Leite et al. *EPJ Data Science* (2025) 14:16
https://doi.org/10.1140/epjds/s13688-025-00534-0

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Weakly supervised veracity classification with LLM-predicted credibility signals

João A. Leite[1*], Olesya Razuvayevskaya[1], Kalina Bontcheva[1] and Carolina Scarton[1]

*Correspondence:
jaleite1@sheffield.ac.uk
[1]Department of Computer Science, The University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom

Abstract

Credibility signals represent a wide range of heuristics typically used by journalists and fact-checkers to assess the veracity of online content. Automating the extraction of credibility signals presents significant challenges due to the necessity of training high-accuracy, signal-specific extractors, coupled with the lack of sufficiently large annotated datasets. This paper introduces PASTEL (**P**rompted we**A**k **S**upervision wi**T**h cr**E**dibility signa**L**s), a weakly supervised approach that leverages large language models (LLMs) to extract credibility signals from web content, and subsequently combines them to predict the veracity of content without relying on human supervision. We validate our approach using four article-level misinformation detection datasets, demonstrating that PASTEL outperforms zero-shot veracity detection by 38.3% and achieves 86.7% of the performance of the state-of-the-art system trained with human supervision. Moreover, in cross-domain settings where training and testing datasets originate from different domains, PASTEL significantly outperforms the state-of-the-art supervised model by 63%. We further study the association between credibility signals and veracity, and perform an ablation study showing the impact of each signal on model performance. Our findings reveal that 12 out of the 19 proposed signals exhibit strong associations with veracity across all datasets, while some signals show domain-specific strengths.

Keywords: Veracity classification; Large language models; Weak supervision; Credibility signals

1 Introduction

In the era of rapidly spreading misinformation, the task of automatic veracity classification of online content has emerged as a prominent field of research [1]. Despite significant progress, several limitations and challenges persist. State-of-the-art methods typically rely on supervised learning, and thus require high-quality, manually annotated datasets. The creation of such datasets is time-consuming, and the evolving nature of misinformation necessitates the continuous development of new datasets [2–4]. Additionally, supervised methods often struggle to generalise across different misinformation domains (e.g., politics and celebrity gossip) due to the domain-specific nature of the statistic patterns present in the training data, resulting in considerable decrease in performance if in-domain data is unavailable [5, 6].

Springer

To address these challenges, prior work has used weakly supervised methods, which rely on indirect signals to assess the truthfulness of content, rather than needing large amounts of annotated data. These methods make use of available information that isn't explicitly labelled, helping to classify content as true or false without the need for manual annotations. Current methods use weak signals as a combination of simple syntactic features (e.g., count of words) and user engagement with the misinformation content (e.g., number of shares) [7–9]. User engagement signals, such as shares, likes, and comments, are often readily available and can reflect patterns of virality or audience interaction, making them appealing for detecting misinformation. However, such signals are inherently platform-dependent and, more importantly, require the content to be disseminated and interacted with before detection. By the time sufficient engagement data is available, the misinformation narrative has often already caused significant harm. These limitations highlight the need for approaches like PASTEL (presented in this paper), which rely solely on the textual content of articles to assess veracity, enabling platform-independent, early detection of misinformation. In spite of the simplicity of the aforementioned signals, the challenge of integrating more sophisticated information (e.g., credibility signals[1] defined by experts) poses a paradox: complex signals demand specialised models and annotated datasets for accurate extraction [11], which undermines the premise of employing weak supervision in the first place.

Pretrained large language models (LLMs) offer promising opportunities to address the aforementioned challenges. While further research is necessary to fully understand their potential and limitations, LLMs have demonstrated remarkable zero-shot performance in various NLP tasks, including common sense reasoning, reading comprehension, and closed-book question answering [12], at times even surpassing state-of-the-art supervised approaches [13]. LLMs exhibit strong recall of factual knowledge without fine-tuning [14], suggesting that the external knowledge acquired during pretraining could be harnessed to extract complex signals from textual content without requiring further fine-tuning with annotated datasets.

Our contribution with this work is the proposal of Prompted weAk superviSion wiTh crEdibility signaLs (PASTEL), an approach modelled on the verification process typically adopted by journalists and fact-checkers, who assess the veracity of online content using a wide range of credibility signals. We leverage the task-agnostic capabilities of LLMs to extract nineteen sophisticated credibility signals from news articles in a zero-shot setting (i.e., without training the model with ground truth labels). These signals are then aggregated into a binary (*misinformation/non-misinformation*) veracity label using weak supervision.

Our comprehensive experiments demonstrate that PASTEL outperforms zero-shot veracity classification by 38.3%, and attains 86.7% of the performance of the state-of-the-art supervised model, which relies on domain-specific training data. Moreover, PASTEL outperforms the state-of-the-art supervised model by 63% in cross-domain settings, underscoring its applicability to real-world scenarios where misinformation rapidly evolves and domain-specific training data is limited. Lastly, we investigate the role of each credibility signal in predicting content veracity by inspecting their statistical association with the

[1] See the report by W3C-CWCG [10] for an overview of credibility signals.

human-annotated veracity labels, and through an ablation study in which PASTEL's performance is measured after individual signals are removed. Our analysis provides valuable insights highlighting the importance of domain-specific credibility signals, and how a diverse range of credibility signals is key in enhancing the model's performance.

The remainder of this paper is structured as follows: Sect. 2 presents an overview of relevant previous work. Section 3 describes our proposed method. The experimental setup is presented in Sect. 4, whilst results are discussed in Sect. 5. In Sect. 6, we analyse and discuss the predicted credibility signals. Section 7 presents a discussion of implications of this work, points to future work, and makes concluding remarks. We make our code and data fully available [15].

2 Related work

2.1 Article-level veracity classification

Building models to automatically assess content veracity generally relies on human-annotated datasets. Most benchmark corpora focus on short claims [16–19] or social media data such as Facebook posts [20–22], tweets [23–25], and Reddit threads [26]. However, article-level veracity assessment relies on more context and nuance, making annotation more challenging and less scalable, therefore fewer datasets are available [5, 27–29]. This section describes four article-level datasets commonly employed in works studying automatic veracity detection and cross-domain generalisation. We also present the key classification approaches used.

Pérez-Rosas et al. [5] introduced two datasets: FakeNewsAMT and Celebrity, annotated with binary veracity labels. FakeNewsAMT contained political news articles from six topics with deceptive versions created by crowdsourced workers. The Celebrity dataset included web articles about celebrities verified against gossip-checking sites. Both datasets achieved annotation agreement scores of 70% and 73%, respectively. Also, the authors performed a cross-domain analysis by training their best model with one of the datasets and testing on the other. Results showed a drop in performance of 13.5% for the FakeNewsAMT dataset, and 34.2% for the Celebrity dataset. Studies on these datasets used models such as SVMs with word embeddings, grammatical features, and word-level attention with multi-layer perceptrons [30, 31]. Transfer learning models such as RoBERTa, GPT-2, XLNet, DeBERTa, and BERT surpassed feature-based methods, with the best reported $F1_{macro}$ scores of 0.99 for FakeNewsAMT and 0.82 for Celebrity using RoBERTa. However, they struggled in cross-domain settings, dropping 40% in performance [6].

Shu et al. [27] presented PolitiFact and GossipCop, two binary article-level datasets. PolitiFact included politically themed articles assessed by journalists, while GossipCop focused on celebrity stories verified by a rating system. Previous methods evaluated on these datasets include CNNs, knowledge-aware attention networks, and convolutional Tselin Machines [27, 32, 33]. The current state-of-the-art results were achieved by Rai et al. [34], who fine-tuned BERT model, achieving $F1_{macro}$ scores of 0.88 for PolitiFact and 0.89 for GossipCop. They experimented with an LSTM layer on top of BERT, which slightly improved performance by 0.02 for PolitiFact but did not affect GossipCop.

2.2 Credibility signals

The term *credibility signals* refers to a wide range of measurable heuristics that collectively help journalists assess the overall trustworthiness of information. Examples of credibility signals include the analysis of article titles [35], writing style [36], rhetorical structure

[37, 38], linguistic features [39], emotional language [40], biases [41], and logical fallacies and inferences [42]. Additionally, credibility signals comprise meta-information that extends beyond the textual content of the article, such as the author's reputation and external references [43].

The W3C Credible Web Community Group (CWCG) [10] performed the most extensive attempt to date at cataloguing credibility signals by defining and documenting hundreds of signals. Dimou et al. [11] selected 23 credibility signals defined by the W3C CWCG and built a modular evaluation pipeline for the task of predicting the credibility of content. Their signals were a mixture of (i) simple syntatic features (e.g., word length, word count, exclamation marks), (ii) metadata (e.g., author name, URL domains), and (iii) a smaller set of complex features extracted by specialised classifiers trained for each of them (e.g., sentiment, clickbait). These signals were grouped into 10 modules, and each module was manually assigned an importance weight that defined its contribution to the overall credibility of the web page. The authors found that morphological, syntactic, and emotional features demonstrated the highest predictive capability for determining the credibility of web content.

To the best of our knowledge, the only dataset annotated with several credibility signals was introduced by Zhang et al. [44]. They employed six trained annotators to label articles with 17 different content indicators and 11 context indicators based on the W3C CWCG definitions. However, their dataset was a feasibility study with a small sample size of only 40 annotated articles, which severely limits its utility for training supervised machine learning models.

2.3 Veracity classification with weak supervision

Programmatic weak supervision (PWS) is a semi-supervised learning paradigm that encodes noisy probabilistic labels using multiple *labeling functions* that are correlated with the objective task [45–48]. Several prior works applied weak supervision techniques to detect the veracity of online content. A common theme among these works was the use of social media metadata, syntactic features, and user interactions with misinformation content as weak signals.

Shu et al. [7] incorporated multiple weak signals from user engagements with content. Their weak signals included (i) *sentiment*, which considered the average sentiment scores inferred from users sharing a given news piece; (ii) *bias*, which was modelled by inspecting how closely the user's interests matched those of people with known public biases; and (iii) *credibility*, which considered the size of the cluster containing the user. This was modelled on the hypothesis that low-credibility users were likely to coordinate and form large clusters, while high-credibility users tended to form small clusters. Their best classifier trained exclusively with weak signals was a RoBERTa model that achieved an average $F1_{macro}$ score of 0.535 across two datasets.

Helmstetter and Paulheim [8] applied weak supervision for misinformation detection on Twitter. They used five sets of features as weak signals: (i) a total of 53 *user-level* features, such as the frequency of tweets, ratio of retweets, number of followers, etc.; (ii) a total of 69 *tweet-level* features, such as word count and the ratio of question and exclamation marks; (iii) *text-level* features comprising TF-IDF encoded vectors representing the tweet text; (iv) *topic-level* features consisting of automatically derived topics using LDA; and (v) *sentiment-level* features representing the ratio of positive, negative, and neutral words in the text. Their best configuration used an XGBoost classifier trained with the proposed

features, achieving an $F1_{macro}$ of 0.77 for detecting a set of misinformation tweets labelled by themselves.

Wang et al. [9] proposed *WeFEND*, a reinforced weakly-supervised fake news detection framework. Their approach leveraged user feedback on known misinformation articles as weak signals. They trained a classifier using these signals and applied it to predict misinformation in articles with unknown veracity, but for which user feedback was available. They evaluated their approach on a dataset of news articles published by WeChat official accounts, along with the corresponding user feedback. Their model achieved an F1-score of 0.880 for misinformation articles and 0.810 for non-misinformation articles.

In conclusion, our approach differs from previous works in two key aspects. First, PASTEL does not rely on any metadata related to user engagement with the misinformation article, but operates exclusively on the textual content of the article. This distinction is crucial because models that depend on engagement features require that the content is spread and interacted with before the model can accurately detect it, by which time the misinformation narrative has already caused harm. Additionally, PASTEL leverages signals defined by specialists from the W3C Credible Web Community Group (CWCG), which encompass more sophisticated concepts (e.g., whether the content presents evidence) compared to user engagement statistics (e.g., number of shares) or syntactic features (e.g., word count) used in previous works. To annotate these complex signals without relying on annotated data, we employ LLMs to predict the weak signals in a zero-shot setting (i.e., without any fine-tuning with annotated data).

3 Prompted weAk Supervision wiTh crEdibility signaLs (PASTEL)

PASTEL draws inspiration from the verification practices employed by journalists and fact-checkers, such as assessing the presence of evidence, identifying bias, and evaluating the credibility of sources, as outlined by frameworks like the W3C Credible Web Community Group [10]. These practices involve using a diverse array of credibility indicators, including linguistic features, citation patterns, and emotional language, to determine the truthfulness of online content. Our method harnesses the task-agnostic abilities of large language models (LLMs) to identify nineteen nuanced credibility signals from news articles in a zero-shot setting, meaning the model operates without training with ground truth labels. Subsequently, we integrate these signals to perform a binary classification (*misinformation* or *non-misinformation*) through a process of weak supervision. Figure 1 provides an overview of the approach, illustrating it with three examples of credibility signals. In the following sections, each component is described in greater detail.

3.1 Credibility signals considered

We leverage nineteen credibility signals, all of which have been shown to be relevant for assessing content veracity. Table 1 displays these signals, that provide a solid foundation of well-defined and validated indicators of content credibility. Note that all our signals are formulated so that their presence in the content indicates a lack of credibility.

The vast majority of the signals used in our experiments were proposed by the W3C (the Web Standards Organisation) Credible Web Community Group [10], who defined numerous credibility indicators to help users and machines identify trustworthy content, i.e. content that is reliable, accurate, and shared in good faith (see Table 1). The aim of our work is not to propose new credibility signals but to use those already established by subject matter experts and demonstrate how they can enhance automatic veracity classification.

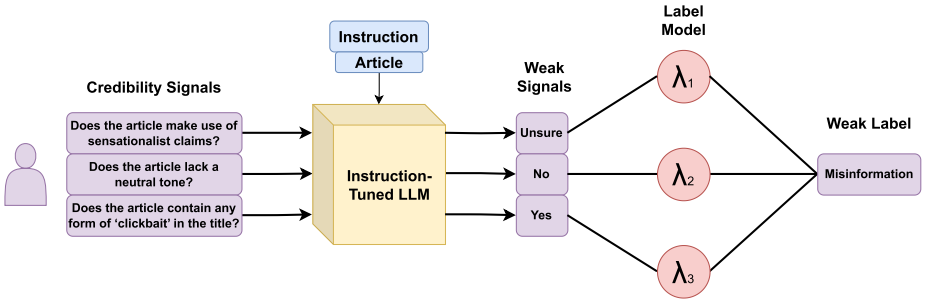

Figure 1 Illustration of PASTEL

Table 1 Credibility signals and their respective definitions

Credibility Signal	Definition
Evidence*	Fails to present any supporting evidence or arguments to substantiate its claims.
Bias‡	Contains explicit or implicit biases (e.g. confirmation bias, selection bias, framing bias).
Inference#	Makes claims about correlation and causation.
Polarising Language#	Uses polarising terms or makes divisions into sharply contrasting groups or sets of opinions or beliefs.
Document Citation#	Lacks citations of studies or documents to support its claims.
Informal Tone#	Uses all caps or consecutive exclamation or question marks.
Explicitly Unverified Claims†	Contains claims that are explicitly lack confirmation.
Personal Perspective†	Includes the author's own personal opinions about the subject.
Emotional Valence†	Language carries emotional valence that is predominantly negative or positive rather than neutral.
Call to Action†	Contains language that can be understood as a call to action, requesting readers to follow through with a particular task or telling readers what to do.
Expert Citation†	Lacks citations of experts in the subject.
Clickbait†	Title contains sensationalised or misleading headlines in order to attract clicks.
Incorrect Spelling†	Contains significant misspellings and/or grammatical errors.
Misleading About Content†	Title emphasises different information than the body topic.
Incivility†	Uses stereotypes and/or generalisations of groups of people.
Impoliteness†	Contains insults, name-calling, or profanity.
Sensationalism†	Presents information in a manner designed to evoke strong emotional reactions.
Source Credibility†	Cites low-credibility sources.
Reported by Other Sources†	Presents a story that was not reported by other reputable media outlets.

*Musi and Reed [42]
‡Dufraisse et al. [41]
#Zhang et al. [44]
†W3C-CWCG [10]

3.2 Signal extraction (LLM prompting)

Instruction-tuned LLMs operate in a question-answering manner through the use of prompts. A prompt is a specific query given to the model to instruct it to perform a task. With carefully crafted prompts, the LLM's capabilities can be harnessed to extract the credibility signals. Fig. 2 displays the prompt template employed to extract a single credibility signal in a question-answering approach using an instruction-tuned LLM.

The prompt uses the Alpaca template [49], and contains 3 distinct sections: 'Instruction', 'Input', and 'Response'. The 'Instruction' section of the prompt guides the model towards

```
### Instruction:
You are a helpful and unbiased news verification assistant. You will be provided with
the title and the full body of text of a news article. Then, you will answer further
questions related to the given article. Ensure that your answers are grounded in
reality, truthful and reliable. You are expected to answer with 'Yes' or 'No', but you
are also allowed to answer with 'Unsure' if you do not have enough information or
context to provide a reliable answer.

### Input:
{title}
{text}

{question} (Yes/Unsure/No)

### Response:
```

Figure 2 Pastel's prompt template

extracting each credibility signal in an unbiased, grounded in reality, truthful, and reliable manner, and also ensures that the model only outputs valid answers (*Yes*, *No*, and *Unsure*). The 'Input' section is filled with the title and body of text of the input news article, followed by a question associated to a credibility signal. This is essentially a mapping from the definition of the respective signal to a question. For example, the definition for the *Inference* signal (see Table 1) is mapped to the following question: "Does this article make claims about correlation and causation?". Immediately following the question, we explicitly state the three candidate answers (i.e., *Yes*, *Unsure*, and *No*) to reinforce that the model should only output these answers. Finally, the 'Response' section is left blank to allow the LLM to perform text completion. Note that this template allows for the extraction of one credibility signal at a time. Therefore, for each input news article, nineteen prompts are created, each with a different question corresponding to a distinct credibility signal. These prompts are fed sequentially to the LLM, with no additional context carried over from previous interactions.

3.3 Weak supervision

After extracting the credibility signals, our objective is to combine the signals into binary veracity labels (*misinformation* or *non-misinformation*). The simplest approach is to apply a majority voting heuristic: if the majority of signals are triggered, the outcome is classified as *misinformation*; otherwise, it is classified as *non-misinformation*. However, this approach has limitations, as all signals are treated equally, whereas ideally, signals with higher accuracy should influence the outcome more than those with lower accuracy. Moreover, signals can be highly correlated, leading to duplicated or nearly identical outputs (i.e., double voting) which can bias the final prediction.

To address these challenges, we employ weak supervision to determine signal weights without relying on annotated data. Instead, weights are estimated from empirical statistics derived from their distribution. Our goal is to train a parameterised classification model, denoted as h_θ, where, for a given news article $x \in X$, the model predicts its veracity label $y \in Y$ (where $Y \in \{0, 1\}$). In a supervised learning setting, h_θ is trained on a dataset comprising pairs of inputs and ground truth labels, denoted as (x_{train}, y_{train}). However, in weakly supervised learning, we lack access to y_{train}. Instead, we generate training labels using a set of labeling functions $\lambda : X \rightarrow Y \cup \{-1\}$, where '$-1$' indicates abstention (in our setting, the 'Unsure' class).

Each labeling function λ is expected to exhibit some correlation with Y, although they may be noisy, meaning they do not necessarily provide highly accurate predictions for Y individually. Assuming we have m inputs and n labeling functions, Λ_{ij} represents the output of labeling function λ_j for input x_i, resulting in a matrix as follows:

$$\Lambda = \begin{bmatrix} \lambda_0(x_0) & \dots & \lambda_{n-1}(x_0) \\ \vdots & \ddots & \vdots \\ \lambda_0(x_{m-1}) & \dots & \lambda_{n-1}(x_{m-1}) \end{bmatrix}_{m \times n} \tag{1}$$

Next, the goal is to transform Λ into a vector of probabilistic weak labels $\tilde{Y} = (\tilde{y}_0, \dots, \tilde{y}_{m-1})$, with $\tilde{y}_i \in [0, 1]$. To do so, we train a generative model $p_\theta(\Lambda, Y)$ to obtain weights θ_j that calibrate the contribution of λ_j towards $\tilde{Y}$. Specifically, we use the approach by Ratner et al. [50], which defines factor types representing the labeling propensity, and pairwise correlations between labeling functions j and k for the ith input:

$$\begin{aligned} \phi^{Lab}_{i,j}(\Lambda, Y) &= \mathbb{1}\{\Lambda_{i,j} \neq -1\} \\ \phi^{Corr}_{i,j,k}(\Lambda, Y) &= \mathbb{1}\{\Lambda_{i,j} = \Lambda_{i,k}\} \end{aligned} \tag{2}$$

The factor types are concatenated into a single vector ϕ_i for each input x_i, and the parameters of the model are defined as $w \in \mathbb{R}^{2n+|C|}$, where C is a set of potentially correlated pairs of labeling functions. The label model is defined by Equation (3), where Z_w is a normalising constant:

$$p_w(\Lambda, Y) = Z_w^{-1} \exp\left(\sum_{i=0}^{m-1} w^T \phi_i(\Lambda, y_i)\right) \tag{3}$$

The model learns without access to the ground truth labels Y, thus the objective is to minimise the negative log marginal likelihood given the observed outputs of the labeling functions Λ:

$$\hat{w} = \arg\min_w - \log \sum_Y p_w(\Lambda, Y) \tag{4}$$

The trained label model is then used to infer the probabilistic weak labels $\tilde{Y} = p_{\hat{w}}(Y|\Lambda)$, and the discrete predictions (*misinformation*/*non-misinformation*) are obtained by taking the *argmax* of each weak label $\tilde{y}_i \in \tilde{Y}$.

4 Experimental setup

In this section we describe the datasets, metrics, models, and techniques employed to assess the performance of our method in comparison to other strong baselines. The classification setting is the following: given the title and body of a news article, predict it's veracity as either *misinformation* or *non-misinformation*. Initially, we assess the models' performance within the same domain, where both the train and test sets are derived from the same dataset. Subsequently, we evaluate the models' cross-domain performance, where the train and test sets originate from different datasets.

Table 2 Datasets used throughout the experiments along with their label distributions and average number of tokens

Dataset	# Misinformation	# Non-misinformation	# Tokens (avg.)
PolitiFact	308 (44.6%)	383 (55.4%)	2605.2
GossipCop	3924 (22.4%)	13,596 (77.6%)	981.3
FakeNewsAMT	240 (50%)	240 (50%)	178.4
Celebrity	250 (50%)	250 (50%)	635.5

4.1 Datasets

We experiment with four English article-level misinformation datasets: PolitiFact and GossipCop by Shu et al. [51], and FakeNewsAMT and Celebrity by Pérez-Rosas et al. [5]. These datasets are chosen because they cover two distinct domains: GossipCop and Celebrity focus on entertainment news, whereas PolitiFact and FakeNewsAMT concentrate on politics. This distinction allows us to assess the model's ability to generalise beyond its training data domain. Furthermore, these datasets exhibit unique characteristics that may impact model performance and are therefore crucial for evaluation.

- GossipCop's classes are considerably imbalanced towards the negative class (77.6%). Other datasets are near to perfectly balanced.
- Gossipcop has more than 10 times the number of articles than the combination of the three other datasets.
- PolitiFact's average document length is considerably larger than other datasets, with 2605.2 average tokens per article. Contrastingly, FakeNewsAMT has only 178.4 average tokens per article, which is notably fewer than others.

Although PolitiFact and GossipCop contain additional social-context data in the form of tweets, we only use content-related attributes (title and body) as input to the models. This choice ensures that the models are evaluated purely on their ability to handle the news content, without the influence of external social-context signals, such as user interactions or engagement patterns. Table 2 presents the class distributions and average number of tokens for each dataset.

4.2 Evaluation

Similar to the previous works that experimented with the four datasets [5, 27, 30–34, 52], we use the $\mathrm{F1}_{\mathrm{macro}}$ score as the main evaluation metric. The $\mathrm{F1}_{\mathrm{macro}}$ score is defined in Equation (5), where N is the number of classes, and TP_i, FP_i, and FN_i correspond to the number of true positives, false positives, and false negatives, respectively, for class i. This metric is particularly suitable for datasets with skewed class distributions, as it returns the average of the $\mathrm{F1}_{\mathrm{macro}}$ scores for each class, and thus does not favour the majority class. We report the mean and standard error of $\mathrm{F1}_{\mathrm{macro}}$ scores using a 10-fold cross-validation strategy. This ensures every sample is used for training and evaluation, providing a robust and generalisable estimate of model performance, while reducing bias from unrepresentative splits in fixed train-test setups.

$$F1_{\mathrm{macro}} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 * TP_i}{2 * TP_i + FP_i + FN_i} \tag{5}$$

4.3 Large language model

We conduct our experiments using LLaMa2, an open-source LLM developed by Meta AI, pretrained on a publicly available dataset of 2 trillion tokens. Given that our framework

relies on the extraction of complex credibility signals in a zero-shot setting, we prioritized selecting the version of the model capable of accurately interpreting prompts and generating reliable outputs without requiring task-specific fine-tuning. Specifically, we employ LLaMa2-Platypus-70B, a variant of LLaMa2 with 70 billion parameters that was fine-tuned using the Open-Platypus dataset [53], which focuses on enhancing the logical reasoning skills of the LLM. LLaMa2-Platypus-70B, which is fully open-source, achieved remarkable performance across several popular LLM benchmark datasets.[2] More specifically, LLaMa2-Platypus-70B achieves competitive results on benchmarks such as ARC (70.65) [54], HellaSwag (87.15) [55], MMLU (70.08) [56], and TruthfulQA (52.37) [57], averaging to a score of 70.06/100 across these corpora.

4.4 Baselines

We compare PASTEL against the state-of-the-art models for the four datasets described in Table 2. Also, we distinguish between supervised and non-supervised (i.e., unsupervised and weakly supervised) baselines to provide a fair assessment of the methods, as supervised models are trained with access to high-quality in-domain annotated data, and thus have a significant methodological advantage over the non-supervised models. Therefore, the supervised baselines serve as an upper bound reference for comparison against the non-supervised methods. The baselines are described in detail below:

Weakly-supervised approach

- *Prompted weAk Supervision wiTh crEdibility signaLs (PASTEL)*: Our method is described in detail in Sect. 3. The LLM extracts nineteen credibility signals for each news article in the dataset. We train Snorkel's label model [50] for 500 epochs using the credibility signals extracted from the training split. Given the lightweight computational demand of training the label model, we experimented with varying the number of epochs, ranging from 100 to 1000, in steps of 50. We did not observe further performance gains past epoch 500. Finally, the signal weights estimated from the training set distribution are applied to aggregate the signals from the test split, producing the final binary (*misinformation*/*non-misinformation*) veracity predictions.

Unsupervised approaches

- *LLaMa Credibility Signals Chain-of-Thought (LLaMa-CS-CoT)*: We employ a Chain-of-Thought [58] approach to extract and aggregate the signals using the same LLM, without employing weak supervision. First, the credibility signals are extracted in the exact same manner as in PASTEL. Next, instead of employing weak supervision to aggregate the signals, we use the same LLM to predict veracity based on the article text, the list of 19 credibility signals, and the LLM answers from the previous step, which indicate the presence or absence of each signal. No fine-tuning is performed. The prompt used for this baseline is shown in Fig. 3.
- *LLaMa Zero-Shot (LLaMa-ZS)*: No credibility signals are extracted. The LLM directly assesses the veracity of articles in the test split without any fine-tuning. The prompt for this baseline is the same as *LLaMa-CS-CoT*, except that only the title and text of the article are provided, without the credibility signals.

[2] See https://huggingface.co/garage-bAInd/Platypus2-70B for more details.

```
### Instruction:
You are a veracity classification model. Your task is to determine whether the article
contains misinformation or not. You will be provided with a list of 19 credibility
signals. Each credibility signal, if present, represents a potential indicator of misin-
formation. However, their presence is not a guarantee of misinformation. You are
expected to consider the credibility signals in aggregate, and the article content, to
make an informed decision.

### Input:
{title}
{text}

Credibility Signals:
{signals}: {signal_predictions}

Does this article contain misinformation? (Yes/No)

### Response:
```

Figure 3 LLaMa-CS-CoT prompt template

Supervised approaches

- *LLaMa Credibility Signals Supervised (LLaMa-CS-S)*: In this approach, credibility signals are first extracted in the same manner as in PASTEL. These signals are then used to train a supervised logistic regression model, which learns to map the 19 extracted signals into binary veracity labels. The model is optimised using default hyperparameter settings provided by the Scikit-Learn Python library.[3]
- *LLaMa Fine-Tuned (LLaMa-FT)*: The LLM is fine-tuned with the causal language modeling objective using articles from the train split alongside their ground truth annotations. We employ LoRA [59] to fine-tune LLaMa2-Platypus, using the same settings as in Lee et al. [53]: a learning rate of 3×10^{-4}, a batch size of 4, and a microbatch size of 1, and the cutoff length is set to 4096 tokens. The training includes 100 warmup steps, spans 1 epoch, and employs no weight decay. The learning rate scheduler is set to cosine. For LoRA settings, we use an alpha value of 16, a rank of 16, and a dropout rate of 0.05. Following fine-tuning, the LLM directly assesses the veracity of articles in the test split, identically to LLaMa-ZS.
- *RoBERTa*: As discussed in Sect. 2, the RoBERTa model by Goel et al. [52] is the state-of-the-art model for both FakeNewsAMT and Celebrity, with $F1_{macro}$ scores of 0.99 and 0.82, respectively. The authors employed a single train and test split of 70% train and 30% test in their experimental setup, thus, we reproduce their model and evaluate it using our more robust methodology with 10-fold cross validation to ensure that the results are comparable with our other proposed baselines. We reproduce their work using the hyperparameters and settings provided in their paper: `RoBERTa-Base` pretrained model, Adam optimizer with β_1 of 0.9 and β_2 of 0.999, learning rate of $2e^{-5}$, weight decay of $1e^{-1}$, batch size of 8, and 5 training epochs. Sentences are truncated to a maximum of 512 tokens.
- *BERT*: The BERT model by Rai et al. [34] is the state-of-the-art model for PolitiFact and GossipCop, with $F1_{macro}$ scores of 0.88 and 0.89, respectively. However, we were not able to reproduce their experiments as they did not specify the hyperparameters

[3] https://scikit-learn.org/.

used to finetune the model, nor did they release their code. Also, they employed a single train and test split evaluation methodology (80% train and 20% test), and we employ a more robust 10-fold cross validation strategy, thus, our experimental setup is not directly comparable to theirs. Therefore, we finetune a `BERT-Base-Uncased` architecture with the default hyperparameters specified in the HuggingFace deep learning framework [60]: Adam optimizer with β_1 of 0.9 and β_2 of 0.999, learning rate of $5e^{-5}$, batch size of 8, and 5 training epochs. Sentences are truncated to a maximum of 512 tokens.

All experiments are conducted using a single NVIDIA A100-80GB GPU. For experiments with LLaMa2 (LLaMa-ZS, LLaMa-FT, LLaMa-CS-S, LLaMa-CS-CoT, and Pastel), we apply 4-bit quantisation [61]. This is a practical measure to reduce the computational and memory requirements of running the LLaMa2-Platypus-70B model in a NVIDIA A100-80GB GPU.

5 Results

5.1 In-domain classification

In the in-domain scenario, models are trained and evaluated with in-domain data, i.e., the train and test sets are derived from the same dataset. Table 3 presents the classification results for the proposed baselines.

First, we compare the supervised baselines: LLaMa-FT, BERT, RoBERTa, and LLaMa-CS-S. We find that both BERT and RoBERTa significantly outperform LLaMa-FT by 0.11 and 0.22 in $F1_{macro}$, respectively, despite LLaMa-FT having a much larger number of parameters (LLaMa2 has 70 billion parameters, while BERT and RoBERTa each have fewer than 150 million parameters). This performance gap may be attributed to the relatively small size of the training data for all datasets ($< 1K$ samples) except GossipCop, as larger models often require larger training sets for optimal performance [62]. For the GossipCop dataset ($17K$ samples), LLaMa-FT outperforms BERT and is only 0.05 behind RoBERTa. Comparing the similarly sized models, BERT and RoBERTa, we find that RoBERTa, on average, outperforms BERT by 0.11 ($\uparrow$ 14.1%) in $F1_{macro}$, despite a statistical overlap (indicated by their standard deviations) in all datasets except FakeNewsAMT. Meanwhile, LLaMa-CS-S demonstrates competitive performance, achieving an average $F1_{macro}$ score of 0.79, which is comparable to BERT and only slightly below RoBERTa. This indicates that supervised aggregation of credibility signals can effectively leverage the information contained in these signals to achieve robust in-domain performance.

Next, we compare the non-supervised baselines (i.e., unsupervised and weakly supervised): LLaMa-ZS, LLaMa-CS-CoT, and Pastel. Pastel consistently outperforms both

Table 3 Classification results ($F1_{macro}$). Highest scores for each setting are in bold. Means and standard deviations obtained with 10-fold cross-validation

Setting	Approach	PolitiFact	GossipCop	FNAMT	Celebrity	Mean
Supervised	BERT	0.89 ± 0.03	0.67 ± 1.6	0.75 ± 0.08	0.79 ± 0.06	0.78
	RoBERTa	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.80 \pm 0.1}$	$\mathbf{0.97 \pm 0.03}$	$\mathbf{0.87 \pm 0.05}$	**0.89**
	LLaMa-FT	0.68 ± 0.02	0.75 ± 0.01	0.79 ± 0.05	0.43 ± 0.03	0.67
	LLaMa-CS-S	0.82 ± 0.04	0.67 ± 0.01	0.84 ± 0.04	0.81 ± 0.07	0.79
Unsupervised	LLaMa-ZS	0.61 ± 0.02	0.55 ± 0.01	0.65 ± 0.02	0.45 ± 0.02	0.57
	LLaMa-CS-CoT	0.72 ± 0.05	0.55 ± 0.01	0.64 ± 0.05	0.58 ± 0.08	0.62
Weakly Supervised	Pastel	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.81 \pm 0.02}$	**0.78**

LLaMa-ZS and LLaMa-CS-CoT, with an average increase of 0.21 in $F1_{macro}$ over LLaMa-ZS and 0.16 over LLaMa-CS-CoT across all four datasets, representing an average increase of 38.3% and 25.8% in performance, respectively. Specifically, PASTEL outperforms LLaMa-ZS by 0.16 (↑ 22.2%), 0.14 (↑ 25.5%), and 0.17 (↑ 26.1%) for PolitiFact, GossipCop, and FakeNewsAMT, respectively, and by similar margins over LLaMa-CS-CoT. The most substantial improvement is observed for the Celebrity dataset, with an increase of 0.36 (↑ 80%) in $F1_{macro}$ compared to LLaMa-ZS, and 0.23 (↑ 65.7%) compared to LLaMa-CS-CoT. We highlight that the results obtained with the LLaMa-ZS baseline are consistent with Hu et al. [63], who used ChatGPT-3.5 to assess veracity for the GossipCop dataset and obtained an $F1_{macro}$ score of 0.57. These results underscore PASTEL's substantial superiority over both zero-shot prompting and chain-of-thought aggregation approaches for veracity assessment.

Finally, we compare PASTEL with RoBERTa, the state-of-the-art supervised model. As discussed in Sect. 4.4, we use the supervised models as upper bound references to PASTEL, as they are trained with access to ground truth labels, while PASTEL is not. Therefore, we compare RoBERTa and PASTEL in terms of PASTEL's ability to approach the scores obtained by the RoBERTa model. We find that PASTEL achieves 86.7% of RoBERTa's performance averaging across the four datasets. Specifically, PASTEL achieves 82.8%, 86.3%, 84.5%, and 93.1% of the performance of the RoBERTa model for PolitiFact, GossipCop, FakeNewsAMT, and Celebrity, respectively.

5.2 Cross-domain classification

Supervised models often experience a decline in performance when there is a mismatch between the training set distribution and the test set distribution, a phenomenon known as domain shift [64]. In this experiment, we evaluate the cross-domain robustness of the state-of-the-art supervised model, RoBERTa, in comparison to PASTEL. For each of the four datasets $i \in D$, both models are trained with i, and evaluated on the three remaining datasets $j \in D \mid j \neq i$. Table 4 presents the cross-dataset $F1_{macro}$ scores for both RoBERTa and PASTEL.

On average, PASTEL achieves a mean $F1_{macro}$ score of 0.75 compared to RoBERTa's 0.46 (an increase of 63%). When evaluated on the PolitiFact, GossipCop, FakeNewsAMT, and Celebrity datasets, PASTEL attains average $F1_{macro}$ scores of 0.75, 0.78, 0.71, and 0.74, respectively. In contrast, RoBERTa achieves lower average $F1_{macro}$ scores of 0.38, 0.57, 0.33, and 0.62 on the corresponding datasets.

Although PASTEL consistently outperforms RoBERTa, the difference is less pronounced for datasets within the same domain, particularly entertainment news. For instance, when RoBERTa is trained on GossipCop and tested on Celebrity, it achieves an $F1_{macro}$ score

Table 4 Cross-dataset $F1_{macro}$ for RoBERTa (RoB) vs. PASTEL (PAS)

		Train							
		PolitiFact		GossipCop		FakeNewsAMT		Celebrity	
		RoB	PAS	RoB	PAS	RoB	PAS	RoB	PAS
Test	PolitiFact	x	x	0.45	**0.69**	0.40	**0.67**	0.65	**0.74**
	GossipCop	0.25	**0.69**	x	x	0.21	**0.67**	**0.69**	**0.69**
	FakeNewsAMT	0.54	**0.76**	0.52	**0.84**	x	x	0.52	**0.78**
	Celebrity	0.34	**0.80**	0.74	**0.81**	0.37	**0.78**	x	x
	Mean	0.38	**0.75**	0.57	**0.78**	0.33	**0.71**	0.62	**0.74**

of 0.74, which is 0.07 lower than PASTEL. When trained on Celebrity and evaluated on GossipCop, both models score 0.69. In the political domain, the performance gap is more significant. Training on PolitiFact and evaluating on FakeNewsAMT results in an $F1_{macro}$ score of 0.54 for RoBERTa, 0.22 lower than PASTEL. Similarly, training on FakeNewsAMT and testing on PolitiFact yields a score of 0.40 for RoBERTa, which is 0.27 below PASTEL.

When the training and testing datasets originate from different domains, the performance difference between the models becomes more substantial. Training on political datasets and evaluating on entertainment datasets poses the most significant challenge for RoBERTa. For instance, when trained on PolitiFact and tested on GossipCop and Celebrity, RoBERTa trails PASTEL by 0.44 and 0.46, respectively. Similar gaps are observed when training on FakeNewsAMT and testing on these datasets, with RoBERTa falling behind by 0.46 on GossipCop and 0.41 on Celebrity. This trend persists when training on entertainment news and testing on political news, albeit with a smaller gap between the two models. Training on GossipCop and testing on PolitiFact and FakeNewsAMT results in gaps of 0.24 and 0.32, respectively. Training on Celebrity and testing on the same two datasets results in gaps of 0.09 and 0.26.

These results underscore the superior robustness of PASTEL to domain shift compared to the supervised state-of-the-art model. This characteristic is crucial for applications where in-domain training data is unavailable, or for dynamically changing domains and emergent topics.

5.3 Error analysis

To gain deeper insights into the performance of our method, we conduct a detailed error analysis to systematically identify the types of errors made by PASTEL. Figure 4 displays the confusion matrices averaged over 10-fold cross-validation for test sets in each dataset.

As each dataset has different sizes and label distributions, we further calculate the False Positive Rate (FPR) and the False Negative Rate (FNR) (see Equation (6)). Table 5 displays the FNR and FPR for each dataset.

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{6}$$

Across all datasets, PASTEL yields a higher rate of false negatives over false positives, with averages of 32.8% and 10.8%, respectively. PASTEL's FNR is notably high for the GossipCop dataset (49.3%), which is possibly a result of its label skewnewss, as the negative class comprises 77.6% of the dataset. Contrastingly, the FNR for the other three datasets is considerably lower, with 32.1%, 28.2%, and 21.5% for FakeNewsAMT, PolitiFact, and Celebrity, respectively.

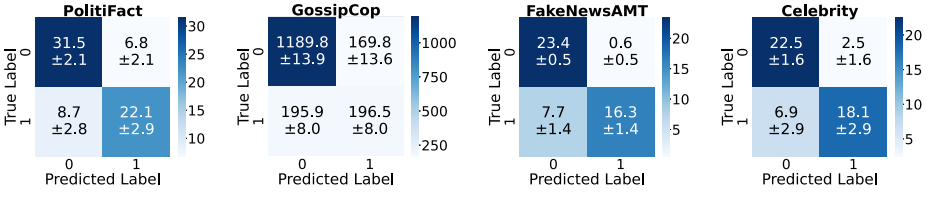

Figure 4 Mean confusion matrices obtained with PASTEL. Means and standard deviations reported across 10-fold cross-validation. Labels 0 and 1 refer to *non-misinformation* and *misinformation*, respectivelly

Table 5 False Negative Rate (FNR) and False Positive Rate (FPR)

Dataset	FNR (%)	FPR (%)
PolitiFact	28.2	17.7
GossipCop	49.3	13.2
Celebrity	21.6	10.0
FakeNewsAMT	32.1	2.4
Mean	32.8	10.8

Table 6 Relative frequency of credibility signals triggered in True Positive (TP) and False Negative (FN) predictions. Percent decrease indicated within parenthesis

Credibility Signal	TPs	FNs
Emotional Valence	0.52	$0.00_{(\downarrow 100\%)}$
Clickbait	0.29	$0.00_{(\downarrow 100\%)}$
Expert Citation	0.55	$0.00_{(\downarrow 100\%)}$
Evidence	0.56	$0.00_{(\downarrow 100\%)}$
Source Credibility	0.25	$0.00_{(\downarrow 100\%)}$
Bias	0.39	$0.01_{(\downarrow 97.4\%)}$
Document Citation	0.73	$0.02_{(\downarrow 97.3\%)}$
Incivility	0.26	$0.01_{(\downarrow 96.2\%)}$
Sensationalism	0.69	$0.03_{(\downarrow 95.7\%)}$
Polarising Language	0.39	$0.02_{(\downarrow 94.9\%)}$
Misleading about content	0.57	$0.05_{(\downarrow 91.2\%)}$
Explicitly Unverified Claims	0.32	$0.03_{(\downarrow 90.6\%)}$
Incorrect Spelling	0.19	$0.02_{(\downarrow 89.5\%)}$
Impoliteness	0.08	$0.01_{(\downarrow 87.5\%)}$
Informal Tone	0.34	$0.08_{(\downarrow 76.5\%)}$
Personal Perspective	0.38	$0.09_{(\downarrow 76.3\%)}$
Reported by Other Sources	0.78	$0.38_{(\downarrow 51.3\%)}$
Call to Action	0.04	$0.02_{(\downarrow 50.0\%)}$
Inference	0.24	$0.14_{(\downarrow 41.7\%)}$
Total	7.57	$0.91_{(\downarrow 88.0\%)}$

In the context of PASTEL's method, false negative errors occur when one or more signals are not triggered. To examine such errors, we compare the distribution of credibility signals in true positive (TP) and false negative (FN) examples. In Table 6, we present the relative frequency (the number of times the credibility signal was triggered, divided by the number of articles) of each credibility signal in TP and FN predictions, averaged across the four datasets.

The statistics indicate that all 19 signals occur less frequently in FN predictions compared to TP. On average, 7.57 credibility signals are triggered in TP predictions, whereas only 0.91 signals are triggered in FN predictions, representing a significant decrease of 88.0%. A reduction of more than 70% in frequency is observed for 16 signals, while *Reported by Other Sources*, *Call to Action*, and *Inference* show smaller decreases of 51.3%, 50.0%, and 41.7%, respectively.

6 Analysis of credibility signals

This section examines the effectiveness of LLM-extracted credibility signals in predicting content veracity through two research questions: (i) Is there a statistical association between credibility signals and the article's veracity? (ii) Which credibility signals contribute the most towards PASTEL's classification performance?

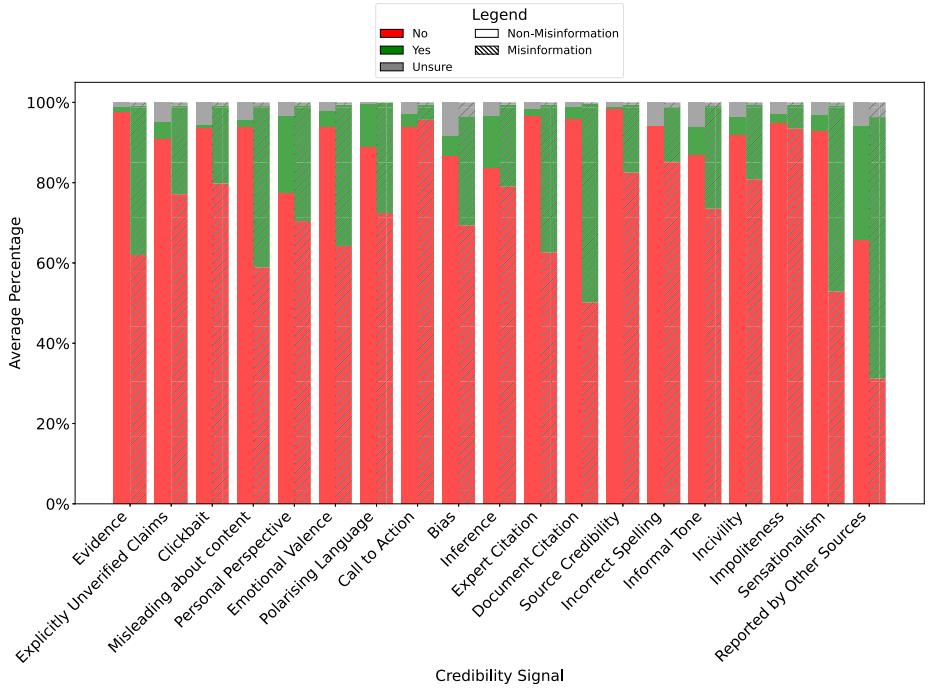

Figure 5 Distribution of LLM responses per credibility signal for *non-misinformation* articles (solid bars) and *misinformation* articles (hashed bars) averaged across all datasets

6.1 Credibility signals and veracity

Figure 5 compares the proportion of LLM responses ('Yes', 'No', or 'Unsure') for each credibility signal in *misinformation* and *non-misinformation* articles.

Firstly, we note that the percentage of 'Unsure' answers is relatively small across all credibility signals, composing less than 10% of the answers. Also, the rate of 'Unsure' answers is higher for *non-misinformation* articles. These statistics may indicate that the model is overconfident, or in other words, is often not capable of identifying when there is not enough information to confidently decide between 'Yes' or 'No'. Nevertheless, all 19 credibility signals are found more frequently in *misinformation* articles than in *non-misinformation* articles.

In order to verify if there is a statistically significant association between the credibility signals and the article's veracity, we perform a Pearson's chi-squared statistical test. Our null hypothesis H_0 is that there is no association between the credibility signals and the veracity of the article. We reject the null hypothesis H_0 if $p < 0.05$. This test is done for each credibility signal independently, and for each dataset separately. Additionally, we analyse the χ^2 statistic as a measure of the strength of association between the credibility signal and the veracity label. A higher χ^2 statistic suggests a significant deviation in the observed distribution of a given signal between *misinformation* and *non-misinformation* articles. For ease of visualisation, the χ^2 statistics are normalised between 0 and 1. Figure 6 illustrates the test outcomes.

When averaging across all datasets (All), we reject H_0 for 12 credibility signals, that therefore have a statistically significant association with the veracity of the articles across all four datasets: *Document Citation*, *Misleading about content*, *Evidence*, *Sensationalism*,

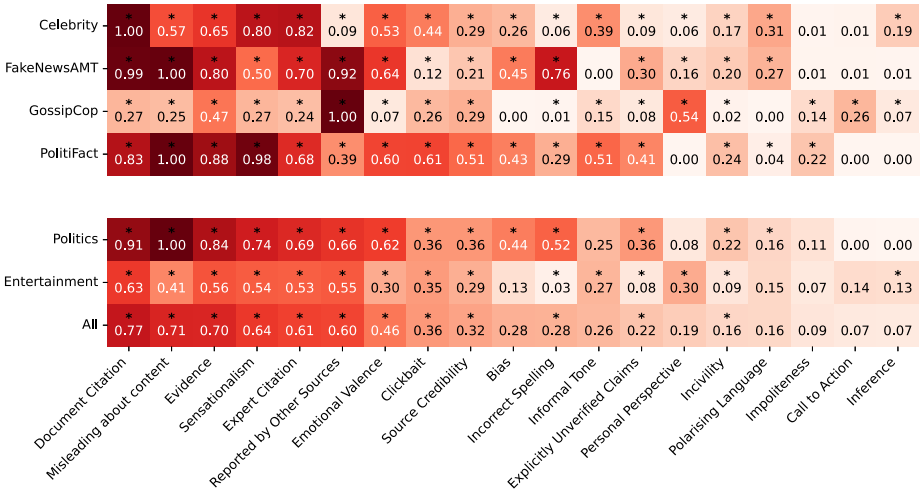

Figure 6 Normalised Pearson's χ^2 statistics per credibility signal. Credibility signals where the null hypothesis H_0 is rejected ($p < 0.05$, 1 degree of freedom) are marked with an asterisk (∗). Results are shown for each dataset, and aggregated by domain; 'Politics' displays the average of FakeNewsAMT and PolitiFact, and 'Entertainment' shows the average of Celebrity and GossipCop. All four datasets are averaged into 'All'. For aggregate results, we reject H_0 if H_0 is rejected in all the aggregated datasets. Credibility signals are sorted in descending order based on the overall average ('All')

Expert Citation, *Reported by Other Sources*, *Emotional Valence*, *Clickbait*, *Source Credibility*, *Incorrect Spelling*, *Explicitly Unverified Claims*, and *Incivility*. Out of these 12 signals, 6 display a particularly high average normalised χ^2 (≥ 0.6), indicating a strong association: *Document Citation*, *Misleading about content*, *Evidence*, *Sensationalism*, *Expert Citation*, and *Reported by Other Sources*. For the other remaining 6 signals, H_0 is rejected only within specific domains. For instance, H_0 is rejected for the signals of *Inference*, *Personal Perspective*, and *Informal Tone* in the Entertainment domain, but not in Politics. Conversely, we only reject H_0 for the signals *Polarising Language* and *Bias* in the Politics domain. Lastly, for some signals, H_0 is rejected only in specific datasets: *Impoliteness* for GossipCop and PolitiFact, and *Call to Action* for GossipCop.

In conclusion, all 19 signals show a statistically significant association with the article's veracity in at least one dataset, with the majority (12 signals) demonstrating a strong association across all four datasets. Additionally, domain-specific signals exist where H_0 is only rejected within either the Politics or Entertainment domains, but not both.

6.2 Ablation study

In this experiment, we evaluate the contribution of each credibility signal to PASTEL's performance through an ablation study. We iteratively remove each of the 19 credibility signals from the dataset, training the label model on the remaining 18 signals. We then compare the performance of this modified model against the model trained with all 19 signals. Table 7 shows the percentage change in $F1_{macro}$ when each signal is excluded.

Overall, individual signals exhibit a relatively small impact on the model's performance. The most influential signal, *Document Citation*, reduces the model's performance by an average of 1.1% across all datasets. The top nine signals positively impacting performance, i.e., those that lead to lower $F1_{macro}$ scores when removed, are: *Document Citation*, *Sensationalism*, *Misleading about content*, *Incorrect Spelling*, *Clickbait*, *Informal Tone*, *Source*

Table 7 Ablation study results. Scores are the percentage change in performance when a certain credibility signal is excluded from the dataset. Signals are sorted increasingly by the mean score

Signal Removed	PolitiFact	GossipCop	FNAMT	Celebrity	Entert.	Politics	Mean
Document Citation	–0.6	–1.0	–1.4	–1.5	–1.2	–1.0	–1.1
Sensationalism	–0.6	–0.5	–0.5	–2.0	–1.2	–0.5	–0.9
Misleading about content	–0.4	–0.3	–2.7	0.6	0.2	–1.6	–0.7
Incorrect Spelling	–0.2	0.1	–1.7	0.0	0.0	–0.9	–0.4
Clickbait	–0.3	–0.1	0.0	–0.4	–0.3	–0.2	–0.2
Informal Tone	–0.6	0.0	0.2	–0.5	–0.2	–0.2	–0.2
Source Credibility	–0.2	–0.1	0.0	0.0	0.0	–0.1	–0.1
Explicitly Unverified Claims	–0.6	0.2	–0.5	0.7	0.4	–0.6	–0.1
Impoliteness	0.0	0.0	0.0	–0.3	–0.2	0.0	–0.1
Expert Citation	–0.5	–0.2	0.6	0.1	–0.1	0.0	0.0
Call to Action	0.4	0.0	0.0	–0.3	–0.1	0.2	0.0
Inference	0.5	–0.1	0.0	–0.2	–0.2	0.3	0.1
Reported by Other Sources	0.5	0.0	0.0	0.3	0.2	0.3	0.2
Incivility	0.9	0.2	–0.3	0.0	0.1	0.3	0.2
Bias	0.9	0.1	0.0	0.0	0.1	0.5	0.3
Personal Perspective	1.6	0.2	–0.3	0.0	0.1	0.7	0.4
Emotional Valence	1.6	0.2	0.1	–0.2	0.0	0.8	0.4
Evidence	–0.3	–0.3	0.8	1.3	0.5	0.2	0.4
Polarising Language	2.0	0.3	–0.3	0.3	0.3	0.8	0.6

Credibility, *Explicitly Unverified Claims*, and *Impoliteness*. Except for *Informal Tone* and *Impoliteness*, all show a statistically significant association with veracity (see Fig. 6).

In contrast, eight signals reduce PASTEL's performance on average, as indicated by an increase in $F1_{macro}$ scores when removed. These, in descending order of impact, are: *Polarising Language*, *Evidence*, *Emotional Valence*, *Personal Perspective*, *Bias*, *Incivility*, *Reported by Other Sources*, and *Inference*. Despite their negative average effect, some signals demonstrate domain-specific benefits, such as *Expert Citation*, *Call to Action*, and *Inference* in Entertainment, and *Misleading about content*, *Incorrect Spelling*, *Source Credibility*, and *Explicitly Unverified Claims* in Politics.

These findings underscore that PASTEL's strength lies in its ability to aggregate multiple credibility signals, as no single signal significantly affects the overall performance on its own. Although some signals, such as *Document Citation* and *Sensationalism*, demonstrate utility across multiple domains, the degree of effectiveness of credibility signals is often domain-specific. For example, while signals like *Source Credibility* and *Misleading about Content* improve performance primarily in the political domain, others such as *Expert Citation* and *Call to Action* show benefits in entertainment.

7 Discussion and conclusion

In this work, we proposed PASTEL, a novel approach that uses LLMs to extract a wide range of credibility signals, which are then aggregated with weak supervision to predict veracity. Extensive experiments show that PASTEL significantly outperforms the unsupervised baseline (LLaMa-ZS) by 38.3%. Additionally, PASTEL achieves 86.7% of the performance of the supervised state-of-the-art RoBERTa model by Goel et al. [52], without using any form of human supervision (neither labelled data nor user interactions as in previous work [7–9]). In cross-domain classification, PASTEL outperforms the supervised state-of-the-art model by a large margin (63%). These results demonstrate the usefulness of our method mainly in scenarios where no in-domain labelled data is available.

PASTEL's ability to leverage credibility signals in a zero-shot setting enables it to maintain high performance across diverse domains, making it well-suited for dynamically changing environments and emergent topics. For example, during the early stages of the COVID-19 pandemic in late 2019, misinformation about the virus spread rapidly, while labelled datasets for training supervised models were not available until mid to late 2020 [65–67]. Additionally, PASTEL offers a key advantage over other weakly supervised methods for misinformation detection that rely on user interactions [7–9]. These approaches depend on users engaging with harmful content before detection is possible, by which time the misinformation may have already caused significant damage. In contrast, PASTEL operates directly at the content level, allowing it to detect misinformation in its early stages of dissemination.

We studied the association between the LLM-predicted credibility signals and the human-annotated veracity labels, revealing that 12 out of the 19 signals exhibit a statistically significant association across all four datasets. Moreover, we observed domain-specific credibility signals that demonstrate higher degrees of association with datasets related to Politics compared to Entertainment, and vice versa. This finding can guide future work in crafting more specialised sets of credibility signals for specific domains. Next, we conducted an ablation study to measure the contribution of each credibility signal towards PASTEL's performance in predicting veracity. We found that the contribution of individual signals is relatively small, and that PASTEL's performance depends on the collective influence of it's wide range of credibility signals rather than in one signal in specific.

8 Limitations and future work

Plenty of research opportunities arise from the implications of this work. While the domain coverage explored in this paper provides valuable insights into PASTEL's performance, we acknowledge the importance of expanding to additional critical domains, such as scientific misinformation or health misinformation [67], in future research. These domains present unique challenges and credibility indicators that would further test PASTEL's robustness and applicability.

Future research may also explore the usefulness of multi-modal credibility signals. For instance, the report by W3C-CWCG [10] describes credibility signals associated with images, such as the originality of the photo and whether it has been manipulated or not. Signals related to audio, video, and even the structure of the content, such as the ads presented, can be considered. Another promising research direction is to explore and mitigate the overconfidence of the LLM when extracting credibility signals, as seen in Fig. 5, where the LLM seldom responds with 'Unsure', which can degrade performance.

The relatively high false negative rate (32.8%) is a limitation that reflects the inherent trade-offs in a weakly supervised framework. The reliance on noisy and imperfect signals can lead to under-detection of misinformation, particularly in edge cases where the signals are insufficiently triggered, as demonstrated in Sect. 5.3. To mitigate the false negative rate, future work could explore incorporating additional domain-specific signals or leveraging multimodal data, such as images and videos, to provide richer input for veracity prediction.

Finally, the current implementation of PASTEL extracts the credibility signals sequentially and independently. This design ensures that the contribution of each individual signal can be rigorously evaluated without introducing interdependencies or assumptions

that might affect the interpretability of results. However, this approach incurs in computational overhead, as the input article and corresponding instruction must be processed repeatedly for each signal. Nonetheless, as each extracted signal produces only a single word (Yes/No/Unsure), the cost associated with generating output tokens, which is typically higher compared to processing input tokens, is considerably mitigated. Future work can explore the feasibility of extracting multiple signals simultaneously by exploiting synergies between signals, and thus reducing redundant processing. For instance, methods such as chain-of-thought or tree-of-thought [68] could provide structured pathways to derive related signals, thus reducing computational costs while potentially improving the accuracy of the extracted signals.

Abbreviations

PASTEL, Prompted Weak Supervision with Credibility Signals; LLM, Large Language Model; NLP, Natural Language Processing; QA, Question-Answering; CWCG, Credible Web Community Group; W3C, World Wide Web Consortium; PWS, Programmatic Weak Supervision; TP, True positives; TN, True negatives; FP, False positives; FN, False negatives; FNR, False negative rate; FPR, False positive rate; ZS, Zero-shot; FT, Fine-tuned; RoBERTa, Robustly Optimized BERT Approach; BERT, Bidirectional Encoder Representations from Transformers; PAS, PASTEL; RoB, RoBERTa; FNAMT, FakeNewsAMT; Entert., Entertainment.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00534-0.

Additional file 1. (ZIP 110 kB)

Acknowledgements

We thank Freddy Heppel, Ivan Srba, and Ben Wu for their valuable feedback. We also acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

Author contributions

JL developed the PASTEL method, participated in the literature review, performed the experiments, and wrote the manuscript. OR participated in the literature review. OR, CS and KB participated in writing the manuscript and provided direction with conceptualisation and methodology. All authors edited and submitted the final manuscript. All authors read and approved the final manuscript.

Funding

This work has been co-funded by the UK's innovation agency (Innovate UK) grant 10039055 (approved under the Horizon Europe Programme as vera.ai, EU grant agreement 101070093) under action number 2020-EU-IA-0282. João Leite is supported by a University of Sheffield EPSRC Doctoral Training Partnership (DTP) Scholarship.

Data Availability

Our code to reproduce the experiments is made fully available at https://github.com/joaoaleite/PASTEL. The datasets used in the experiments are publicly available: (1) FakeNewsAMT and Celebrity (https://lit.eecs.umich.edu/downloads.html), and (2) PolitiFact and GossipCop (https://github.com/KaiDMML/FakeNewsNet).

Declarations

Ethical considerations

LLMs are known to inherit biases from their training data [69], which can manifest in their interpretations and judgements regarding the presence or absence of credibility signals in textual content. These biases may lead to inaccuracies or disparities in signal detection, potentially favouring certain types of content or perspectives over others. Moreover, the deployment of LLM-based systems in real-world applications must navigate concerns around fairness, transparency, and accountability. Researchers and developers are therefore urged to mitigate biases through rigorous testing, data preprocessing, and continuous monitoring.

Also, although efforts aimed at mitigating misinformation are crucial in combating its harmful effects, it is important to acknowledge that these efforts can inadvertently empower malicious actors [70]. By gaining insights into which credibility signals are more easily detected by LLMs, and which correlate more strongly with veracity, malicious users could potentially exploit this knowledge to enhance their misinformation tactics and circumvent automatic detection systems. Therefore, we strongly urge researchers to apply our methodology with caution and in accordance with best practice ethics protocols.

Ethics approval and consent to participate

Not applicable.

Consent for publication
The authors provide their full consent for the publication of this manuscript in EPJ Data Science.

Competing interests
The authors declare no competing interests.

Received: 4 November 2024 Accepted: 14 February 2025 Published online: 21 February 2025

References

1. Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv 53(5). https://doi.org/10.1145/3395046
2. Fu C, Pan X, Liang X, Yu S, Xu X, Min Y (2023) Feature drift in fake news detection: an interpretable analysis. Appl Sci 13(1):592
3. Ksieniewicz P, Zyblewski P, Choraś M, Kozik R, Giełczyk A, Woźniak M (2020) Fake news detection from data streams. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207498
4. Silva R, Almeida T (2021) How concept drift can impair the classification of fake news. In: Anais do IX symposium on knowledge discovery, mining and learning. SBC, Porto Alegre, pp 121–128. https://doi.org/10.5753/kdmile.2021.17469. https://sol.sbc.org.br/index.php/kdmile/article/view/17469
5. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2018) Automatic detection of fake news. In: Bender EM, Derczynski L, Isabelle P (eds) Proceedings of the 27th international conference on computational linguistics. Association for Computational Linguistics, Santa Fe, pp 3391–3401. https://aclanthology.org/C18-1287
6. Goel P, Singhal S, Aggarwal S, Jain M (2021) Multi domain fake news analysis using transfer learning. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp 1230–1237. https://doi.org/10.1109/ICCMC51019.2021.9418411
7. Shu K, Zheng G, Li Y, Mukherjee S, Awadallah AH, Ruston S, Liu H (2020) Early detection of fake news with multi-source weak social supervision. In: Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020. Proceedings, part III. Springer, Berlin, pp 650–666. https://doi.org/10.1007/978-3-030-67664-3_39
8. Helmstetter S, Paulheim H (2018) Weakly supervised learning for fake news detection on Twitter. In: 2018 IEEE/ACM international conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, Los Alamitos, pp 274–277
9. Wang Y, Yang W, Ma F, Xu J, Zhong B, Deng Q, Gao J (2020) Weak supervision for fake news detection via reinforcement learning. In: Proceedings of the AAAI conference on artificial intelligence 34(01), pp 516–523. https://doi.org/10.1609/aaai.v34i01.5389
10. W3C-CWCG (2024) W3C Credible Web Community Group. https://github.com/w3c/credweb. Accessed 2024-02-14
11. Dimou A, et al (2022) Evaluating web content using the w3c credibility signals. In: Towards a knowledge-aware AI: SEMANTiCS 2022—proceedings of the 18th international conference on semantic systems, Vienna, Austria, 13–15 September 2022, vol 55. IOS Press, Amsterdam, p 3
12. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971
13. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates, Red Hook, pp 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
14. Petroni F, Rocktäschel T, Riedel S, Lewis P, Bakhtin A, Wu Y, Miller A (2019) Language models as knowledge bases? In: Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, pp 2463–2473. https://doi.org/10.18653/v1/D19-1250. https://aclanthology.org/D19-1250
15. Leite J (2024) PASTEL Repository. https://github.com/joaoaleite/PASTEL. Accessed 2024-02-14
16. Vlachos A, Riedel S (2014) Fact checking: task definition and dataset construction. In: Proceedings of the ACL 2014 workshop on language technologies and computational social science, pp 18–22
17. Ferreira W, Vlachos A (2016) Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. ACL
18. Wang WY (2017) "Liar, liar pants on fire": a new benchmark dataset for fake news detection. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers). Association for Computational Linguistics, Vancouver, pp 422–426. https://doi.org/10.18653/v1/P17-2067. https://aclanthology.org/P17-2067
19. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A (2018) FEVER: a large-scale dataset for fact extraction and VERification. In: Walker M, Ji H, Stent A (eds) Proceedings of the 2018 conference of the North American, Long Papers. Chapter of the association for computational linguistics: human language technologies, vol 1. Association for Computational Linguistics, New Orleans, pp 809–819. https://doi.org/10.18653/v1/N18-1074. https://aclanthology.org/N18-1074
20. Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B (2018) A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Melbourne, pp 231–240. https://doi.org/10.18653/v1/P18-1022. https://aclanthology.org/P18-1022
21. Santia G, Williams J (2018) Buzzface: a news veracity dataset with Facebook user commentary and egos. In: Proceedings of the international AAAI conference on web and social media, vol 12, pp 531–540

22. Tacchini E, Ballarin G, Della Vedova ML, Moret S, Alfaro L, et al (2017) Some like it hoax: automated fake news detection in social networks. In: CEUR workshop proceedings, pp 1–15. CEUR-WS
23. Zubiaga A, Liakata M, Procter R, Wong Sak Hoi G, Tolmie P (2016) Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE 11(3):0150989
24. Mitra T, Gilbert E (2015) 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, pp 258–267
25. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) Eann: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '18. Association for Computing Machinery, New York, pp 849–857. https://doi.org/10.1145/3219819.3219903
26. Nakamura K, Levy S, Wang WY (2020) Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection. In: Proceedings of the 12th language resources and evaluation conference. European Language Resources Association, Marseille, pp 6149–6157. https://aclanthology.org/2020.lrec-1.755
27. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2020) Fakenewsnet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 8(3):171–188
28. Li Y, Jiang B, Shu K, Liu H (2020) Toward a multilingual and multimodal data repository for covid-19 disinformation. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, Los Alamitos, pp 4325–4330
29. Hossain MZ, Rahman MA, Islam MS, Kar S (2020) BanFakeNews: a dataset for detecting fake news in Bangla. In: Proceedings of the 12th language resources and evaluation conference. European Language Resources Association, Marseille, pp 2862–2871. https://aclanthology.org/2020.lrec-1.349
30. Saikh T, De A, Ekbal A, Bhattacharyya P (2019) A deep learning approach for automatic detection of fake news. In: Proceedings of the 16th international conference on natural language processing. NLP Association of India, International Institute of Information Technology, Hyderabad, pp 230–238. https://aclanthology.org/2019.icon-1.27
31. Gautam A, Jerripothula KR (2020) Sgg: spinbot, grammarly and glove based fake news detection. In: 2020 IEEE sixth international conference on multimedia big data (bigMM). IEEE, Los Alamitos, pp 174–182
32. Dun Y, Tu K, Chen C, Hou C, Yuan X (2021) Kan: knowledge-aware attention network for fake news detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 81–89
33. Bhattarai B, Granmo O-C, Jiao L (2022) ConvTextTM: an explainable convolutional Tsetlin machine framework for text classification. In: Proceedings of the thirteenth language resources and evaluation conference. European Language Resources Association, Marseille, pp 3761–3770. https://aclanthology.org/2022.lrec-1.401
34. Rai N, Kumar D, Kaushik N, Raj C, Ali A (2022) Fake news classification using transformer based enhanced lstm and bert. Int J Cogn Comput Eng 3:98–105
35. Horne B, Adali S (2017) 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, pp 759–766
36. Afroz S, Brennan M, Greenstadt R (2012) Detecting hoaxes, frauds, and deception in writing style online. In: 2012 IEEE symposium on security and privacy. IEEE, Los Alamitos, pp 461–475
37. Rashkin H, Choi E, Jang JY, Volkova S, Choi Y (2017) Truth of varying shades: analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 2931–2937
38. Nikolaidis N, Piskorski J, Stefanovitch N (2024) Exploring the usability of persuasion techniques for downstream misinformation-related classification tasks. In: Calzolari N, Kan M-Y, Hoste V, Lenci A, Sakti S, Xue N (eds) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). ELRA and ICCL, Torino, pp 6992–7006. https://aclanthology.org/2024.lrec-main.613
39. O'Brien N, Latessa S, Evangelopoulos G, Boix X (2018) The language of fake news: opening the black-box of deep learning based detectors. In: Workshop on "AI for social good", NIPS 2018, Montreal, Canada. http://hdl.handle.net/1721.1/120056
40. Giachanou A, Rosso P, Crestani F (2019) Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. SIGIR'19. Association for Computing Machinery, New York, pp 877–880. https://doi.org/10.1145/3331184.3331285
41. Dufraisse E, Treuillier C, Brun A, Tourille J, Castagnos S, Popescu A (2022) Don't burst blindly: for a better use of natural language processing to fight opinion bubbles in news recommendations. In: Proceedings of the LREC 2022 workshop on natural language processing for political sciences. European Language Resources Association, Marseille, pp 79–85. https://aclanthology.org/2022.politicalnlp-1.11
42. Musi E, Reed C (2022) From fallacies to semi-fake news: improving the identification of misinformation triggers across digital media. Discourse Soc 33(3):349–370. https://doi.org/10.1177/09579265221076609
43. Sitaula N, Mohan CK, Grygiel J, Zhou X, Zafarani R (2020) Credibility-based fake news detection. In: Disinformation, misinformation, and fake news in social media: emerging research challenges and opportunities, pp 163–182
44. Zhang AX, Ranganathan A, Metz SE, Appling S, Sehat CM, Gilmore N, Adams NB, Vincent E, Lee J, Robbins M, Bice E, Hawke S, Karger D, Mina AX (2018) A structured response to misinformation: defining and annotating credibility indicators in news articles. In: Companion proceedings of the web conference 2018. WWW '18. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp 603–612. https://doi.org/10.1145/3184558.3188731
45. Fu D, Chen M, Sala F, Hooper S, Fatahalian K, Ré C (2020) Fast and three-rious: speeding up weak supervision with triplet methods. In: International conference on machine learning, pp 3280–3291. PMLR
46. Varma P, Sala F, Sagawa S, Fries J, Fu D, Khattar S, Ramamoorthy A, Xiao K, Fatahalian K, Priest J, et al (2019) Multi-resolution weak supervision for sequential data. Adv Neural Inf Process Syst 32
47. Ratner AJ, De Sa CM, Wu S, Selsam D, Ré C (2016) Data programming: Creating large training sets, quickly. Adv Neural Inf Process Syst 29
48. Smith R, Fries JA, Hancock B, Bach SH (2022) Language models in the loop: Incorporating prompting into weak supervision. arXiv preprint arXiv:2205.02318
49. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang P, Hashimoto TB (2023) Alpaca: a strong, replicable instruction-following model. Stanf Cent Res Found Model 3(6):7. https://crfm.stanford.edu/2023/03/13/alpaca.html

50. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C (2017) Snorkel: rapid training data creation with weak supervision. In: Proceedings of the VLDB endowment. International conference on very large data bases, vol 11. NIH Public Access, p 269
51. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newsl 19(1):22–36
52. Goel P, Singhal S, Aggarwal S, Jain M (2021) Multi domain fake news analysis using transfer learning. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp 1230–1237. https://doi.org/10.1109/ICCMC51019.2021.9418411
53. Lee AN, Hunter CJ, Ruiz N (2023) Platypus: Quick, cheap, and powerful refinement of llms. arXiv preprint arXiv:2308.07317
54. Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, Tafjord O (2018) Think you have solved question answering? Try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457
55. Zellers R, Holtzman A, Bisk Y, Farhadi A, Choi Y (2019) HellaSwag: can a machine really finish your sentence? In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, pp 4791–4800. https://doi.org/10.18653/v1/P19-1472. https://aclanthology.org/P19-1472
56. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2021) Measuring massive multitask language understanding. In: Proceedings of the International Conference on Learning Representations (ICLR)
57. Lin S, Hilton J, Evans O (2022) TruthfulQA: measuring how models mimic human falsehoods. In: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Dublin, pp 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229. https://aclanthology.org/2022.acl-long.229
58. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le QV, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A (eds) Advances in neural information processing systems, vol 35. Curran Associates, Red Hook, pp 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
59. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2022) LoRA: low-rank adaptation of large language models. In: International conference on learning representations. https://openreview.net/forum?id=nZeVKeeFYf9
60. HuggingFace Trainer Documentation. https://huggingface.co/docs/transformers/main_classes/trainer. Accessed 2024-02-14
61. Dettmers T, Zettlemoyer L (2023) The case for 4-bit precision: k-bit inference scaling laws. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol 202. PMLR, pp 7750–7774. https://proceedings.mlr.press/v202/dettmers23a.html
62. Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, Las Casas D, Hendricks LA, Welbl J, Clark A, Hennigan T, Noland E, Millican K, Driessche G, Damoc B, Guy A, Osindero S, Simonyan K, Elsen E, Vinyals O, Rae J, Sifre L (2022) An empirical analysis of compute-optimal large language model training. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A (eds) Advances in neural information processing systems, vol 35. Curran Associates, Red Hook, pp 30016–30030. https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf
63. Hu B, Sheng Q, Cao J, Shi Y, Li Y, Wang D, Qi P (2023) Bad actor, good advisor: Exploring the role of large language models in fake news detection. arXiv preprint arXiv:2309.12247
64. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. Mach Learn 79:151–175
65. Chen E, Lerman K, Ferrara E (2020) Tracking social media discourse about the covid-19 pandemic: development of a public coronavirus Twitter data set. JMIR Public Health Surveill 6(2):19273. https://doi.org/10.2196/19273
66. Hossain T, Logan IV RL, Ugarte A, Matsubara Y, Young S, Singh S (2020) COVIDLies: detecting COVID-19 misinformation on social media. In: Proceedings of the 1st workshop on NLP for COVID-19 (part 2) at EMNLP 2020. Association for Computational Linguistics, Online. https://aclanthology.org/2020.nlpcovid19-2.11. https://doi.org/10.18653/v1/2020.nlpcovid19-2.11
67. Cui L, Lee D (2020) Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885
68. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2023) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 55(9). https://doi.org/10.1145/3560815
69. Nadeem M, Bethke A, Reddy S (2021) StereoSet: Measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416. https://aclanthology.org/2021.acl-long.416
70. Xu D, Fan S, Kankanhalli M (2023) Combating misinformation in the era of generative ai models. In: Proceedings of the 31st ACM international conference on multimedia. MM '23. Association for Computing Machinery, New York, pp 9291–9298. https://doi.org/10.1145/3581783.3612704

Publisher's Note