Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

News Source Credibility Assessment: A Reddit Case Study

Arash Amini, Yigit Ege Bayiz, Ashwin Ram, Radu Marculescu, and Ufuk Topcu

The University of Texas at Austin
{a.amini, egebayiz, ashwin.ram, radum, utopcu}@utexas.edu

Abstract

We present a transformer-based model for credibility assessment, CREDiBERT (CREDibility assessment using Bidirectional Encoder Representations from Transformers), fine-tuned for Reddit submissions focusing on political discourse. We adopt a semi-supervised training approach for CREDiBERT, leveraging the community structure of Reddit. By encoding submission content using CREDiBERT and integrating it with a classification neural network, we improve the credibility assessment for Reddit submission by 3% in F1 score compared to existing methods. Additionally, we introduce a new version of the post-to-post network in Reddit that efficiently encodes user interactions to enhance the credibility assessment task by 8% in the $F1$ score. We demonstrate CREDiBERT's applicability by evaluating the susceptibility of Reddit communities to different topics and assessing the credibility score of unseen sources.

Introduction

Social media and online news platforms have drastically altered the landscape of news consumption. Individuals now encounter a diverse array of news through social media platforms, compared to the time when people relied on newspapers and television channels. While the digital age offers unprecedented personalization and speed in news delivery, it also presents a significant risk of receiving inaccurate information. Discerning fake news on social media platforms is a formidable challenge, necessitating sophisticated methodologies to distinguish it from authentic reporting.

Recent advances in natural language processing have shown great potential for identifying fake news. Through detailed analysis of language patterns and textual features in news content, natural language processing techniques have demonstrated high precision in differentiating fake news from legitimate reports (Raza and Ding 2022). This breakthrough has catalyzed new research avenues, thus empowering scientists to delve deeper into the characteristics of fake news and devise robust countermeasures.

Despite advances in fake news detection algorithms through content, the reputation of the article's source is still one of the key components in assessing the trustworthiness of the article. As evidenced by Pehlivanoglu et al. (2021), historical patterns in reporting accuracy often correlate to potential misinformation or propagandistic intents.

This paper diverges from the predominant focus on identifying fake news in the literature. Instead, we concentrate on discerning the *credibility* of news sources shared to combat the spread of misinformation. Our objective is to ascertain the credibility of the sources rather than categorically labeling specific news as true or fake. Although there is no direct, infallible correlation between a source's credibility and news veracity, the news source historical track record can offer valuable insights into its general reliability and editorial practices. Such an analysis, especially when conducted over an extended period of time, can reveal patterns and tendencies that indicate the source commitment to accuracy and journalistic integrity.

Source credibility refers to the perceived believability of a source, which significantly influences how receivers processed the messages it shares. It is primarily comprised of perceptions of the source's trustworthiness (the confidence in the source's intent to communicate valid information) and expertise (the perception that the source is capable of providing valid assertions). Credibility is fundamentally a perceptual judgment, distinct from, but related to, the factual accuracy of the information presented (Hocevar, Metzger, and Flanagin 2017).

Throughout this paper, we define the credibility of a news source as a numerical score that sums up the reliability and trustworthiness of the information it disseminates. Political bias and credibility are crucial indicators of how individuals perceive a news source, significantly impacting how people engage with the media. The rise of social networks has challenged the traditional dominance of conventional media over the information ecosystem, with content creators, influencers, and political elites now having a more direct impact on public opinion than ever. Unlike conventional media outlets, assessing the credibility of digital content is challenging due to the wide variety and volume of online content. To address this, we train a language model to detect features related to credibility in textual information. This framework automates the assessment of credibility and improves our understanding of the quality of information that communities receive on various subjects.

Misinformation and fake news can be ambiguous and have different meanings depending on the context. For con-

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

sistency, we define *fake news* and *misinformation* as any news or article contaminated by fake stories or false narratives intended to influence public opinion. Note that a high credibility score for a media source does not imply it never disseminates unreliable information. Rather, it indicates that the likelihood of information shared from this source being contaminated by misinformation is low.

Reddit, a prominent social news website, exemplifies the challenges inherent in news source verification. Reddit hosts various *subreddits*, each focused on a specific topic. Users frequently post submissions in political subreddits, including links to news articles for discussion. Reddit provides a distinctive platform where users can anonymously post content in various topic-specific communities, providing a conducive environment for disseminating fake news. The substantial proliferation of misinformation about COVID-19 in Reddit and QAnon incidents underscores the importance of mitigating misinformation across Reddit before it spreads to other platforms.

We call a website (or platform), *source*, when it frequently disseminates information; thus, in this manuscript, the definition of a source includes conventional media outlets and small content creators such as YouTube channels or Twitter profiles.A source on Reddit is *verifiable* if its information history, like that of conventional news outlets, can be tracked. The credibility score of a verifiable source is known, while an unverifiable source has an unknown credibility score. We denote the credibility of a submission by the credibility score for the source of the submission in question.

Expanding news sources to social media platforms such as Twitter and YouTube made verifying the trustworthiness of the sources of news shared often challenging. Although some submissions link to verifiable sources, such as major news websites, many do not. Our analysis indicates that more than 60% of submissions posted in major political subreddits do *not* have a verifiable source.

Recognizing that identical news stories are often reported with distinct biases across various communities, we leverage Reddit's community-based structure to generate a comprehensive dataset. This dataset comprises over 1.5 million pairs of Reddit submissions, each pair referencing the same event but sourced from different outlets, thus having different credibility scores. To this end, we employ sentence transformers (Thakur et al. 2020) to assemble a dataset of submission pairs referencing the same event based on their textual similarity. We train a bi-encoder, CREDiBERT, using this dataset to capture credibility-related features by aligning the textual similarity of submission pairs with their credibility score discrepancies. An artificial neural network classifies the submissions as credible or non-credible by their CREDiBERT-generated embeddings. The proposed framework outperforms BERT-based sentence classifiers by over 3% in the F1 score in the submission credibility assessment task.

As the last contribution, we encode the user interaction information to enhance the credibility assessment. We introduce a weighted post-to-post network (Hurtado, Ray, and Marculescu 2019) that efficiently represents the social interactions among Reddit users. In the original post-to-post network, submissions are represented as nodes in a graph, with edges between submissions if they share users who comment on them. We enhance this network by introducing a weighted version of the post-to-post network that incorporates user reactions to the comments, providing additional context and depth to the interactions between submissions. Of note, this network does *not* rely on user profiling, which often raises privacy and security concerns. We then employ a Graph Convolutional Network to infer the credibility of submissions. This approach ensures a balance between insightful analysis and user privacy, thus paving the way for more ethical credibility assessment in social media.

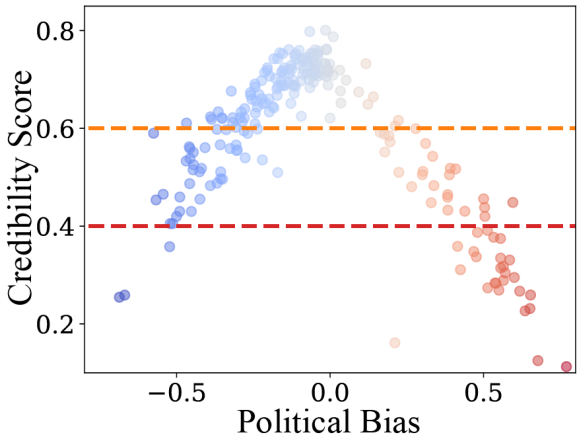

Figure 1: Credibility score with respect to Political Bias for 223 news sources. The data is normalized and sourced to the Ad-Font media website.

We cross-validate our results with human evaluations in two stages. Given the large data requirements for training a bi-encoder, we automate the pairing process and validate its accuracy by having humans manually evaluate random samples selected from the results of the automated pairing algorithm. Additionally, we validate the credibility assessment task by comparing the results with human evaluations for 100 samples.

This paper is structured as follows: It starts with a review of related works. Next, we describe our dataset, which comprises approximately 1.2 million Reddit submissions and over 12 million pairs of submissions. In the first half of the Methods section, we present the CREDiBERT and discuss the direct credibility assessment task. The second half introduces a new post-2-post network, incorporating user reactions from comments to enhance the results further. Following this, we compare our framework with other baselines in the Results section and discuss its applications and limitations in the Discussion section. The paper concludes by highlighting our main contributions.

Related Work

Over the last decade, the issue of fake news detection has garnered substantial interest, particularly highlighted by its

Submission content	Subreddit	Credibility Score	Overall Score
Poll: Majority of Republicans would support Trump in 2024	r/politics	0.74	78
For everyone who thinks he is going away: Majority of Republicans would support Trump in 2024	r/Republican	0.73	85
Trump remains the most popular Republican in the country and the leading candidate for the 2024 GOP nomination	r/Conservative	0.24	963

Table 1: Example of submissions referencing the same event with various representations in different subreddits

implications during major events such as the 2016 U.S. presidential election. Initial strides in the domain were made by Castillo et al. (2013), who focus on assessing the credibility of Twitter content during crises, thus laying the groundwork for understanding digital misinformation. The momentum was significantly amplified post-2016 by attracting a wide array of research, including social impact assessments by Allcott and Gentzkow (2017) and broader misinformation detection techniques by Shu et al. (2017). Chan and Donovan (2017) further extended this realm into strategies for countering misinformation. A persistent challenge noted across studies, particularly by Torabi Asr and Taboada (2019), is the scarcity of high-quality labeled data, which impedes the development of robust detection mechanisms.

Automated Fact Checking

Automated fact-checking has significantly evolved with the advancement of artificial intelligence, increasingly relying on the intricate features of news content and social context to discern authenticity. Notably, the introduction of Bidirectional Encoder Representations from Transformers (BERT) marked a substantial advancement in natural language processing, enhancing the capability to process and understand news content. Devlin et al. (2018) spearheaded this development, which was later built upon by Jwa et al. (2019), who proposed exBAKE, a BERT-based architecture specifically tailored for fake news identification. This model notably improved detection efficacy, as measured by the F1 score. However, the journey of refinement continued as Przybyła (2020) critically examined the text style role in automated fact-checking, revealing that while BERT-based methods are potent, they tend to overfit and might underperform in style capturing tasks of large texts compared to Stylometry and BiLSTM. Addressing some of these concerns, Raza and Ding (2022) proposed an enhanced transformer model that leverages both content and social media traces, pushing the boundary further in automated fact-checking. This ongoing evolution, including the rise of pre-trained large language models, continually presents new opportunities and challenges in the field, as evidenced by the work of Chen and Shu (2024) and Leite et al. (2023), who explore these cutting-edge developments in combating misinformation.

Credibility Assessment

The quest to understand and counteract misinformation has led researchers to examine the role of source credibility closely. Pehlivanoglu et al. (2021) suggest that historical patterns in a source reporting accuracy can indicate misinformation or propagandistic content, setting a foundation for further studies. Spezzano et al. (2021) explore how users recognize and react to news from low-credibility sources on social media, while Mosleh and Rand (2022) shift focus towards measuring user exposure to misinformation rather than identifying false news directly. Measuring user exposure to misinformation necessitates a robust method to assess media quality, a controversial and challenging task. Bachmann et al. (2022) contribute to this by proposing a quantitative measure for news media source quality. Further technological advancements are employed by Chiang et al. (2022), who apply BERT and artificial neural networks to automate the identification of source credibility. Lastly, Chipidza et al. (2022) provide an interesting angle by investigating the interplay between the ideological stance and perceived source credibility, particularly in the context of COVID-19 discussions on Reddit. These research studies illuminate various facets of source credibility, highlighting its multifaceted nature and the diverse methodologies employed to understand and evaluate it.

Misinformation on Reddit

Reddit's unique combination of anonymity and community-driven content, exemplified by incidents like QAnon, fosters an environment conducive to spreading fake news. Understanding the mechanics of this spread is crucial for effective detection and mitigation. Glenski et al. (2018) shed light on this issue by highlighting how user engagement metrics and subreddit-specific norms significantly influence the perceived credibility of posts on Reddit. Their findings underscore the importance of contextual and community factors in content credibility. Building upon the need for robust data in studying these dynamics, Sakketou et al. (2022) introduced the FACTOID dataset, a comprehensive collection of Reddit submissions with annotated credibility and bias information. This dataset facilitates a deeper analysis of how misinformation spreaders operate within the Reddit ecosystem. Bond and Garret's (2023) study paved the way for more nuanced research into users' temporal and contextual interactions with both true and false news submissions on the platform, contributing to the broader understanding of misinformation propagation in online communities.

This paper distinguishes itself from the existing body of literature by introducing two significant innovations: First we leverage Reddit's intricate community structure to auto-

matically curate a novel dataset comprising pairs of submissions referencing identical events, providing a unique vantage point for understanding source credibility. Secondly, we introduce the *CREDiBERT* model, a semi-supervised approach utilizing a Siamese network architecture specifically designed to mitigate the overfitting issues prevalent in standard BERT-based models for this task (Przybyla 2020). This adaptation allows for a more robust and generalizable source credibility assessment, addressing a critical challenge in the field.

Data

We compile a comprehensive dataset from five major political subreddits: r/politics, r/Conservative, r/Republican, and r/democrats. Our dataset, covering the period from January 2016 to December 2022, includes detailed information such as submission text, author IDs, source domains, submission times, associated subreddits, overall submission scores, and comment counts for each submission[1]. We then collect all available comments from these submissions. We select these specific subreddits based on their activity volume to ensure a diverse range of political discourse and viewpoints. In collecting this data, we adhere to ethical guidelines for research involving online communities, ensuring the anonymity and privacy of the users.

We utilize a dataset from the Ad Fontes Media website[2], which provides political bias and credibility scores for 223 major news sources. Ad Fontes Media assigns credibility scores and political bias ratings to media outlets, with credibility scores ranging from 0 (least credible) to 64 (most credible), and political bias ratings from -42 (extreme left) to 42 (extreme right). In this paper, we normalize the credibility and bias scores by dividing them to 64 and 42, respectively. Figure 1 depicts the credibility scores concerning political bias for 223 media sources. We note that the curve is asymmetric, indicating that left-leaning sources are perceived as more credible than right-leaning ones.

We ensure the reliability of this data by cross-validating the scores with additional datasets from FACTOID date set (Sakketou et al. 2022) and Media-Bias-Fact-Check website[3] for the common sources. We consider a news source as *verified* if one can access the history of information shared by the news source in question. While the Ad-front media does not share details on the establishment of credibility scores, they state that the credibility score reflects how frequently information shared by news outlets contains misinformation or deceptive narratives.

After thorough data cleaning, the dataset comprises a total of $1,312,853$ submissions. We organize comments based on their reply level: first-tier comments are direct responses to submissions, second-tier comments are replies to first-tier comments, and so on. We limit our analysis to comments up to the fourth tier, as we observed that comments beyond this level often diverge significantly from the original submission content. To maintain data quality, we exclude comments when the author's account has been deleted, removed, or if the comment text is unavailable. The experiments utilize data collected from the 2022, which includes $85,291$ submissions. This temporal split helps to evaluate the model performance on recent, previously unseen data, providing a more reliable assessment of its predictive capabilities.

[1]All Reddit data collected from https://pushshift.io

[2]Details available in https://adfontesmedia.com/

[3]Accessible here https://mediabiasfactcheck.com/

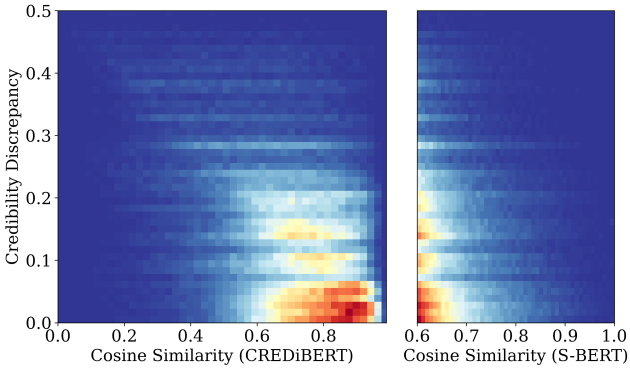

Figure 2: The distribution of credibility score discrepancies with respect to cosine similarity between embeddings provided by CREDiBERRT (left) and S-BERT (right) for $300,000$ pairs of submissions from 2022. Pairs are selected to have at least 0.6 similarity score for S-BERT embedding, thus only parts with data are depicted.

Method

We develop CREDiBERT based on the premise that a text encoder capable of capturing features related to credibility should map submissions referencing the same event with significant credibility discrepancies far apart. Inspired by the effectiveness of bi-encoders (Reimers and Gurevych 2019; Thakur et al. 2020) in producing sentence embeddings that surpass BERT-based models, we train CREDiBERT as a bi-encoder that aligns textual similarities of submissions with their credit discrepancies. The initial step in training CREDiBERT involves identifying pairs of submissions referencing the same event. To this end, we employ Sentence Transformers to assess textual similarity between pairs and identify submissions referencing the same event from different sources. After training CREDiBERT, we use the embeddings it generates to train another artificial neural network for credibility assessment task. It is important to note that when we discuss the credibility score for a submission, we specifically refer to the credibility score for the source of that submission.

Submission Pairing Moderators play a pivotal role in shaping the content of each subreddit, thus ensuring that submissions align with the intended narrative and bias. This moderation influences the diversity of news sources and perspectives presented within their domains and provides us with submissions that reference the same event from various sources. We identify submissions referencing the same event through their textual similarity measured by sentence transformer (S-BERT). Table 1 shows an example of a sub-

mission referencing the same event with different credibility scores.

While the BERT-based model excels at creating context-aware word embeddings through training with a masked language model and next-sentence prediction, it falls short in efficiently embedding entire sentences in the latent space. To overcome this challenge, Reimers et al. 2019 introduced sentence transformers, which use a Siamese network architecture to enrich sentence-level embedding. The resulting S-BERT model outperforms traditional BERT-based models in sentence embedding tasks. In this study, we employ the pre-trained "all-distilroberta-v1" model from Huggingface website.

Let us define the complete set of collected submissions as $\mathcal{S} := \{s_1, s_2, \cdots, s_n\}$, where n represents the total number of submissions. For each submission s_i, we denote its text, credit score, and posting time as p_i, c_i, and t_i, respectively. To determine the similarity between two news submissions, we encode the text p_i through a sentence transformer, denoted by $\mathbf{T}$. This model maps the content of submission s_i to the embedding e_i, by $e_i = \mathbf{T}(p_i)$, where e_i is a vector of real numbers with 768 dimensions. We utilize a bi-encoder to identify similar submissions, which assesses the textual similarity between two submissions, s_i and s_j, by the cosine similarity of their corresponding embeddings. We calculate the similarity by

$$\cos(e_i, e_j) = \frac{\langle e_i, e_j \rangle}{\|e_i\| \|e_j\|}, \tag{1}$$

where $\|\cdot\|$ and $\langle\cdot\rangle$ denote the Euclidean norm and the inner product, respectively. This method allows us to quantitatively compare submissions based on their textual content, enabling a more precise identification of similar news stories across different subreddits.

We opt for bi-encoders over cross-encoders to efficiently compare the vast number of submission pairs, primarily due to their faster processing capabilities. Reimers and Gurevych (Reimers and Gurevych 2019) note that comparing $10,000$ pair of sentences using cross-encoders would take about 65 hours while generating embeddings and computing cosine similarities would take less than 6 second. Given that our dataset encompasses over one million submissions, this choice significantly reduces the computational burden.

Recognizing that submissions referencing the same event are likely posted in close temporal proximity, we implement a time constraint, denoted as Δ, to refine our search. We let Δ equal to two days across this manuscript unless otherwise stated. This constraint means we only consider pairing submissions if their posting times, represented by $t_{ij} = |t_i - t_j|$, are within Δ.

Two submissions are similar if they meet the temporal criteria and their similarity score exceeds the similarity threshold $\bar{e} = 0.60$. For each qualified pair of similar submission (s_i, s_j), we then calculate the credit score difference $c_{ij} = |c_i - c_j|$, which is the absolute difference in their credit scores. This approach streamlines our process, ensuring we focus on our analysis's most relevant and temporally aligned submission pairs. We define the pool of all pairs of similar submissions, $\mathcal{P}$, by

$$\mathcal{P} := \{(s_i, s_j, c_{ij}) |\ \bar{e} < \cos(e_i, e_j),\ |t_i - t_j| \leq \Delta\}. \tag{2}$$

This approach effectively expands the dataset from 1.2 million individual submissions to a comprehensive set of 12 million uniquely paired submissions. Automated submission pairing facilitates the substantial dataset expansion and allows us to train a bi-encoder, CREDiBERT, that captures credibility-related features. This process also enables regular, unsupervised training of the bi-encoder, ensuring it stays current with recent trends.

To ensure a balanced training set, we meticulously selected $1,383,385$ samples from the pool of $12,171,894$ pairs. We balance the data set by dividing the set of all pairs based on their credibility score discrepancies into $K = 5$ groups. If a pair (s_i, s_j, c_{ij}) has their credibility discrepancy satisfies $\frac{k-1}{K} < \varrho c_{ij} < \frac{k}{K}$, then this pair belongs to group $k \in \{1, \cdots, K\}$ and $\varrho = 2$ is the scaling factor. We balance the dataset by limiting the maximum number of samples selected from each group to $300,000$. If a group has a population smaller than the ideal number of samples, we select all the populations of that group. We perform this selection to achieve representative and diverse training data, enhancing the reliability and generalizability of the CREDiBERT.

We validated the automated pairing by randomly selecting $1,000$ pair samples from the submission pair pool, $\mathcal{P}_\tau$, and having two humans assess whether pairs reference the same event. The results indicate that the proposed algorithm accurately predicted 92.7% of the pairings, with the pairs almost always covering the same subject. For transparency, we include the data and the results of expert evaluations in the code files.

CREDiBERT To generate embeddings that reflect properties associated with the credibility scores of submissions, we train a bi-encoder, CREDiBERT, that aligns the embeddings of submission pairs based on their credibility score discrepancies. We subsequently use these embeddings to train a secondary artificial neural network for credibility assessment tasks. The underlying objective is to align the embeddings from CREDiBERT such that cosine similarity between pairs with minor credibility score differences have values close to 1, indicating similarity, while those with significant discrepancies are marked as dissimilar.

To train the CREDiBERT model, we adopt a Siamese architecture similar to S-BERT training. This choice is motivated by the architecture effectiveness in processing and comparing text embeddings when pairs of labeled sentences are available. The training is guided by the following objective function,

$$\sum_{(i,j)\in\mathcal{P}_\tau} |\cos(\mathbf{C}(p_i), \mathbf{C}(p_j)) - \big(1 - \varrho c_{ij}\big)|^2, \tag{3}$$

where $\mathbf{C}(\cdot)$ denotes the embedding generated by CREDiBERT, $\mathcal{P}_\tau$ represent the set of training pairs and $\varrho = 2$ is the scaling factor. The objective function aims to minimize the discrepancy between the cosine similarity of the embeddings and the credit score differences. We trained the CREDiBERT model using the 'all-distilroberta-v1' S-BERT model

Model Type	Model	F1		
		Non-Credible	Credible	Overall
Random	Random	0.482	0.513	0.504
	Majority	0.000	1.000	0.583
Word Embedding	word2vec-google-news-300	0.591 ± 0.002	0.735 ± 0.002	0.674 ± 0.001
	fasttext-wiki-news-300	0.597 ± 0.003	0.739 ± 0.001	$\mathbf{0.679 \pm 0.001}$
BERT-based Text Classifier	BERT-base-uncased	0.710 ± 0.005	0.806 ± 0.006	$\mathbf{0.762 \pm 0.006}$
	DistilBERT-base-uncased	0.704 ± 0.008	0.805 ± 0.004	0.758 ± 0.006
Sentence Transformer	S-BERT: all-distilroberta-v1	0.637 ± 0.002	0.766 ± 0.001	0.712 ± 0.001
	CREDiBERT (*Ours*)	0.735 ± 0.002	0.824 ± 0.001	$\mathbf{0.791 \pm 0.001}$

Table 2: CREDiBERT surpasses other text classification models in binary credibility classification of submissions, as evidenced by cross-validation results. We train and validate all models using $56,491$ and $12,088$ samples and reporter results on $12,088$ test samples. We show the best F1 performance in each class in bold.

as the starting point with a dataset of $1,312,853$ pairs of submissions. The entire training process took nearly 8 hours on a single NVIDIA A100 GPU.

Figure 2 illustrates the distribution of credibility score discrepancies of submission pairs with respect to the cosine similarity of embeddings generated by CREDiBERT and S-BERT. Recall that we chose the minimum similarity score of $\bar{e} = 0.60$. Therefore, the cosine similarity of all the S-BERT-generated embeddings for the selected submission pairs is greater than 0.60. It is evident from Figure 2 that the embeddings from CREDiBERT demonstrate a strong correlation with the credibility discrepancies between the submissions compared to S-BERT.

Credibility Assessment We train an artificial neural network equipped with a three-layer perceptron and a softmax output layer to assess the credibility of submissions. We give the classifier network CREDiBERT-generated embeddings of submissions and train it to predict the corresponding submission labels. To ensure a fair comparison with baselines, we maintain temporal separation between the data used to train CREDiBERT and the data for the credibility assessment task.

Post-to-Post Network User interaction with submission provides rich information regarding the submission contents. However, user monitoring can potentially leads to privacy violations and erode the trust between the community and the platform. We develop a post-to-post network based on user interaction patterns to overcome this challenge and enhance the result of the credibility assessment task. The underlying premise is that if users exhibit similar reactions to two submissions, these submissions imply some similarity. Ideally, one could encode these reactions alongside text embeddings and analyze them through a model to discern subtle similarities. However, this approach poses a significant challenge due to the sheer volume of user engagement, with tens of thousands of active authors commenting on submissions. To address this issue, we explore an alternative method to distill meaningful patterns from user interactions by focusing on key indicators that reflect the essence of users reactions without getting overwhelmed by the data volume.

The post-to-post network is based on the premise that submissions with similar commenting patterns, particularly regarding their authors, are likely to be more closely related. While the original network introduced by (Hurtado, Ray, and Marculescu 2019) considers submissions connected if they share at least one commenter, Bond and Garret (2023) suggest that authors may exhibit different reactions depending on the news authenticity. We employ S-BERT to encode the comment content to estimate the users reactions toward the submissions. For instance, if a comment q_k^i on submission s_i responds to a parent comment q_l^i, which has similarity toward the post by α_l^i, we can measure the similarity between these comments by the cosine similarity of their embeddings, $\cos(\mathbf{T}(q_l^i), \mathbf{T}(q_k^i))$, recalling $\mathbf{T}(\cdot)$ represent the S-BERT encoding model. We can then calculate the submission-similarity for comment q_k^i toward the submission s_i by

$$\alpha_k^i := \alpha_l^i \times \cos(\mathbf{T}(q_l^i), \mathbf{T}(q_k^i)), \tag{4}$$

where q_l^i is the parent comment for q_k^i and $\alpha_0^i = 1$. This method enriches the analysis by considering the depth of user engagement and its impact on the perceived similarity between various submissions.

We initiate the analysis with first-tier comments, assigning a submission-similarity score based on the cosine similarity between the submission and the comment, calculated as $\cos(\mathbf{T}(p_i), \mathbf{T}(q^i))$ for some first tier comment q^i, posted on submission s_i. This process is then iteratively applied to second-tier comments and subsequent tiers, allowing us to determine the post-similarity score for all comments across different levels. To encapsulate the spectrum of user reactions for each post, we construct a vector, r_i, which aggregates the reactions of all users within the set $\mathcal{A}$. For an author a who has commented on submission s_i, the reaction is quantified in the ath element of the vector r_i by

$$\{r_i\}_a = \frac{1}{|\mathcal{C}_a^i|} \sum_{k \in \mathcal{C}_a^i} \alpha_k^i, \tag{5}$$

where $\mathcal{C}_a^i$ is the set of all comments author a posted in reply to submission s_i. We denote the cardinality of the set $\mathcal{C}_a^i$,

the number of distinct elements in the set, by $|\mathcal{C}_a^i|$. In cases where an author has not commented on submission s_i, we set $\{r_i\}_a = 0$. Given the large number of authors, represented by $|\mathcal{A}|$, we navigate the challenge of combining the embedding vector and the reaction vector for each submission by introducing a weighted version of the post-to-post network.

To better capture the intricate interplay between user interactions and submissions, we define the post-to-post graph $\mathcal{G} := (\mathcal{V}, \mathcal{E}, \mathcal{W})$. In this graph, $\mathcal{V}$ denotes the nodes (submissions), with $\mathcal{E}$ and $\mathcal{W}$ representing the edges and weights, respectively. We consider the set of users who commented on post $i \in \mathcal{V}$ as $\mathcal{A}_i$. An edge is established between submissions i and $j \in \mathcal{V}$ if they share more than m common users, leading to the edge set $\mathcal{E}$. Thus $\mathcal{E}$ is defined as

$$\mathcal{E} := \{(s_i, s_j) \mid |\mathcal{A}_i \cap \mathcal{A}_j| > m\}. \tag{6}$$

This structure allows us to analyze the relationships and similarities between submissions based on user engagement patterns. While the original post-to-post network captures the basic structure and offers insights into the connections between posts, it does not differentiate between positive and negative user reactions to submissions. We enhance the network by assigning weights to its edges to address this. Specifically, for an edge $(i, j) \in \mathcal{E}$, we calculate the weight w_{ij} by

$$w_{ij} = \langle r_i, r_j \rangle, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denote the inner product between two vectors. This weight reflects the degree of similarity in user reactions to both submissions. The weights are then stored in $\mathcal{W}$.

Assigning edge weights to incorporate user responses enriches the network's analytical depth. For instance, if a user exhibits negative and positive reactions to two distinct submissions, our model interprets this as indicating a strong separation between these submissions, while the original network presented by Hurtado, Ray, and Marculescu (2019) considers them strongly connected. This nuanced analysis allows us to map the network of submissions more accurately, considering the presence of user interactions and their qualitative nature. On the other hand, this method avoids profiling individual users, thus upholding user privacy and safety.

Result

We conduct four distinct experiments to test the accuracy, reliability, and applicability of the proposed framework. We study the credibility assessment task in the first experiment and cross-validate results with baseline models and human assessment. In the second experiment, we investigate the effect of integrating the post-to-post network on the credibility assessment task. We demonstrate the applicability of the proposed framework for various tasks in the third and fourth experiments. We challenge the framework to estimate the credibility of six unseen sources for the third task. All the information linked to these sources was removed from the training data at all stages. Finally, we perform a case study to detect the susceptibility of communities to misinformation with respect to different topics. In the rest of the paper,

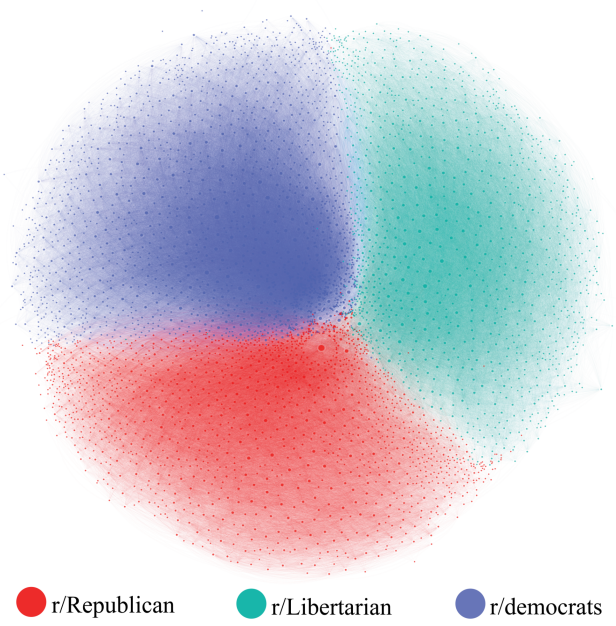

Figure 3: The post-to-post network for $4,371$ submissions. The network shows a strong separation between different communities. For brevity, only edges with weights over 0.2 are shown.

we pair two submissions if they have a cosine similarity of at least $\bar{e} = 0.6$ and a posting time difference of no more than $\Delta = 2$ days.

Credibility Assessment

We assess submission credibility by predicting their labels using the proposed framework. The experiments include scenarios where credibility is binary (credible or non-credible) and where multiple levels of credibility can be assigned to a submission. We perform a sensitivity analysis for the credibility threshold for binary classification tasks and cross-validate the binary classification results with baselines and human assessments.

Binary Classification Task For binary classification, we call a source credible if it has a credibility score greater than a certain threshold Υ. To study how different choices for credibility threshold would affect the results of binary classification, we experiment with five different values of Υ. Figure 4 shows the results of the binary classification corresponding to the values of $\Upsilon = \{0.40\,, 0.50\,, 0.55\,, 0.60\,, 0.7\}$. It is evident that the task becomes more challenging for $\Upsilon = \{0.55\,, 0.60\}$ because the labels are distributed evenly for these thresholds. The S-BERT-based model performed close to the majority for $\Upsilon = \{0.40\,, 0.70\}$, indicating that high and low threshold values make the data unbalanced. However, CREDiBERT maintains its superior performance for these threshold values, demonstrating its ability to distinguish between different levels of credibility. We set the credibility threshold for the rest of the study to $\Upsilon = 0.55$. This decision would ensure

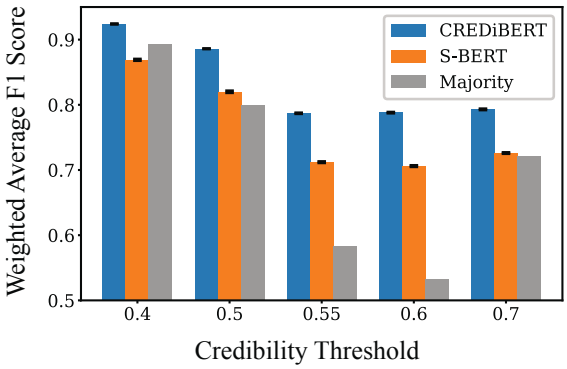

Figure 4: Average weighted F1 score for binary classification task with various credibility thresholds.

a balanced data set, which is important in learning baseline models. We did not choose a higher threshold since the distribution of sources with a credibility score greater than 0.6 becomes extremely dense, as depicted in Figure 1

Cross Validation We focus on comparing the proposed model, CREDiBERT, against the word2vec embedding model (Mikolov et al. 2013), standard BERT text classification model (Devlin et al. 2018), and the S-BERT embedding model (Reimers and Gurevych 2019). We emphasize that submission texts are often short and include only one sentence. Thus, methods such as bag-of-words and stylometry, which require medium to large text corpus, are incompatible with this task (Przybyla 2020). The experiment data set is collected exclusively from 2022, comprising a total of $80,248$ submissions, divided into training sets (70%), validation (15%), and testing (50%) sets. All models are trained over the identical training data set and the results are reported over the test data set. Notably, we train CREDiBERT on submission pairs posted before 2022, ensuring that our training and test data sets are novel and independent. The details of each baseline model are available in Appendix A. We obtain the confidence interval by repeating the training of each model 20 time.

Table 2 presents the cross-validation results for the binary classification task. The proposed framework outperforms the other baselines in terms of the F1 score by 4%. Although BERT-based text classifiers outperformed other baselines, training BERT requires significantly more computational power than alternative models. As expected, word2vec-based models cannot perform properly, as they disregard the style and tone of the text.

Human Assessment We cross-validate the results with human assessment. To do this, we randomly selected 100 submissions from February 2022 and employed four humans to assess the credibility label of the submissions. To ensure that human credibility annotation is consistent, we selected a submission 20 from the same time period and trained annotators by evaluating the learning dataset. After the training phase, each human assesses the credibility of the selected submissions. Table 3 shows the human assessment of the credibility results for $\Upsilon = 0.55$, against the models trained for binary classification. Expert evaluation results in an accuracy of 0.633 ± 0.17. The overall agreement of expert was 66.7%, which is defined by the average proportion of all possible pair of raters that agree on an item. We emphasize that this task is difficult for humans because most submission texts are news headlines and sources with low credibility. Although CREDiBERT outperforms human assessments by more than 14%, other models perform close to experts. We emphasize that distinguishing credible from non-credible news is challenging for humans due to several factors highlighted in the literature(Horne, Khedr, and Adali 2018; Horne, Nørregaard, and Adali 2019). The sheer volume and speed of the information online often overwhelm traditional vetting processes and individual capacity for evaluation. People frequently consume news passively via social media, leading them to rely on mental shortcuts rather than careful analysis, making them susceptible to engaging but misleading content characterized by simpler language or negativity. Furthermore, echo chambers can normalize hyper-partisan or even false narratives, blurring the lines between legitimate, biased, and unreliable sources. This difficulty is compounded by tactics employed by some sources, such as strategically mixing true and false information or copying content from credible outlets, which can obscure the source's true nature and intent.

Multi-Class Classification The Ad-Font recognizes a source as low credible if it has a normalized credibility score of less than 0.4, mixed if its credibility score falls between 0.4 and 0.6, and credible if it has a score greater than 0.6. We demonstrate the strengths and limitations of the proposed model by examining the framework performance on multi-class classification task. We first divided the credibility score into 3 classes identical to the Ad-Font website annotations. Figure 5 shows the results for this task. It is evident that CREDiBERT outperforms other models in all classes; while all models struggle with low credibility class, the CREDiBERT performs significantly better than other models in the detection of low credibility submissions, further showcasing the ability of CREDiBERT in encoding credibility-related features. The results obtained from CREDiBERT have tighter confidence bounds, indicating robustness, especially on low-credible submissions. We emphasize that the average weighted F1 score for this task is significantly lower than the binary classification task. This phenomenon indicates that predicting credibility labels with high resolution is challenging even for CREDiBERT-based models.

Post-to-Post Network

We integrate the post-to-post network with a Graph Convolutional Network (GCN) (Kipf and Welling 2016) for the binary classification task. To this end, we feed the textual embedding of the submissions to the GCN as the node features along with the network information. We compare the CREDiBERT model against the node2vec embedding and baseline models discussed in the previous section. The experiment details involving node2vec are available in Ap-

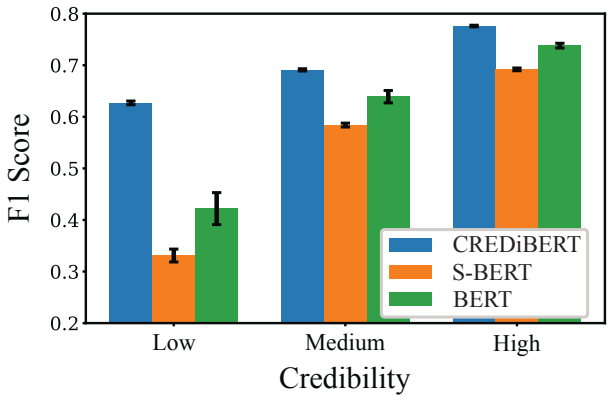

Figure 5: F1 score for multi-class classification task. The predictions based on CREDiBERT embeddings outperform other models in detecting low credibility submissions.

pendix A. We remark that the node2vec experiment does not have access to textual information and asses credibility solely through post-to-post network information. We focus on submissions from 2022 across three specific subreddits: r/democrats, r/Libertarian, and r/Republican. The reason behind this selection is that r/politics and r/conservative have a substantial size, posing significant computational challenges.

Figure 3 illustrates the resulting post-to-post network for the $4,371$ submissions encoding information of more than $100,000$ comments. Figure 3 representation allows us to observe how well a post-to-post network can distinguish between communities and reveal the submissions interaction. To select submissions, we first choose the authors with more than 5 comments during 2022 and then select the posts with at least 2 comments from the selected authors to ensure meaningful interaction. After cleaning data, we are left with $4,371$ submissions.

Node2vec Node2vec, introduced by Grover and Leskovecis (2016), is a graph embedding technique. It effectively represents graph nodes as vectors in a latent space, capturing both the structural characteristics and homophilic tendencies of the graph. We pass the graph embeddings to a classifier layer for binary classification of the submissions. We emphasize that in this technique, we do *not* utilize the submission content to demonstrate the post-to-post graph representativeness.

Table 5 provides the cross-validation results for the different baseline models. We divide the data into the train (80%), validation (10%), and test (10%) sets. All models undergo a validation process, and we report the results for the test set. Combining CREDiBERT with the Graph Convolutional Network (GCN) demonstrates superior performance compared to the baseline models. We compare the results with the binary classification models discussed in the previous section to ensure comparable results. The results underscore more than 9% improvement in the F1 score compared to the BERT-classification model, illustrating the effectiveness of

Model	**Accuracy**
Majority	0.543
Expert Assessment	0.633
S-BERT + Classifier	0.650
BERT Classifier	0.683
CREDiBERT + Classifier (Ours)	$\mathbf{0.775}$

Table 3: Credibility classification results compared with expert assessment for 100 random samples.

the post-to-post network in encoding user interactions with submissions. It also emphasizes that BERT-classifiers struggle with unseen data, while CREDiBERT-based classifiers perform similarly. Remarkably, Node2vec achieves an F1 score of 0.747 without textual embeddings, indicating that the post-to-post network is proficient in encoding user reactions to detect low-credibility submissions.

Unseen Source Credibility Estimation

In this experiment, we show that the proposed framework is not limited to Reddit submissions and can estimate the credibility of unseen sources. We use the binary classifiers developed in the previous section to estimate the credibility score of unseen sources. To this end, we first predict all submission labels referencing the source in question and then estimate the score by the ratio of submissions labeled as credible to all submissions. In other words, we measure the frequency at which a source disseminates non-credible information, which aligns with the definition of credibility provided earlier.

Table 4 shows the results of this experiment. While the S-BERT classifier struggles, the CREDiBERT-based and BERT classifiers can completely distinguish between credible and non-credible sources. The results provided by CREDiBERT show more accuracy and robustness in estimating source credibility scores.

This experiment confirms that the proposed framework can effectively detect the credibility of unverified sources and can be used to detect low-credibility information across Reddit.

Topic-based Susceptibility Analysis

To demonstrate the capabilities of the CREDiBERT, we conduct a case study on the susceptibility of different subreddit communities toward low-credible sources with respect to different topics. By analyzing the credibility of submissions in these communities, we aim to gauge their exposure to low credible information (Mosleh and Rand 2022). Reddit voting system, where users upvote or downvote submissions, offers insights into community responses to different information sources. We use submission scores, which reflect the average of upvotes and downvotes and the diverse submissions covering major events in each subreddit, to determine each community's susceptibility to specific topics. For this study, we identify the major topics discussed in the submissions for all the subreddits (verified and unverified)

Source	Score	CREDi.	BERT	S-BERT
patch.com	0.76	−0.13	−0.16	+0.08
usnews.com	0.71	+0.01	+0.08	+0.20
vice.com	0.60	+0.03	+0.10	+0.26
rt.com	0.48	−0.10	+0.05	+0.31
lifesitenews.com	0.34	−0.06	−0.04	+0.28
spectator.com	0.29	−0.07	+0.09	−0.12

Table 4: Credibility score prediction for six unseen sources. The 'Score' column shows the base credibility score. The columns **CREDi.**, **BERT**, and **S-BERT** show the difference between the original predictions of CREDiBERT, BERT Classifier, and S-BERT, respectively, and the base **Score**.

Model	Acc.	F1
Random	0.505	0.506
Majority	0.543	0.352
Node2vec	0.754	0.747
Binary S-BERT	0.673	0.604
Binary BERT	0.729	0.730
Binary CREDiBERT (*Ours*)	0.771	0.770
CREDiBERT + GCN (*Ours*)	**0.818**	**0.817**

Table 5: Cross-validation report for binary classification results of $4,371$ submissions in three major political subreddits.

posted during January 2022 in five major subreddits and utilize the BERTopic model, a method developed by Grootendorst 2022 for topic modeling.

Let us consider topic h and the set of submissions referencing h posted in subreddit z by $\mathcal{S}_{h,z}$. We measure the exposure of users in subreddit z to low credible information, $\gamma_{h,z}$, by averaging the credibility label of the submissions given by

$$\gamma_{h,z} = \frac{1}{|\mathcal{S}_{h,z}|} \sum_{s \in \mathcal{S}_{h,z}} \gamma_s, \tag{8}$$

where $|\mathcal{S}_{h,z}|$ denote the number of submissions addressing topic h in subreddit z, and $\gamma_s \in \{0, 1\}$ stands for binary credibility label for submission s generated through CREDiBERT. To estimate the reaction of the subreddit user to topic h, we calculate the weighted average of the credibility labels of submission in $\mathcal{S}_{h,z}$ with respect to the submission score, given by

$$\rho_{h,z} = \frac{\sum_{s \in \mathcal{S}_{h,z}} \iota_s \lambda_s}{\sum_{s \in \mathcal{S}_{h,z}} \lambda_s}, \tag{9}$$

where λ_s, and $\rho_{h,z}$ denotes the submission score and reaction score. The reaction score, $\rho_{h,z}$, reflects the credibility of the sources users promoted regarding the topic in question.

Figure 6 illustrates the exposure and reaction scores for six topics across four subreddits: r/Conservative, r/Republican, r/Libertarian, and r/politics. We select topics with a minimum of four submissions in each subreddit and ensure a balanced representation across the political spectrum. To assess a community susceptibility to misinformation, we consider two factors: the reaction score $\rho_{h,z}$, which reflects the credibility of sources favored by users, and the exposure score $\gamma_{h,z}$, reflecting the sources available to the community, predominantly controlled by subreddit moderators. A lower exposure score than a reaction score in a subreddit suggests a preference for more credible sources than those provided by moderators. Therefore, the reaction score ($\rho_{z,h}$) and the difference between the reaction and exposure scores ($\rho_{z,h} - \gamma_{z,h}$) are crucial indicators of a community susceptibility to specific topics.

We observe that subreddits generally identified as right-leaning, such as r/Conservative, show lower exposure scores than those identified as left-leaning. Specifically, in r/Conservative, there is a notable trend: users tend to promote submissions from more credible sources when the topics align with right-leaning perspectives (second row of Figure 6). Conversely, users favor less credible sources more frequently when topics contradict their biases. This pattern highlights the influence of political biases on the selection of information sources within these online communities. However, in order to arrive at any concrete conclusions, conducting a more comprehensive study is necessary.

Subreddit r/politics shows the highest exposure to credible information among the subreddits in this study. However, we observe a trend where users chose to promote sources with lower credibility in discussions on topics like '*Highest inflation rate since 1982*' and '*Supreme Court ruling on vaccine mandate*'. This suggests that, for certain topics, the community's exposure to credible sources does not necessarily prevent the selection of less reliable information. This phenomenon indicates a susceptibility within r/politics to favor less credible sources when discussing specific, perhaps more controversial, topics. While these observations from r/politics and r/conservative do not establish a direct correlation with susceptibility to fake news, they suggest a potential for greater susceptibility to these topics compared to other communities. This is particularly notable in cases where users, despite being exposed to low-credibility sources, predominantly vote in favor of more credible ones.

Discussion

The framework presented in this paper demonstrates a notable efficacy in identifying low-credibility submissions on Reddit, outperforming conventional text classification methods. While currently optimized for Reddit data, its design holds potential for adaptation to other social media platforms and news websites, broadening its applicability. Furthermore, the automated pairing of submissions in CREDiBERT facilitates ongoing model refinement, enabling it to stay current with evolving trends and patterns in online misinformation. Our findings indicate that even in communities with access to reliable sources, there is a tendency to favor less credible information for certain topics. This insight opens up avenues for using CREDiBERT to understand and potentially mitigate topic-specific misinformation.

The weighted post-to-post network offers a novel method to analyze user interactions without delving into personal user data. This method encodes extensive user interactions

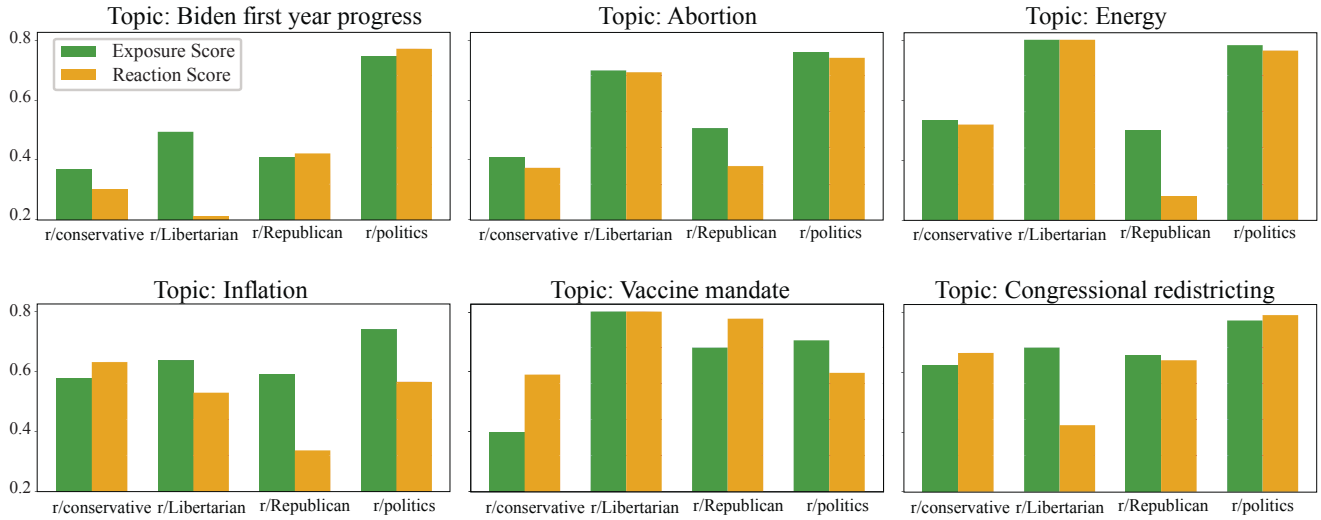

Figure 6: The exposure (green) and reaction (yellow) score of 6 topics in r/Conservative, r/Republican, r/Libertarian, and r/politics. r/politics has the highest exposure score among all subreddits, while r/Conservative and r/Republican have the lowest exposure score. While r/Libertarian shows extreme susceptibility to certain topics for others, it has identical exposure and reaction scores.

in a weighted graph, enabling comprehensive analysis while respecting user privacy. While platforms like Twitter and Facebook present different challenges due to their content structure, adapting the post-to-post methodology could be a promising path forward.

Limitations A major limitation of CREDiBERT lies in its ability to assess only the credibility of news sources rather than the veracity of the content within the articles. This means that while it can effectively identify patterns of misinformation from generally unreliable sources over time, it may not detect false information disseminated through otherwise credible outlets. This gap highlights the challenge of discerning nuanced misinformation.

Unreliable news sources often attempt to mimic the style of credible ones, particularly in their headlines. This presents a second challenge for us: as these distinctions blur over time, text classification can struggle to differentiate between them. However, incorporating a post-to-post network could potentially mitigate this issue. By analyzing users reactions, we can glean additional insights into the credibility of posts, as demonstrated earlier.

Data Biases The current data set is constructed from five major political subreddits. However, nearly 70% of data belongs to r/politics. The analysis indicates that sources referenced to r/politics have a tendency to be left-leaning; thus, the data is skewed toward the left. We emphasize that the extreme right and extreme left have significantly lower credibility scores than unbiased media outlets, depicted in Figure 1. The data in Figure 1 shows that the left-leaning sources has a higher credibility score compared to the right-leaning ones. The implications of this skew are best demonstrated in Figure 6. While subjects are chosen to be equally favored by both sides of the political spectrum, the right-leaning subreddits generally show a lower credibility score, creating a perception that right-leaning has a tendency to engage with low-credible sources more often. While the credibility score of submission sources can be used to estimate the average susceptibility of the communities, it is essential to consider other factors, such as engagement with such content, to make any conclusion about the susceptibility of the community.

Conclusion

In this study, we introduced CREDiBERT, an innovative sentence-level embedding model designed to assess credibility in Reddit submissions. According to our evaluations, CREDiBERT-generated embeddings can outperform existing text classification models over the credibility assessment task. We also developed a weighted post-to-post network to encode Reddit user interactions efficiently without requiring user profiling. When integrated with CREDiBERT, this network enhances the detection of credible sources. By applying CREDiBERT to recent Reddit submissions, we have revealed its capability to estimate community susceptibility to low-credible information on various topics. Exploring the application of CREDiBERT in other social media contexts and refining its methodology to address its current limitations present exciting avenues for future research.

References

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2): 211–236.

Bachmann, P.; Eisenegger, M.; and Ingenhoff, D. 2022. Defining and measuring news media quality: Comparing the

content perspective and the audience perspective. *The International Journal of Press/Politics*, 27(1): 9–37.

Bond, R. M.; and Garrett, R. K. 2023. Engagement with fact-checked posts on Reddit. *PNAS nexus*, 2(3): pgad018.

Castillo, C.; Mendoza, M.; and Poblete, B. 2013. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5): 560–588.

Chan, M.-p. S.; Jones, C. R.; Hall Jamieson, K.; and Albarracín, D. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11): 1531–1546.

Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3): 354–368.

Chiang, T. H.; Liao, C.-S.; and Wang, W.-C. 2022. Investigating the Difference of Fake News Source Credibility Recognition between ANN and BERT Algorithms in Artificial Intelligence. *Applied Sciences*, 12(15): 7725.

Chipidza, W.; Krewson, C.; Gatto, N.; Akbaripourdibazar, E.; and Gwanzura, T. 2022. Ideological variation in preferred content and source credibility on Reddit during the COVID-19 pandemic. *Big Data & Society*, 9(1): 20539517221076486.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Glenski, M.; Weninger, T.; and Volkova, S. 2018. Identifying and understanding user reactions to deceptive and trusted social news sources. *arXiv preprint arXiv:1805.12032*.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*, 855–864.

Hocevar, K. P.; Metzger, M.; and Flanagin, A. J. 2017. Source credibility, expertise, and trust in health and risk messaging. In *Oxford Research Encyclopedia of Communication*.

Horne, B.; Khedr, S.; and Adali, S. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Horne, B. D.; Nørregaard, J.; and Adali, S. 2019. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(1): 1–23.

Hurtado, S.; Ray, P.; and Marculescu, R. 2019. Bot detection in reddit political discussion. In *Fourth International Workshop on Social Sensing*, 30–35.

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

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

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

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mosleh, M.; and Rand, D. G. 2022. Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, 13(1): 7144.

Pehlivanoglu, D.; Lin, T.; Deceus, F.; Heemskerk, A.; Ebner, N. C.; and Cahill, B. S. 2021. The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications*, 6(1): 1–12.

Przybyla, P. 2020. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 490–497.

Raza, S.; and Ding, C. 2022. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4): 335–362.

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

Sakketou, F.; Plepi, J.; Cervero, R.; Geiss, H.-J.; Rosso, P.; and Flek, L. 2022. Factoid: A new dataset for identifying misinformation spreaders and political bias. *arXiv preprint arXiv:2205.06181*.

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

Spezzano, F.; Shrestha, A.; Fails, J.; and Stone, B. 2021. That's fake news! Investigating how readers identify the reliability of news when provided title, image, source bias, and full articles. *ACM, Human Computer Interaction journal*, 5.

Thakur, N.; Reimers, N.; Daxenberger, J.; and Gurevych, I. 2020. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.

Torabi Asr, F.; and Taboada, M. 2019. Big Data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1): 2053951719843310.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures?
   Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? Yes.
   (g) Did you discuss any potential misuse of your work? Yes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we have uploaded the CREDiBERT model, the data set we employed for training it, and the training and test data set in an anonymous Google drive. Codes are available in the Appendix B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, however, we address our method limitation and blindness toward individual news articles varsity in Discussion and Introduction.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes.
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, because all the data we use are publicly available, and we did not creat a new data set.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes please see Ethics Statement and Data
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR ? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Ethics Statement

The research presented in this paper strictly utilizes publicly available data from Reddit. In line with ethical standards, our

methodology prioritizes user anonymity; we do not track or profile any users. The post-to-post framework is designed to analyze user interactions while fully respecting user privacy, avoiding the inclusion of any personally identifiable information. It is crucial to address the ethical concerns related to the data used to evaluate credibility, specifically the bias-credibility curve we use throughout this paper. A skewed curve could substantially affect the definition of credibility and favor one party over another as non-credible. Therefore, we recommend users of CREDiBERT to disclose this information along with the model.

We recognize the potential for misusing the proposed framework. For example, sources with low credibility might attempt to leverage CREDiBERT to modify their content, making it appear more credible. To mitigate such risks, we propose a set of specific ethical guidelines for the use of CREDiBERT and similar tools.

Transparency: Users of CREDiBERT-like tools should be required to publish transparency reports that detail their use of the tools, the nature of the data processed, and the purposes for which it is used.

Independent audit: We advocate for regular independent audits of these tools to ensure that the predictions and operations remain unbiased.

Stakeholder engagement: Developing the guidelines should involve a broad spectrum of stakeholders from both parties. This would also motivate such organizations to increase their credibility and reputation.

Appendix A

Evaluation Methods

Word2Vec Word2Vec was developed as a technique to estimate the vector representation of words efficiently (Mikolov et al. 2013). We compute the average word embeddings of the submission text and then feed the embedding to a classification layer to assess the credibility of the submissions.

BERT-based Text Classifier The BERT-based text classifiers fine-tune the underlying transformer for the classification task and utilize the first token as the aggregate sentence representation for classification, allowing them to over-perform naive classification of average embeddings pooling generated by BERT-based models.

Sentence Transformers The S-BERT model was developed to address the challenges the BERT-based model faces in sentence-level embedding in sentence comparison tasks. We integrate the sentence-level embedding with two different classification models for better comparison. We emphasize that while the CREDiBERT is primarily created for classification tasks, the model is trained to estimate Credibility score discrepancies; thus, it is considered a sentence transformer model. For the first approach, we integrate sentence-level embeddings with a classification layer for binary classification.

Node2vec Node2vec, introduced by Grover and Leskovecis (2016), is a graph embedding technique. It effectively represents graph nodes as vectors in a latent space, capturing both the structural characteristics and homophilic tendencies of the graph. We pass the graph embeddings to a classifier layer for binary classification of the submissions. We emphasize that in this technique, we do *not* utilize the submission content to demonstrate the post-to-post graph representativeness.

Appendix B

Code and Data Access

The code and data for training and evaluating CREDiBERT are accessible through:
https://drive.google.com/drive/folders/1hxA5eaz-zz88ftq6RbvS73DODl69YnbU?usp=sharing