and verification task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, Vol. 3473 of CEUR Workshop Proceedings, CEUR-WS.org, 2023.
URL https://ceur-ws.org/Vol-3473/paper32.pdf

[113] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. Passaro, Dataset for multimodal fake news detection and verification tasks, Data in Brief 54 (2024) 110440. doi:https://doi.org/10.1016/j.dib.2024.110440.
URL https://www.sciencedirect.com/science/article/pii/S2352340924004098