

Elevating Cyber Threat Intelligence against Disinformation Campaigns with LLM-based Concept Extraction and the FakeCTI Dataset

Domenico Cotroneo^a, Roberto Natella^a, Vittorio Orbinato^a

^aDIETI, Università degli Studi di Napoli Federico II, via Claudio 21, Naples, 80125, Naples, Italy

Abstract

The swift spread of fake news and disinformation campaigns poses a significant threat to public trust, political stability, and cybersecurity. Traditional Cyber Threat Intelligence (CTI) approaches, which rely on low-level indicators such as domain names and social media handles, are easily evaded by adversaries who frequently modify their online infrastructure. To address these limitations, we introduce a novel CTI framework that focuses on high-level, semantic indicators derived from recurrent narratives and relationships of disinformation campaigns. Our approach extracts structured CTI indicators from unstructured disinformation content, capturing key entities and their contextual dependencies within fake news using Large Language Models (LLMs). We further introduce FakeCTI, the first dataset that systematically links fake news to disinformation campaigns and threat actors. To evaluate the effectiveness of our CTI framework, we analyze multiple fake news attribution techniques, spanning from traditional Natural Language Processing (NLP) to fine-tuned LLMs. This work shifts the focus from low-level artifacts to persistent conceptual structures, establishing a scalable and adaptive approach to tracking and countering disinformation campaigns.

Keywords: Cyber Threat Intelligence, Disinformation, Fake News, Large Language Models

1. Introduction

In recent years, the proliferation of fake news has emerged as a critical threat to global information integrity. Fabricated with misleading and false content, fake news is designed to misinform, manipulate, and deceive audiences to serve specific agendas [1, 2, 3]. Malicious actors craft these narratives to influence public discourse, sway political decisions, and disrupt societal harmony. The widespread adoption of social media platforms has further aggravated this issue, enabling rapid dissemination of disinformation at an unprecedented scale [4, 5, 6, 7]. As a result, distinguishing between authentic and deceptive content has become increasingly challenging.

To combat this growing threat, security analysts leverage *Cyber Threat Intelligence* (CTI) to track and mitigate disinformation campaigns. This CTI includes links to web articles and social media posts classified by analysts as fake content, along with the domains and social media accounts that created them. However, this kind of CTI is affected by the same issues as the “traditional” CTI used for network attacks, which includes web domains, IP addresses, and hostnames. Attackers can easily modify their attacks to make them unrecognizable by CTI, e.g., by buying new domains. The same logic applies to disinformation and fake news, where attackers can create new websites or social media accounts to evade detection.

This issue is highlighted by the *Pyramid of Pain* [8], a framework designed for network-based CTI that classifies low-level indicators as trivial to bypass, while higher-level intelligence is significantly harder for adversaries to evade. Figure 1 depicts how we adapt the Pyramid of Pain to the disinformation landscape, illustrating how different types of intelligence vary

in their persistence and resistance to manipulation. To be effective, CTI for fake news must move beyond easily modifiable artifacts and capture persistent elements of disinformation operations. Our adaptation of the Pyramid of Pain for disinformation shifts the focus from technical artifacts to content-based indicators. While traditional CTI relies on low-level indicators such as IP addresses, domains, and hashes, which adversaries can rapidly modify, disinformation campaigns similarly exploit website domains and social media handles that are easily replaced. To develop a more resilient strategy, our model shifts the focus to persistent narrative structures and behavioral patterns that support disinformation operations. Unlike network artifacts, core narratives cannot be easily rewritten, as they structure the fundamental viewpoint promoted by threat actors.

This work aims to extract the fundamental elements that characterize a fake news campaign, in order to recognize new fake content from the same attacker in a way that is robust to textual modifications and variations of domains and accounts. To address the limitations of traditional CTI approaches in tackling disinformation, we introduce a concept-based CTI framework that captures the narratives and relationships within disinformation campaigns. Unlike conventional CTI indicators, which primarily rely on low-level technical artifacts, our methodology extracts high-level, content-based intelligence that persists even as threat actors modify superficial elements of their campaigns. Our approach is based on the structured representation of disinformation narratives through tuples that encapsulate the key entities, their relationships, and contextual dependencies within fake news articles. These struc-

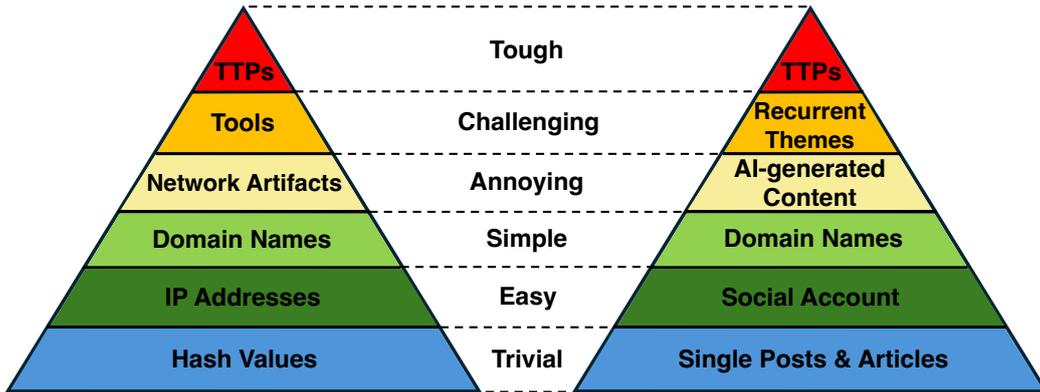


Figure 1: Traditional and disinformation-oriented Pyramids of Pain.

tured representations allow for more robust attribution and variant detection, as they focus on the core meaning and intent rather than specific linguistic expressions. We employ Large Language Models (LLMs) to automate the extraction and analysis of these structured indicators, effectively converting unstructured narratives into structured intelligence. In addition, we analyze multiple analytical techniques, ranging from lexical and semantic similarity methods to fine-tuned LLMs, to recognize new variants of fake news that align with existing disinformation campaigns, even when the language and presentation of the content have been altered.

The primary contributions of this study are as follows:

1. *Concept-based CTI indicators extraction*: we propose a novel methodology to extract structured CTI indicators that characterize fake news narratives through key entities, relationships, and their contextual connections. This structured representation enables robust and adaptable attribution by focusing on thematic and relational consistency. By encoding the core meaning of fake news content, these indicators enable the detection of new campaign variants and the association of evolving disinformation efforts with previously identified attack groups. The extracted intelligence is represented as structured tuples that capture the essential actions and relationships in disinformation narratives, ensuring a scalable and consistent format for further analysis.
2. *FakeCTI dataset*: to facilitate research in this domain, we introduce FakeCTI, the first dataset that systematically links fake news articles to known disinformation campaigns and threat actors. FakeCTI includes 12,155 articles from 43 distinct campaigns, each annotated with metadata specifying the associated campaign, threat actor, and dissemination medium. FakeCTI allows researchers and practitioners to analyze disinformation trends, develop automated detection systems, and evaluate attribution methodologies at scale, offering a structured foundation for attribution studies.
3. *Experimental evaluation of fake news attribution techniques*: using the FakeCTI dataset, we assess multiple techniques for attributing fake news articles to their originating disinformation campaigns. Our evaluation spans

several approaches, achieving accuracy in fake news attribution up to 94% (using LLMs), highlighting how well different techniques generalize across disinformation content and their capability to identify new and evolving fake news narratives.

This study demonstrates how concept-based CTI indicators can help improving disinformation detection, attribution, and prevention, offering a more effective approach to countering manipulated narratives and adversarial disinformation strategies. We publicly share our artifacts¹ to encourage further research on this topic.

The remainder of the paper is structured as follows. Section 2 discusses related work; Section 3 presents the methodology for the extraction of concept-based CTI indicators; Section 4 introduces the FakeCTI dataset; Section 5, 6 illustrate the experimental analysis on the indicators extraction and fake news attribution; Section 7 concludes the paper.

2. Background and Related Work

2.1. Fake News Detection

Sharma *et al.* [9] performed a comprehensive analysis of fake news detection and mitigation techniques, categorizing existing methodologies into three main approaches: content-based identification, feedback-based identification, and intervention-based solutions.

Content-based identification. Content-based fake news detection techniques analyze the text and linguistic features of news articles to extract distinguishing characteristics between fake and true news and identify deceptive patterns [10, 11, 12]. These methods include linguistic-based techniques, which try to identify deception through stylistic markers and factual inconsistencies, and deep learning-based solutions, which aim to extract complex textual patterns without manual feature engineering. However, adversaries can evade this kind of detection by rewording or slightly modifying content. Simple paraphrasing or reformatting headlines can bypass content-based classifiers, making such models ineffective against large-scale disinformation campaigns.

¹GitHub repo

Feedback-based identification. Feedback-based methods leverage user interactions, propagation patterns, and engagement metrics on social media to detect fake news. These strategies include propagation pattern, temporal pattern, and user response and stance analysis, which assume that the way information spreads can indicate its veracity [13, 14, 15]. A significant limitation of many existing approaches is their focus on single social media platforms such as Twitter/X or Facebook. Consequently, adversaries easily evade detection by shifting narratives across different ecosystems, such as migrating from one platform to another, e.g., from Twitter/X to Telegram or Reddit, when content moderation policies increase. Strengthening detection requires integrating multi-platform tracking systems that correlate narratives, domain registrations, and user interactions across different platforms.

Intervention-based solutions. Intervention-based solutions represent mitigation strategies to prevent the spread of disinformation. These approaches include proactive disinformation interventions, such as fact-checking bots, which analyze claims against verified knowledge bases to disprove false information, credibility scores assigned to sources, content moderation, consisting of machine learning-based techniques to reduce the reach of fake news by limiting algorithmic amplification, and user awareness and debunking strategies [16, 17, 18]. However, fact-checking and debunking approaches are typically slow, while content moderation strategies must balance accuracy with freedom of speech concerns.

As highlighted above, most existing detection techniques focus on identifiable artifacts such as suspicious domain registrations, automated social media accounts, or content propagation anomalies. While these features provide valuable clues about the spread of disinformation, they remain reactive indicators that adversaries can quickly replace or modify. Indeed, adversaries can easily generate new social media accounts or register new domains once previous ones are flagged. To effectively disrupt disinformation, cybersecurity teams must analyze the TTPs used by adversaries, such as narrative amplification techniques, AI-generated content, and cross-platform coordination tactics.

Another significant shortcoming is the lack of attribution and integration with traditional CTI workflows. While frameworks like STIX [19, 20] and TAXII [21] enable structured CTI sharing for cyber threats, disinformation analysis remains fragmented. Current approaches, such as fact-checking databases and disinformation-debunking sites, operate in isolation from security operations centers (SOCs) and intelligence-sharing platforms, e.g., MISP [22]. Detection and attribution efforts should focus on the highest levels of the Pyramid of Pain, *Recurrent Themes* and *TTPs*, where adversaries rely on known strategies but change execution methods. To strengthen attribution, CTI workflows must incorporate standardized frameworks, such as DISARM, that map disinformation actors and TTPs into structured intelligence feeds. By aligning disinformation intelligence with standardized cybersecurity methodologies, organizations can develop multi-layered defense mechanisms, allowing for real-time intelligence-sharing across governments, tech platforms, and fact-checking networks.

2.2. Disinformation-oriented CTI

In addition to traditional fake news identification and mitigation techniques, the *DISARM* (Disinformation Analysis and Response Measures) framework [23] has been proposed as a standardized approach for integrating disinformation-oriented CTI. *DISARM* provides a structured approach to analyzing and countering online disinformation by categorizing threat actors, attack techniques, and mitigation strategies similar to established CTI frameworks like MITRE ATT&CK [24]. The framework aims to enhance the interoperability of threat intelligence across organizations, social media platforms, and cybersecurity agencies by defining a common taxonomy for disinformation campaigns. Key elements of *DISARM* include actor profiling (e.g., state-sponsored campaigns, cybercriminals, ideological groups), tactics (e.g., narrative amplification, deepfake propagation, social engineering), and countermeasures (e.g., AI-driven detection, cross-platform attribution, content moderation). Unlike traditional fake news detection, which often focuses on individual articles or social media posts, *DISARM* treats disinformation as a systemic cyber threat requiring multi-layered intelligence sharing and automated threat hunting. However, challenges remain in operationalizing the framework, particularly in real-time attribution of adversarial actors and alignment of disinformation CTI with existing cybersecurity workflows.

To this end, Gonzalez *et al.* [25] proposed *DISINFOX*, an open-source framework to integrate disinformation intelligence into existing CTI workflows, leveraging *DISARM* to model disinformation TTPs and translate them into STIX objects for structured representation and interoperability. *DISINFOX* facilitates the storage, management, and exchange of disinformation incidents, providing integration with OpenCTI [26]. The framework was validated using a dataset of 118 real-world disinformation incidents, demonstrating its feasibility for structured intelligence-sharing and analysis. However, challenges such as manual annotation, limited automation, and lack of standardized evaluation frameworks may hinder its large-scale operationalization.

Despite its structured approach, *DISARM* is not yet ready for practical deployment due to critical limitations. Limited adoption and institutional backing prevent its widespread integration into cybersecurity workflows, as most CTI platforms are not yet configured to handle disinformation intelligence. Moreover, *DISARM* lacks standardized evaluation frameworks to assess the impact and severity of disinformation incidents, unlike traditional CTI which benefits from well-established risk-scoring models, e.g., CVSS for vulnerabilities [27]. Determining the virality, influence, and effectiveness of a disinformation campaign is highly context-dependent, making it difficult to assign threat levels unequivocally.

Our solution, focused on extracting and analyzing structured concept-based CTI, aligns with the *Recurrent Themes* level in the disinformation-oriented Pyramid of Pain, the highest abstraction layer just below TTPs. This reflects the persistent nature of themes and narratives across campaign variants, central to attribution and detection. While TTPs remain highly impact-

ful, their standardization is still limited, as highlighted by the challenges faced in operationalizing DISARM.

2.3. Fake News Datasets

The landscape of fake news analysis has evolved significantly with the availability of structured datasets designed for training machine learning models for disinformation detection. While several datasets provide valuable insights into fake news detection, their applicability to fake news attribution, especially concerning disinformation campaigns or threat actors, remains limited.

Fakeddit [28] is a large-scale, multimodal dataset containing over 1 million samples collected from Reddit, a social media platform where fake news, conspiracy theories, and propaganda often circulate. It is one of the most extensive publicly available datasets for fake news detection, providing textual and visual features alongside metadata and user engagement data. Unlike many datasets that offer binary classifications (true/false), *Fakeddit* employs a more granular labeling system, allowing researchers to categorize fake news into six different classes: true, mostly true, half true, half false, mostly false, and false. Despite its comprehensiveness, the dataset does not explicitly connect disinformation to specific threat actors or coordinated disinformation campaigns, nor does it provide insights into the intent behind the disinformation.

IFND [29] is a multimodal dataset that focuses on Indian fake news, covering news articles spanning from 2013 to 2021. The dataset contains both true and false news, allowing researchers to explore the characteristics of disinformation within an Indian context. A distinctive feature of *IFND* is its use of Latent Dirichlet Allocation (LDA) topic modeling, which helps classify news articles into different topical categories. However, while the dataset enables fine-grained classification and content-based fake news detection, it does not provide any attribution to threat actors or coordinated influence campaigns. The dataset lacks metadata that could link fake news articles to disinformation networks.

The *LIAR* dataset [30] is one of the most widely used benchmark datasets for fake news detection and fact-checking. It consists of 12,836 manually labeled short statements collected from Politifact [31], a Pulitzer Prize-winning fact-checking organization. Unlike *Fakeddit* and *IFND*, *LIAR* is specifically designed to analyze political disinformation: it includes statements made by politicians, government officials, and public figures, along with detailed justifications for each truthfulness label. Each statement in *LIAR* is classified into six fine-grained categories: pants-on-fire (completely false), false, barely true, half-true, mostly true, and true. While *LIAR* provides a structured approach to analyzing political disinformation, it does not capture broader disinformation campaigns or any organized influence operations.

The *PushShift Telegram* dataset [32] is unique since it focuses on disinformation spread through Telegram, a messaging platform known for hosting extremist groups, conspiracy theorists, and disinformation networks. Unlike previous datasets, which collect news articles and political statements,

PushShift Telegram comprises messages from 27,800+ Telegram channels, capturing real-time discussions, propaganda, and conspiracy-driven content. This dataset is particularly relevant for studying coordinated disinformation efforts, as many Telegram groups serve as hubs for extremist and disinformation networks, providing valuable insight into how disinformation spreads within closed communities. Although it includes data from channels linked to disinformation campaigns, *PushShift Telegram* does not provide direct attribution to specific threat actors. Without metadata linking the groups to nation-state actors, troll farms, or intelligence operations, it remains challenging to draw definitive conclusions about the origins and intent behind the disinformation.

The *PushShift Reddit* dataset [33], like the Telegram version, provides a large-scale collection of Reddit discussions and posts. The dataset contains millions of comments and posts from Reddit users. Unlike *Fakeddit*, which focuses on structured fake news classification, *PushShift Reddit* captures unfiltered, real-time conversations, making it valuable for studying disinformation narratives and how they evolve. While the dataset includes discussions on conspiracies, political propaganda, and hoaxes, it does not explicitly label fake and true news. Similar to *PushShift Telegram*, the dataset is useful for studying how disinformation spreads, but it does not provide direct links between disinformation campaigns and specific threat actors.

BuzzFace [34] is a comprehensive and large-scale dataset designed to study fake news dissemination and user engagement on Facebook. It includes 2,282 annotated news articles spanning mainstream, left-leaning, and right-leaning news sources, providing a diverse representation of disinformation. The dataset categorizes news into four veracity levels: mostly true, mostly false, a mixture of true and false, and no factual content. *BuzzFace* also includes metadata such as engagement metrics, user comments, and reactions. Although the dataset captures patterns of interaction and possible social bot activity, it lacks detailed attribution metadata, such as information on who created or amplified the disinformation.

Table 1 summarizes the comparison between these existing datasets. It is possible to notice how most datasets lack explicit connections to threat actors or coordinated disinformation campaigns, making them unsuitable for attribution. *Fakeddit* and *IFND*, despite their large-scale nature, focus on fake news classification rather than tracing its origin, while *LIAR* is limited to fact-checked political statements, offering no insights into organized influence operations. *BuzzFace*, which integrates Facebook user engagement, provides valuable data on misinformation spread but lacks source tracking or disinformation network links. Among them, *PushShift Telegram* is the most relevant for threat actor association, as it captures propaganda narratives and manipulated content from extremist groups, making it a potential resource for analyzing coordinated disinformation efforts. However, it still requires external intelligence sources to attribute fake news to specific actors definitively. *PushShift Reddit*, though capturing discussions where misinformation evolves, does not explicitly tag or track coordinated campaigns, making its association with threat ac-

Table 1: Comparison of existing fake news datasets.

Dataset	Samples	Source	Threat Actor Association	Usability for Attribution
Fakeddit	1,063,106	Reddit	No	Low
IFND	Unspecified	Indian news websites	No	Low
LIAR	12,836	Politifact	No	Low
PushShift Telegram	Unspecified	Telegram	Yes	Moderate
PushShift Reddit	Unspecified	Reddit	Possible	Moderate
BuzzFace	2,282 articles, 1.6M comments	Facebook	No	Low
FakeCTI	12,155 articles, 43 disinformation campaigns	News websites, Social networks	Yes	High

tors speculative. In terms of usability for fake news attribution, only PushShift Telegram and, to a lesser extent, PushShift Reddit, hold moderate potential. On the contrary, Fakeddit, IFND, LIAR, and BuzzFace are better suited for fake news detection and classification rather than identifying who creates or disseminates misinformation. While these datasets provide valuable insights into how misinformation spreads, they lack critical metadata for direct attribution, such as source credibility, bot network activity, or links to known disinformation operations. PushShift Telegram remains the most promising but requires additional forensic analysis and intelligence sources to confirm attribution. Effective fake news attribution demands more than just classification, it requires datasets that track source origins, coordinated amplification, and intent, none of which are fully addressed by these datasets.

Conversely, the proposed dataset, i.e., *FakeCTI*, offers a significant advantage over existing fake news datasets by systematically linking fake news articles to their corresponding disinformation campaigns and threat actors. While prior datasets primarily focus on binary or multi-class fake news classification, labeling articles as true, false, or misleading, they lack explicit connections to the broader disinformation strategies that spread these narratives. FakeCTI provides structured attribution metadata, allowing researchers to trace the origin and intent of disinformation campaigns. This linkage is crucial for effective fake news attribution, as it allows the detection of new campaign variants and the identification of recurring disinformation strategies across multiple sources.

3. Methodology

Our work aims to define and extract concept-based Cyber Threat Intelligence (CTI) indicators to enhance the attribution and detection of fake news articles linked to disinformation campaigns. Figure 2 illustrates how our proposal is integrated into the disinformation landscape. Disinformation campaigns are orchestrated efforts by threat actors (①) to spread false or misleading information to manipulate public opinion, influence political processes, or discredit individuals or institutions. These actors range from state-sponsored groups to independent propagandists and malicious organizations, leveraging various digital platforms to amplify their narratives. The fake news ar-

ticles they produce form the foundation of these campaigns and are carefully crafted to blend with legitimate news, making detection and attribution increasingly challenging. These articles and social media posts are variants of the same fake content, spread using various web domains and social network accounts and crafted by rephrasing content.

Once published, these fake news articles become part of larger disinformation campaigns (②), each characterized by distinct themes, narratives, and objectives. For example, a campaign may focus on public health disinformation, promoting anti-vaccine content, while another targets political elections, spreading conspiracy theories about electoral fraud. Establishing clear associations between individual fake news articles and their respective campaigns is crucial for attribution. However, this step remains difficult due to the lack of structured, high-level intelligence indicators that describe the narrative and intent behind the disinformation.

Traditional CTI approaches rely on low-level IoCs such as IP addresses, domain names, and email addresses. While these indicators can help track technical infrastructure, they do not capture the semantic and conceptual features of disinformation. This gap necessitates the development of higher-level, content-driven CTI indicators (③) that can characterize fake news articles based on their key claims, entities, and relationships rather than ephemeral technical artifacts. To address this challenge, we introduce a new format of concept-based CTI indicators derived from the textual content of fake news articles. Unlike traditional IoCs, these indicators remain persistent across different campaigns and dissemination channels, making them more effective for long-term tracking and analysis.

Once novel CTI indicators have been defined, the next step is their automatic extraction from newly encountered fake news articles (④). These structured intelligence indicators can help categorize, attribute, and track disinformation campaigns.

The extracted CTI indicators serve multiple purposes (⑤), primarily in the following areas:

- *Detection*: concept-based CTI indicators enable the identification of new fake news articles that share similarities with known disinformation narratives. This allows early intervention before misleading information spreads widely.

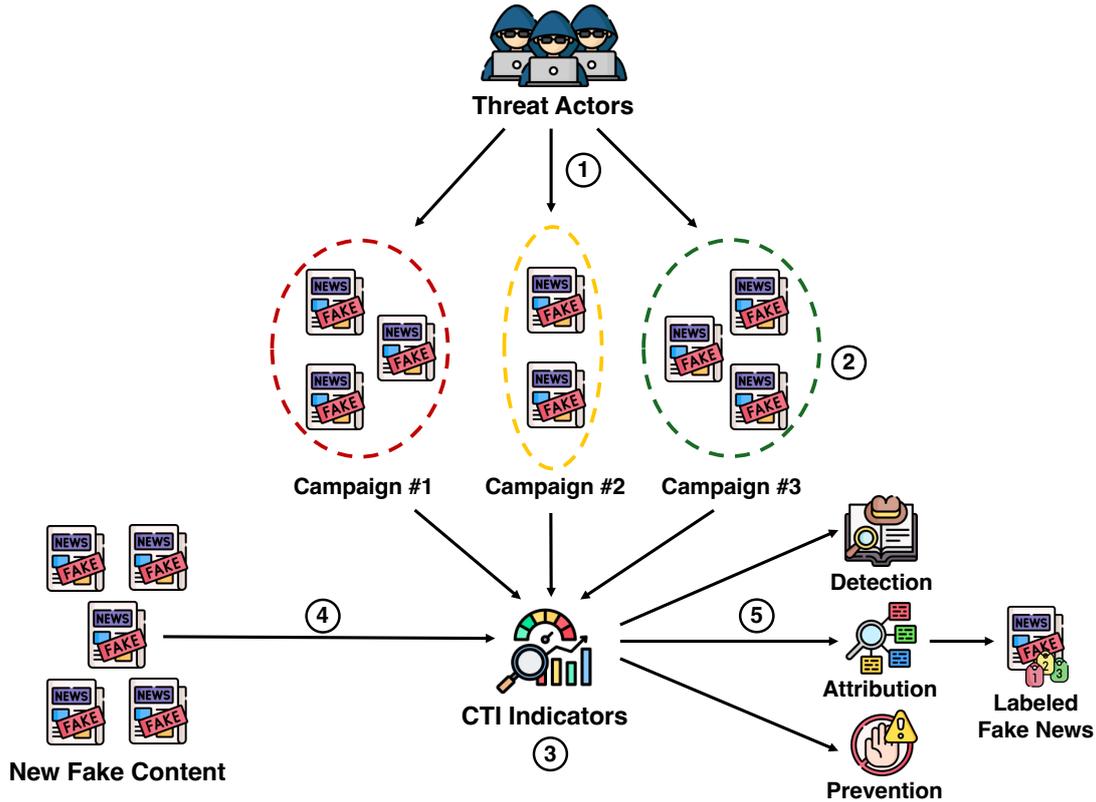


Figure 2: Overview of the proposal.

- *Attribution*: by linking extracted indicators to specific disinformation campaigns, we can trace which actors are responsible for spreading certain types of misinformation, improving situational awareness.
- *Prevention*: identifying recurring patterns in disinformation helps in designing countermeasures to mitigate the impact of future campaigns, such as improving fact-checking mechanisms and enhancing automated content moderation.
- *Suitable for supporting fake news detection*: CTI should be effective in guiding the detection of additional variants of fake news, enabling the identification of disinformation campaigns based on recognized patterns.
- *Lightweight*: CTI should be quick to share and easy to store and process, ensuring efficient dissemination among stakeholders and minimal resource demand, in order to scale well for large volumes of data to analyze.

3.1. Concept-based CTI for fake news

We propose a novel approach to structuring CTI indicators specifically tailored for fake news attribution. We focus on extracting meaningful, content-based indicators by analyzing the core claims, narratives, and references embedded in fake news articles. These indicators remain stable even as superficial details change, making them more effective for long-term attribution. Such indicators should meet the following properties:

- *Non-volatile*: CTI indicators should reflect the contents of a disinformation campaign, without relying on technical artifacts that can be easily changed (domain names, social media accounts), and robust against variants of the same campaign.
- *Interpretable*: CTI must be easily understood by both humans and automated systems, facilitating swift analysis and response.
- *Entities*: key actors, organizations, events, or concepts mentioned in the article.
- *Relations*: the semantic connections between these entities. These relationships provide contextual meaning to the extracted entities.
- *Objects*: the targets or outcomes associated with the entities and relations.

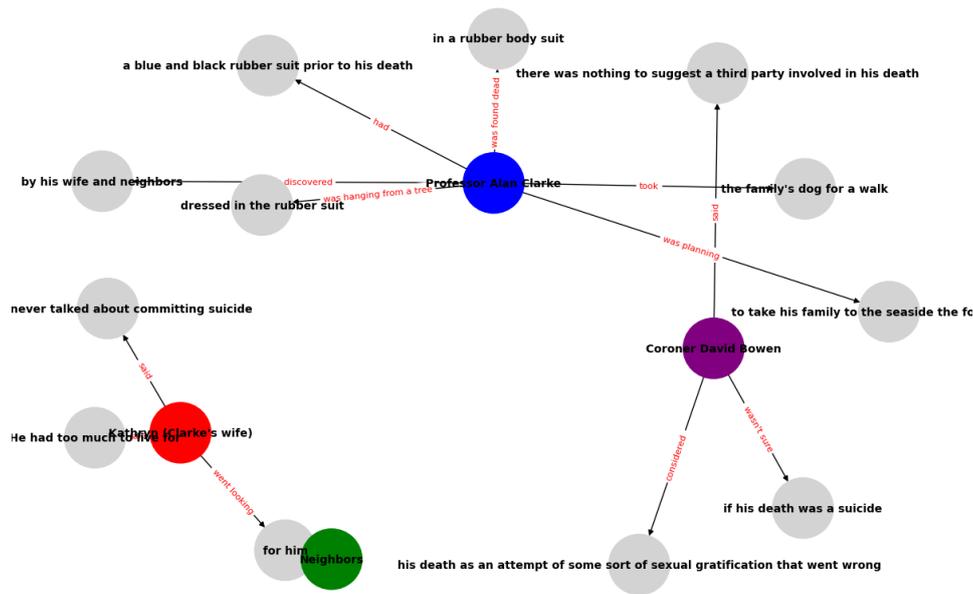


Figure 3: Example of graph extracted from a fake news article.

This information is then grouped into $\langle \text{subject}, \text{relation}, \text{object} \rangle$ triples. In the following, we will also refer to such triples as *tuples*. For example, the sentence “Country X funds Organization Y to spread misinformation” is transformed into the triple $\langle \text{Country X}, \text{funds}, \text{Organization Y} \rangle$. By capturing these core elements, we reduce the complexity of the text while preserving its critical meaning. The primary purposes of this process are i) to structure the unorganized, narrative-based content of fake news into a consistent format, ii) to enable the comparison of content across articles by focusing on entities and relationships, ignoring superficial variations in wording or phrasing, and iii) to identify recurring narratives, themes, and strategies within and across disinformation campaigns.

The use of such triples provides several advantages for attribution tasks. One key benefit is their **non-volatile** nature. Unlike traditional indicators of compromise (IoCs), which threat actors can easily alter, triples focus on the underlying content and intent of a disinformation campaign. The narrative or relationships between entities often remain consistent even as surface-level details change. For instance, while a campaign may migrate from one website to another, the core narrative captured by a triple like $\langle \text{Organization X}, \text{spreads}, \text{anti-vaccine propaganda} \rangle$ is likely to persist.

Another significant advantage is **high interpretability**. Triples offer a structured representation of information that is easily understandable for both human analysts and automated systems. A triple such as $\langle \text{Organization X}, \text{collaborates with}, \text{Country Y} \rangle$ directly conveys the essence of the information, facilitating faster decision-making and more effective response planning.

Additionally, triples enable **relationship identification** through mapping into a **relationship graph**, where nodes rep-

resent entities and edges define relationships. Such graphs allow for advanced analysis by highlighting recurring connections between entities, such as a particular actor repeatedly spreading similar narratives. They also reveal clusters of related campaigns or narratives, providing insights into potential coordination between different disinformation sources.

Figure 3 illustrates an example of a graph extracted from a fake news article. Specifically, this example refers to an article discussing the mysterious death of Professor Alan Clarke [35], a prominent UK cancer scientist and director of the European Cancer Stem Cell Research Institute. The discovery was made by his wife, Kathryn, and neighbors, who went looking for him after he failed to return from walking the family dog. The coroner, David Bowen, ruled out the involvement of a third party and was uncertain if Clarke’s death was a suicide. Instead, Bowen speculated it might have been an attempt at sexual gratification that resulted in misadventure. The narrative ties into a disinformation campaign attributed to the threat actor *Elizabeth Erin*, aiming to fuel distrust in the medical community by falsely linking the deaths of medical professionals to their work on cancer discoveries. The article additionally hints at broader conspiratorial undertones by referencing over 30 “mysterious deaths” of prominent doctors, aligning with a known disinformation campaign titled *Doctors Found Dead After Cancer Discovery*. It frames the situation with speculative and sensational details, such as pointing out Clarke’s outfit and suggesting ulterior motives behind the coroner’s public release of intimate information about the death. The graph visualizes the relationships and facts presented in the article, connecting key subjects, actions, and objects such as Clarke’s attire, the coroner’s verdict, and statements from witnesses.

3.2. Tuples Extraction

To automatically extract the tuples from a given fake news article, we leverage Large Language Models (LLMs). LLMs represent the most recent advancement in state-of-the-art relation extraction from text [36, 37, 38], making them the optimal choice for structuring information from unstructured sources like fake news articles. The effectiveness of LLM-based tuple extraction largely depends on the prompt structure and content. In our approach, we carefully design the prompt to provide clear and structured instructions, following best practices to ensure high-quality tuple extraction. The prompt includes a role description along with an explicit task description, according to prompt engineering best practices [39, 40], guiding the model to focus on the extraction of `<subject, relation, object>` triples while avoiding irrelevant details. Additionally, we define step-by-step extraction criteria to enhance consistency and reduce ambiguity in tuple formation. To improve interpretability, we include an example within the prompt, demonstrating the expected input-output structure. Finally, we append a termination keyword (END LIST) to ensure that the model correctly delimits its output, preventing unnecessary hallucination or extraneous text generation. This structured prompting methodology is designed to be model-agnostic, meaning it can be applied to any LLM capable of relation extraction. By following a clear role definition, explicit task description, structured step-by-step criteria, and a termination keyword, the prompt ensures consistency and accuracy across different models. Whether used with open-source or fine-tuned models, this approach remains adaptable, making it versatile for extracting structured information from unstructured text sources [41, 42].

Role: You are a natural language processing expert specialized in analyzing textual data and extracting structured information. Your task is to identify subject-relation-object relationships from the input text, which represent key actions and relationships between entities. These triples will be used for structured representation of fake news articles.

Context: Subject-relation-object relationships capture the fundamental structure of actions and relationships described in a sentence. In these relationships, the subject represents the entity performing an action, the verb describes the action or relationship, and the object represents the entity affected by the action. Identifying these relationships helps in organizing unstructured textual data into a structured format, enabling easier analysis and interpretation.

Example: Text: “John gave a book to Mary.”

 Tuple: John - gave - a book to Mary

Instructions: Read the following text and identify all the tuples in the subject-verb-object form. The tuples should reflect the main actions and relationships between the entities mentioned in the text. Follow these steps:

1. Identify the subject of the action.
2. Identify the verb that describes the action or relationship.
3. Identify the object or destination of the action.

Return the tuples in this format: **Subject - Relation - Object**. At the end of the process, print “END LIST” to indicate the conclusion of the extraction.

3.3. Fake News Attribution

The final phase of the methodology uses concept-based CTI indicators to attribute newly detected fake news articles to previously identified campaigns. Thus, our approach maps an article to a known disinformation campaign and the threat actor behind it. The primary goal is to determine whether a piece of fake news belongs to an existing campaign and can be linked to specific threat actors. To this end, we consider the following techniques.

- *Syntactic similarity*: detects direct or near-identical matches between articles, measuring the overlap in vocabulary and phrases between potential disinformation content and known fake news. To analyze the syntactic similarity between fake news articles and previously identified disinformation campaigns, we employed **Term Frequency-Inverse Document Frequency** (TF-IDF), widely used in NLP to assess the importance of words within a document relative to a larger corpus [43, 44]. In this context, TF-IDF allows us to measure syntactic similarity between extracted `<subject, relation, object>` triples and reference campaign data, enabling attribution based on textual overlap. To formalize this analysis for fake news attribution, we define the process of matching extracted tuples with those in the ground truth and assigning a campaign based on similarity. Let $T = \{t_1, t_2, \dots, t_n\}$ be the set of extracted tuples from a fake news article under analysis, and $C = \{C_1, C_2, \dots, C_m\}$ be the set of known disinformation campaigns, each characterized by a set of reference tuples T_{C_j} .

Each tuple t_i is transformed into a TF-IDF vector representation \mathbf{v}_i . Similarly, each campaign C_j is represented by the TF-IDF matrix M_{C_j} , where each row corresponds to the TF-IDF vector of a reference tuple associated with the campaign. The cosine similarity between an extracted tuple t_i and a campaign C_j is computed as:

$$\text{sim}(t_i, C_j) = \max_{t_k \in T_{C_j}} \frac{\mathbf{v}_i \cdot \mathbf{v}_k}{\|\mathbf{v}_i\| \|\mathbf{v}_k\|} \quad (1)$$

where \mathbf{v}_i is the TF-IDF vector representation of the extracted tuple t_i , \mathbf{v}_k is the TF-IDF vector representation of a reference tuple t_k from campaign C_j , and $\|\mathbf{v}_i\|$ and $\|\mathbf{v}_k\|$ are the norms of the respective vectors. The maximum operator ensures that the best-matching tuple in the campaign is considered.

A tuple t_i is assigned to a campaign C_j if:

$$\text{sim}(t_i, C_j) \geq \tau \quad (2)$$

where τ is a predefined similarity threshold. If no campaign satisfies this condition, the tuple remains unassigned.

For each fake news article, a **voting mechanism** determines the most probable campaign based on the number of tuples assigned to each campaign. Let V_{C_j} be the vote count for campaign C_j , computed as:

$$V_{C_j} = \sum_{t_i \in T} 1(\text{sim}(t_i, C_j) \geq \tau) \quad (3)$$

where $1(\cdot)$ is the indicator function:

$$1(\text{sim}(t_i, C_j) \geq \tau) = \begin{cases} 1, & \text{if } \text{sim}(t_i, C_j) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The article is ultimately assigned to the campaign with the highest vote count:

$$C^* = \arg \max_{C_j \in \mathcal{C}} V_{C_j} \quad (5)$$

where C^* represents the most probable disinformation campaign associated with the article.

- *Semantic similarity*: leverages *word embeddings* to capture the meaning of the content beyond the exact wording. This approach allows for identifying nuanced connections between potentially misleading news and known disinformation content. This analysis follows a similar approach as in the lexical similarity analysis but with the primary distinction of using semantic embeddings instead of lexical frequency-based representations. Let $T = \{t_1, t_2, \dots, t_n\}$ be the set of extracted tuples from a fake news article, and let $\mathcal{C} = \{C_1, C_2, \dots, C_m\}$ be the set of known disinformation campaigns, each characterized by a set of reference tuples T_{C_j} . Each tuple t_i is transformed into a semantic embedding \mathbf{e}_i in a high-dimensional vector space. Similarly, each campaign C_j is represented by an embedding matrix M_{C_j} , where each row corresponds to the embedding of a reference tuple associated with the campaign.

The cosine similarity between an extracted tuple t_i and a campaign C_j is computed as:

$$\text{sim}(t_i, C_j) = \max_{t_k \in T_{C_j}} \frac{\mathbf{e}_i \cdot \mathbf{e}_k}{\|\mathbf{e}_i\| \|\mathbf{e}_k\|} \quad (6)$$

where \mathbf{e}_i is the embedding representation of the extracted tuple t_i , \mathbf{e}_k is the embedding representation of a reference tuple t_k from campaign C_j , and $\|\mathbf{e}_i\|$ and $\|\mathbf{e}_k\|$ are the norms of the respective embedding vectors.

A tuple t_i is assigned to a campaign C_j if:

$$\text{sim}(t_i, C_j) \geq \tau \quad (7)$$

where τ is a predefined similarity threshold. If no campaign satisfies this condition, the tuple remains unassigned. Again, for each article, a voting mechanism determines the most probable campaign based on the number of tuples assigned to each campaign. Let V_{C_j} be the vote count for campaign C_j , computed as:

$$V_{C_j} = \sum_{t_i \in T} 1(\text{sim}(t_i, C_j) \geq \tau) \quad (8)$$

where $1(\cdot)$ is the indicator function:

$$1(\text{sim}(t_i, C_j) \geq \tau) = \begin{cases} 1, & \text{if } \text{sim}(t_i, C_j) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

The article is ultimately assigned to the campaign with the highest vote count:

$$C^* = \arg \max_{C_j \in \mathcal{C}} V_{C_j} \quad (10)$$

where C^* represents the most probable disinformation campaign associated with the article.

- *Large Language Models*: this approach uses machine learning approaches based on LLMs that are fine-tuned to recognize campaign-specific language patterns and relationships between entities. These models classify fake news articles by disinformation campaign and threat actor. Specifically, we leverage DistilBERT, a lighter, faster variant of BERT, designed to retain most of its capabilities while reducing computational overhead through knowledge distillation [45]. DistilBERT retains the core transformer architecture of BERT, relying on *self-attention mechanisms* and *feed-forward layers* to process and understand text. It follows the standard encoder-only transformer structure, where input text is tokenized and transformed into vector representations before passing through multiple transformer layers.

The self-attention mechanism [46], a key component of transformers, computes contextual relationships between words in a sequence. Given an input sentence tokenized into words $\{w_1, w_2, \dots, w_n\}$, each word is transformed into a high-dimensional embedding vector. The transformer then applies multi-head self-attention to compute contextualized representations. The self-attention mechanism computes a weighted sum of all words in the input sequence to determine how much attention each word should pay to others. This is achieved using three matrices: Query (Q), Key (K), and Value (V).

For each word token x_i , these matrices are computed as:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \quad (11)$$

where $X \in \mathbb{R}^{n \times d}$ is the input matrix (word embeddings), $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable weight matrices, d is the embedding dimension, and d_k is the dimensionality of queries and keys. The attention score between two tokens

is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (12)$$

where QK^T computes the similarity between queries and keys, $\sqrt{d_k}$ scales the values to stabilize training, and the softmax function normalizes the scores so they sum to 1. This allows the model to focus on the most relevant parts of the input sentence. DistilBERT maintains multi-head attention, where multiple independent self-attention computations are performed in parallel, allowing the model to capture different types of relationships between words.

The goal is to let the model learn how to associate extracted tuples with their corresponding disinformation campaigns based on their semantic content. Given a set of fake news articles $A = \{a_1, a_2, \dots, a_n\}$ and a predefined set of disinformation campaigns $C = \{C_1, C_2, \dots, C_m\}$, the model’s task is to classify each article a_i by analyzing the extracted tuples.

Each article a_i consists of a set of tuples $T = \{t_1, t_2, \dots, t_n\}$ where each tuple t_j represents structured information in the <subject, relation, object> form. These tuples serve as the input to the model.

To process the input, each tuple t_j is tokenized and transformed into a numerical representation X_j :

$$X_j = f_{\text{tokenizer}}(t_j)$$

where $f_{\text{tokenizer}}(\cdot)$ maps the tuple into a tokenized sequence compatible with DistilBERT. The model is trained to predict the most probable campaign label C_j for each tuple t_j .

For each input tuple t_j , DistilBERT produces a probability distribution over the possible disinformation campaigns:

$$P(C|t_j) = \text{softmax}(W\mathbf{h}_{\text{CLS}} + b)$$

where W is a learned weight matrix, b is a bias term, \mathbf{h}_{CLS} is the output representation of the DistilBERT classification token, and $P(C|t_j)$ represents the predicted probability distribution over campaigns. The campaign assigned to tuple t_j is:

$$C^*(t_j) = \arg \max_{C_j} P(C_j|t_j)$$

where $C^*(t_j)$ is the campaign with the highest predicted probability.

Since each fake news article a_i consists of multiple tuples, we apply a majority voting mechanism to determine its final campaign classification. Let V_{C_j} be the vote count for campaign C_j :

$$V_{C_j} = \sum_{t_j \in T_i} 1(C^*(t_j) = C_j)$$

where $1(\cdot)$ is the indicator function:

$$1(C^*(t_j) = C_j) = \begin{cases} 1, & \text{if } C^*(t_j) = C_j \\ 0, & \text{otherwise} \end{cases}$$

The final campaign assignment for the article a_i is given by:

$$C^*(a_i) = \arg \max_{C_j} V_{C_j}$$

ensuring that the campaign with the most assigned tuples determines the article’s classification.

These three solutions differ in terms of performance, cost, and accuracy. Syntactic similarity is computationally efficient and straightforward but may fail to capture deeper semantic relationships between articles. Semantic similarity achieves a better understanding of content meaning by utilizing word embeddings, albeit at a higher computational cost. Fine-tuned LLMs offer the highest accuracy and flexibility in identifying complex patterns and relationships. However, their increased computational cost and the requirement for high-quality training data make them the most resource-intensive option.

4. The FakeCTI Dataset

To support our investigation of concept-based CTI indicators, we created a new collection of fake news content in the *FakeCTI* dataset. This structured dataset facilitates the study of disinformation campaigns, grouping fake news that belong to the same campaign and associating the campaigns with threat actors.

We gathered data from diverse and scattered sources linked by Wikipedia in the following archives:

- *Political disinformation website campaigns in the United States* [47];
- *Fake news websites* [48];
- *Fake news troll farms* [49];
- *Corporate disinformation website campaigns* [50].

These archives collect information about several fake news campaigns, in the form of unstructured Cyber Threat Intelligence (CTI) reports [51, 52, 53]. The campaigns have been categorized into various domains such as political disinformation campaigns, troll farms, fake news websites, and corporate disinformation efforts. These reports document the origins, methods, and motivations of fake news, providing insights into how it is created and disseminated.

Political disinformation campaigns often stem from partisan entities or super PACs, with websites like American Action News designed to sway public opinion during elections by spreading conspiracy theories or ideologically charged narratives. Fake news websites, such as Conservative Beaver and The Red Panther, use deceptive tactics like spoofed domains and exaggerated clickbait to generate traffic and ad revenue,

Table 2: Example of FakeCTI dataset entries.

Title	Source	Campaign	Threat Actor	Medium
Vaccinated people are walking biological time bombs and a THREAT to society	Foreign Affairs Intelligence Council	Covid trigger Neurological Degenerative Diseases	Alex Jones	Web
LEAKED: Lady Gaga Halftime Performance to Feature Muslim Tribute	Houston Leader	A Cure for Wellness Promotion	Regency Enterprise	Web
Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement	Ending the Fed News	Viral Fake Election News	N/A	Web

typically driven by independent networks or profit-focused individuals. Troll farms, as seen in operations from Ghana or Myanmar, use coordinated networks of sites to amplify divisive or false stories, frequently backed by anonymous contributors or political actors aiming to manipulate public discourse. Corporate disinformation efforts, exemplified by platforms like Chevron-funded The Richmond Standard, masquerade as community news outlets to promote favorable narratives and suppress criticism, often managed by PR firms.

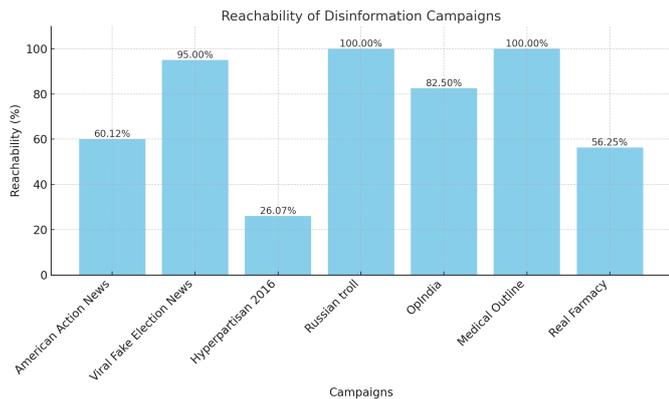


Figure 4: Reachability of disinformation campaigns.

To create the FakeCTI dataset, we extracted data from the websites and platforms used to disseminate fake news and referenced in these archives and systematized the data to make them suitable for research studies. After collecting the raw data from these sources, we removed irrelevant or low-quality entries, such as duplicate articles. Then, using the metadata extracted from the CTI reports, we manually labeled the disinformation articles based on their association with specific disinformation campaigns and threat actors. Finally, we linked fake news to the channels used to spread it, such as social media platforms, blogs, and news websites.

A challenge in building the dataset is the reachability of the disinformation campaigns. Not all fake news articles included in the original campaigns were accessible during the data collection process. Many articles were no longer available, even through the use of web archives that capture historical snapshots of web pages. This limitation arises because mainstream web portals and information agencies often remove false content once it is reported, as part of their efforts to protect readers

from disinformation. As a result, a limitation of the dataset is that it reflects only the accessible portion of the content of the campaigns.

Figure 4 shows the reachability of contents for various disinformation campaigns in the dataset, with respect to the total number of articles reported for the campaigns. The reachability metric quantifies the proportion of articles from each campaign that could still be accessed and retrieved during the data collection process. Campaigns such as *Russian Troll* and *Medical Outline* exhibit full reachability (100%), indicating that all known articles from these campaigns were accessible at the time of writing. Other campaigns, such as *Hyperpartisan 2016* and *Real Farmacy*, have significantly lower reachability at 26.07% and 56.25%, respectively, reflecting the challenges in retrieving articles due to factors like website takedowns or insufficient archival coverage. This variability highlights a limitation in the dataset’s completeness, as campaigns with lower reachability may be underrepresented.

Despite the limitations, we obtained a significant amount of data on fake news campaigns. Table 3 shows some statistics about the FakeCTI dataset. The dataset contains a total of 12,155 fake news samples, encompassing 43 distinct disinformation campaigns orchestrated by 11 identified threat actors. FakeCTI aggregates content from 149 different sources, offering a diverse and comprehensive foundation for studying the dissemination and attribution of fake news across multiple campaigns and actors.

Statistic	Value
Total number of samples	12,155
Number of different threat actors	11
Number of different fake news campaigns	43
Number of different sources	149

Table 3: Summary of FakeCTI dataset statistics.

The resulting entries in the FakeCTI dataset include the following information: the *title* and *text* of the fake news article, a *link* to the article itself, the *disinformation campaign* that spread the fake news, the *threat actor* behind the campaign, and the *dissemination medium* used for spreading fake news. Table 2 illustrates some sample entries from the dataset. The first entry originates from the Foreign Affairs Intelligence Council as

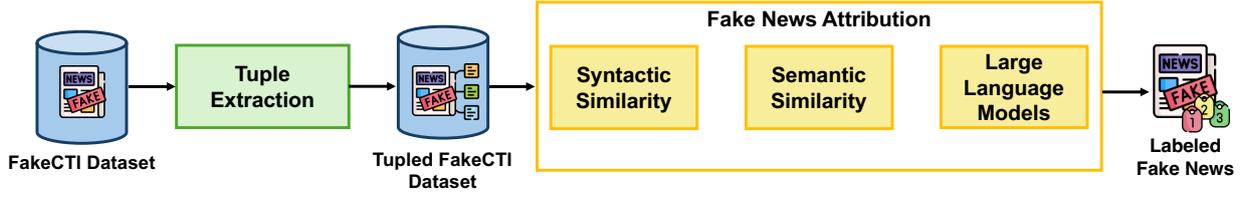


Figure 5: Overview of the experimental analysis.

part of the campaign *Covid triggers Neurological Degenerative Diseases* attributed to Alex Jones, a known figure in spreading health-related disinformation. The second article was published by Houston Leader and is associated with the campaign *A Cure for Wellness Promotion* with its threat actor identified as Regency Enterprise, likely linked to a promotional stunt. Lastly, the third example comes from *Ending the Fed News* and is part of the *Viral Fake Election News* campaign, which refers to the 2016 presidential election in the United States. It is possible to notice how this entry lacks an identified threat actor. This absence highlights that such attribution is not always possible due to the anonymous and decentralized nature of many disinformation efforts.

This dataset bridges a significant gap in existing resources, providing a richer context for fake news attribution tasks. It allows analysts to link disinformation narratives to their broader campaigns and threat actors, making the dataset an invaluable tool for attribution and detection research.

5. Research Questions

We evaluated the effectiveness of the proposed approach by considering the following research questions (RQ).

▷ **RQ1:** *How accurately can LLMs extract tuples from fake news articles, compared to manually annotated ground truth?*

RQ1 investigates the quality of the tuples generated by LLMs in the context of disinformation analysis. The goal is to evaluate the quality of the extracted triples against a manually curated dataset, ensuring that the LLMs can capture the critical relationships and information in fake news articles. This assessment is crucial for determining whether LLMs can reliably standardize and structure unstructured textual data for automated analysis.

▷ **RQ2:** *How effective are the extracted CTI indicators for attributing fake news articles to specific disinformation campaigns and threat actors?*

RQ2 examines whether the CTI indicators can accurately and comprehensively be used to link fake news articles to the appropriate disinformation campaigns or threat actors. By testing attribution techniques (e.g., lexical similarity, semantic similarity, and fine-tuned LLMs), we aim to validate whether the structured triples sufficiently represent the key characteristics of the campaigns to support meaningful and precise attribution efforts. Given the broad focus of RQ2, we further refine this research question into sub-questions RQ2.1 and RQ2.2.

▷ **RQ2.1:** *How effectively can traditional NLP techniques associate extracted tuples with known disinformation campaigns?*

RQ2.1 aims to evaluate the effectiveness of standard NLP methods in mapping extracted `<subject, relation, object>` tuples to their respective disinformation campaigns. The focus is on assessing how well lexical and semantic similarity measures can establish meaningful connections between tuples and campaign narratives.

▷ **RQ2.2:** *Can fine-tuned LLMs improve the accuracy and reliability of fake news attribution compared to traditional NLP-based approaches?*

RQ2.2 investigates whether fine-tuning transformer-based LLMs enhances the attribution process beyond traditional methods. The study examines whether these models, trained specifically on disinformation-related data, can capture more nuanced patterns in tuple-based representations and improve classification performance, ultimately leading to more robust and scalable fake news attribution.

6. Experimental Analysis

To evaluate the effectiveness of CTI indicators, we conducted an experimental analysis of fake news attribution techniques on the *FakeCTI* dataset. These experiments focused on identifying the disinformation campaign behind a given fake news article. Figure 5 illustrates the workflow for the experimental evaluation. First, we start with *tuple extraction* from the *FakeCTI* dataset. Then, leveraging the tupled version of the dataset, we test several techniques for *fake news attribution* to assess the concept-based CTI indicators.

6.1. Tuple Extraction

To extract structured information from fake news articles, we selected a distilled version of the Meta LLaMA model. Model distillation is a machine learning technique that transfers knowledge from a large, complex teacher model to a smaller, more efficient student model, aiming to reduce computational costs while maintaining high performance. We chose LLaMA due to its advanced NLP capabilities, offering deep text understanding, coherent response generation, and adaptability across various contexts. Specifically, we leveraged the `Llama-3-8B-Instruct-v0.10` model [54], equipped with 8 billion parameters, making it highly capable of understanding and generating detailed, context-aware responses.

Despite the efficiency improvements from distillation, executing the model on available hardware resources, particularly GPU memory, remained challenging. To address this, we implemented quantization, a technique that reduces numerical precision in model parameters and calculations. Traditional LLMs

Table 4: Performance comparison across quantized models and temperature settings for tuple extraction. Best performance is **blue**.

Model	Temperature	Precision (%)	Recall (%)	F1-Score (%)	Avg. Extraction Time (s)
Q4_K_M	0.0	63	65	64	25.07
	0.3	67	70	69	29.71
	0.6	60	63	61	24.61
Q5_K_M	0.0	58	61	59	26.14
	0.3	61	64	62	31.53
	0.6	56	60	57	31.45
Q8_0	0.0	54	58	55	29.81
	0.3	57	60	58	23.60
	0.6	50	54	51	44.79

operate using 32-bit floating point representations, but quantization allows these to be reduced to lower precision, such as 8-bit integers, thereby minimizing memory requirements and computational load. While quantization could theoretically degrade model accuracy, modern quantization algorithms mitigate this issue by preserving inference quality [55, 56].

Specifically, we tested three quantized versions of the model:

- **Q4_K_M** (4-bit quantization)
- **Q5_K_M** (5-bit quantization)
- **Q8_0** (8-bit quantization)

These configurations were evaluated based on execution time and precision, ensuring an optimal balance between efficiency and extraction accuracy.

After quantization, we proceeded with the extraction of <subject, relation, object> tuples from fake news articles. This process posed several challenges, including how to optimize the extraction process for better accuracy and how to evaluate the extracted results. A key factor influencing extraction quality was the *temperature* parameter, which controls randomness in model responses. Low temperatures (close to 0) make responses more deterministic, ensuring consistent extractions but potentially reducing the model’s ability to generalize, while higher temperatures (close to 1) increase variability, which can introduce more diverse extractions but at the risk of inconsistency or errors.

To determine the optimal configuration, we tested three temperature values: 0 (fully deterministic), 0.3 (moderate variability), and 0.6 (higher variability). This allowed us to measure the impact of this parameter on the accuracy of tuple extraction.

To evaluate the model, we manually extracted sets of <subject, relation, object> tuples from fake news articles to serve as *ground truth*. This data served as a benchmark to compare model-generated tuples against human-annotated references. We built the ground truth through the following steps:

1. *Random selection of articles*: we randomly sampled a subset of 10 articles from the dataset to ensure diversity in the topic and linguistic style.

2. *Manual tuple extraction*: each article was thoroughly read to manually extract relevant triples, ensuring high-quality, human-verified annotations.
3. *Standardized storage*: we saved the extracted tuples in a structured format, associating each tuple with its source article and corresponding disinformation campaign.

The analysis evaluated the three quantized models across the three temperature settings, resulting in *nine different model configurations*. For each configuration, we calculated precision, recall, and F1-Score based on the alignment between extracted tuples and the ground truth data. Table 4 illustrates the results of the evaluation.

The results indicate that the Q4_K_M model with a temperature of 0.3 achieved the best overall performance, with an accuracy of 67%, recall of 70%, and an F1-score of 69%. This model demonstrated a strong balance between extracting relevant <subject, relation, object> tuples while maintaining precision in identifying meaningful relationships within fake news articles. Comparing different temperature settings, lower temperatures (e.g., 0.0) generally resulted in higher precision but slightly lower recall, as the model tended to be more deterministic and less likely to capture variations in sentence structures. Conversely, increasing the temperature to 0.6 introduced more variability, slightly improving recall but leading to noisier extractions, reducing overall accuracy. Regarding the impact of quantization, the Q5_K_M and Q8_0 models exhibited lower accuracy across all temperature settings, with the Q8_0 model at 0.6 performing the worst, achieving only 50% accuracy and 51% F1-score. This decline in performance suggests that higher-bit quantization, while preserving model precision, may introduce excess noise or reduce the effectiveness of structured extraction tasks. Overall, the Q4_K_M model at 0.3 temperature emerged as the optimal choice, offering the best trade-off between precision, recall, and F1-Score in tuple extraction. Based on these findings, we leveraged this model for large-scale tuple extraction on the entire FakeCTI dataset, extracting structured <subject, relation, object> tuples from each fake news article. The resulting structured data was compiled into a tuple-level version of the FakeCTI dataset, providing a standardized representation of disinformation narratives and enabling further analyses for attribution and detection

Table 5: Comparison between numbers of manually and automatically extracted tuples (Q4_K_M_0.3TEMP).

Title	Ground Truth Tuples	Extracted Tuples	Corresponding Tuples
Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement	11	13	11
Clinton Rape Accuser Punches Back for Trump	4	7	4
What a sore loser! Trump Just Gave The Most PATHETIC Excuse For Losing To Hillary In Debate	1	3	0
#HowToConfuseAMillennial Let them WATCH what Muslims think about women’s rights!	1	1	1
How Can I Cure ADHD Naturally?	7	17	7
What Ointment Is Good For Eczema?	9	11	6
Harrison Butker’s Kansas City Chiefs among the few NFL teams not to celebrate Pride Month publicly	8	9	8
Gotcha: Biden’s Vaccine Mandate Never Materialized Beyond A Press Release	16	7	1
Six Vaccinated People Diagnosed with Mumps	13	23	13
Breaking: Lia Thomas to Swim With A Men’s Team After Extreme Criticism	10	27	7

tasks.

Table 5 reports the number of extracted tuples for the best-performing model against the ground truth data. It is worth noting that, in several cases, the model successfully retrieves a number of tuples close to the ground truth. However, discrepancies emerge in articles where the number of extracted tuples significantly exceeds the ground truth, indicating that the model may tend to overgenerate information.

For completeness, we also report an example of a comparison between manually extracted tuples and LLM-extracted ones for a specific article in Table 6. This comparison reveals both the strengths and limitations of the automated extraction. One notable difference is the number of tuples extracted, with the LLM generating 13 tuples compared to the 11 manually annotated ones. The LLM successfully identifies core entities, relations, and objects in several cases, producing near-perfect matches such as <Pope Francis, has made, an unprecedented decision to endorse a US presidential candidate> and <The FBI, refuse, to recommend prosecution>, demonstrating its ability to capture both syntactic and semantic elements when the structure is straightforward. However, lexical variations are observed in cases where meaning is preserved but phrasing differs, such as <I (Pope Francis), feel, that to not voice my concern would be a dereliction of my duty>, which the LLM extracted as <Pope Francis, feel, that to not voice concern would be a dereliction of duty>. Similarly, <The FBI, admitting, the law had been broken by Secretary Clinton> was transformed into <The FBI, admit, the law had been broken on multiple occasions by Secretary Clinton>, indicating a tendency for slight rephrasing rather than exact extraction.

However, some tuples exhibit errors in entity-relation matching, such as <Support of Donald Trump, was

released, from the Vatican this evening>, which the LLM split into <Pope Francis, endorse, Donald Trump> and <Pope Francis, release, a statement>. While this captures part of the meaning, it generalizes the statement and loses specific details. Additionally, the LLM failed to extract <I (Pope Francis), don’t agree, with Mr. Trump on some issues>, suggesting challenges in handling negations or nuanced disagreements. The model also introduced extra tuples not present in the ground truth, such as <Pope Francis, voice, his concern> and <Pope Francis, offer, any kind of support>, the latter being problematic as it misrepresents the original text. This tendency to infer relationships beyond the given content raises concerns about hallucination. Despite these limitations, the LLM performs well in structuring information from fake news articles, with many tuples closely matching the ground truth and preserving the underlying meaning. Nevertheless, fine-tuning and post-processing validation may be necessary to mitigate hallucinations, improve accuracy in complex linguistic constructs, and ensure reliable extraction for disinformation attribution.

RQ1: How accurately can LLMs extract tuples from fake news articles, compared to manually annotated ground truth?

The results demonstrate that LLMs are capable of effectively identifying and structuring relationships between entities within fake news articles in <subject, relation, object> tuples. In several cases, the LLM successfully extracts a number of tuples close to the ground truth, highlighting its capability to capture key entities, relations, and objects with a high degree of syntactic and semantic accuracy. The model exhibits

Table 6: Comparison between manually extracted tuples and LLM-extracted tuples for *Pope Francis Shocks World, Endorses Donald Trump for President* article.

Ground Truth (Manual Extraction)	LLM-based Extraction
Pope Francis - has made - the unprecedented decision to endorse a US presidential candidate	Pope Francis - has made - an unprecedented decision to endorse a US presidential candidate
Support of Donald Trump - was released - from the Vatican this evening	Pope Francis - endorse - Donald Trump Pope Francis - release - a statement
I (Pope Francis) - feel - that to not voice my concern would be a dereliction of my duty	Pope Francis - feel - that to not voice concern would be a dereliction of duty
I (Pope Francis) - think - a strong and free America is vitally important	Pope Francis - say - a strong and free America is vitally important
I (Pope Francis) - now fear - that the Rule of Law in America has been dealt a dangerous blow	Pope Francis - fear - the Rule of Law in America has been dealt a dangerous blow
I (Pope Francis) - don't agree - with Mr. Trump on some issues	
I (Pope Francis) - feel - that voting against the powerful political forces is the only option	Pope Francis - think - that voting against the powerful political forces is the only option
I (Pope Francis) - ask - that Americans vote for Donald Trump for President of the United States	Pope Francis - ask - that Americans vote for Donald Trump for President of the United States
The FBI - refusing - to recommend prosecution	The FBI - refuse - to recommend prosecution
The FBI - admitting - the law had been broken by Secretary Clinton	The FBI - admit - the law had been broken on multiple occasions by Secretary Clinton
The FBI - expose - itself as corrupted by political forces	The FBI - expose - itself as corrupted by political forces Pope Francis - voice - his concern Pope Francis - offer - any kind of support
Total Tuples: 11	Total Tuples: 13

strong performance in cases where sentence structures are straightforward, often generating tuples that match the ground truth with minimal variation, still preserving both meaning and structure. However, the model may overgenerate tuples, sometimes producing more than the ground truth, raising concerns about hallucination. In some cases, the LLM inferred a relationship not explicitly stated in the original text, introducing factual inaccuracies. Furthermore, the model occasionally splits or rephrases tuples, which captures parts of the meaning but may lose important contextual details. Despite these shortcomings, LLMs effectively structure information from fake news articles, making them a valuable tool for automated disinformation analysis. While the ability to generalize and extract meaningful tuples is clear, improvements through fine-tuning and post-processing techniques are necessary to reduce hallucinations, improve entity-relation accuracy, and ensure reliable extraction of complex linguistic constructs. Overall, LLMs show strong potential for tuple-based CTI extraction, even though human oversight or post-processing remains necessary to ensure high-quality results.

6.2. Fake News Attribution

This evaluation aims to assess the effectiveness of the extracted CTI indicators in attributing fake news articles to specific disinformation campaigns. By utilizing `<subject, relation, object>` triples as structured indicators of disinformation narratives, we aim to investigate whether these extracted relationships capture the distinctive characteristics of each campaign and provide a reliable basis for automated attribution. The objective is to determine how well these CTI indicators can distinguish between different campaigns, ensuring that they serve as meaningful features for classification.

To this end, we conducted a comparative analysis of multiple techniques, evaluating their ability to classify fake news articles based on CTI tuples. The evaluation entails the following steps:

1. Grouping extracted triples by article to ensure consistency in classification.

2. Testing different attribution strategies, including similarity-based voting mechanisms and deep learning-based classifiers.
3. Evaluating the performance to determine which approach best captures campaign-specific disinformation patterns.

The study systematically explores techniques that measure lexical and semantic similarity between triples and reference campaign data (TF-IDF and SBERT). These methods are compared with LLM-based classification, using a DistilBERT model fine-tuned on the dataset. By comparing these techniques, we aim to establish whether the extracted triples serve as robust and representative indicators for disinformation attribution, effectively distinguishing fake news campaigns based on their content structure and relational patterns.

6.2.1. Syntactic Similarity

To analyze the syntactic similarity between fake news articles and previously identified disinformation campaigns, we consider the *Term Frequency-Inverse Document Frequency* (TF-IDF), introduced in Section 3.3. This metric is widespread in NLP to assess the importance of words within a document relative to a larger corpus. In our study, TF-IDF enables measuring *lexical similarity* between extracted `<subject, relation, object>` triples and reference campaign data. We treated each extracted triple as a separate document, while the reference campaign data formed the overall corpus. The goal was to compare the lexical similarity of each test triple with triples from known campaigns to determine the most probable attribution.

We integrated the TF-IDF metric into two different attribution strategies: *TF-IDF voting* and *TF-IDF thresholding*.

TF-IDF voting. We designed the *TF-IDF voting* solution to classify an article based on the cumulative lexical similarity of its grouped tuples to known disinformation campaigns. The process begins with the initialization of campaign counters, where each campaign is assigned a counter to track how many tuples are lexically similar to it. Since tuples extracted from the same article are thematically related, we analyzed them as a collective unit by grouping them based on their article of origin. Each tuple is then transformed into a TF-IDF vector representing its lexical features. We computed the similarity between each test triple and the campaign reference data using cosine

similarity, as described in Section 3.3. If a tuple’s similarity score exceeds a predefined threshold, the counter for the corresponding campaign is incremented. After processing all tuples from an article, the campaign with the most votes is assigned as the source of disinformation. This approach proves particularly effective in cases where disinformation campaigns rely on repetitive language patterns, making lexical similarity a strong indicator of attribution. Specifically, we define *accuracy* as:

$$Accuracy = \frac{N_{CA}}{N_A}$$

where N_{CA} is the number of articles correctly linked to the corresponding disinformation campaign, while N_A is the total number of articles.

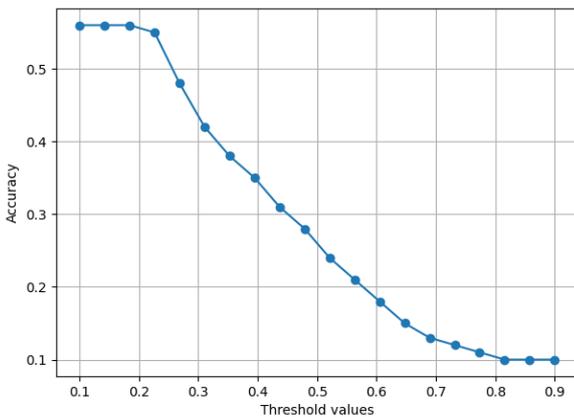


Figure 6: Evaluation of TF-IDF voting.

We divided the dataset into a training set (66%) and a test set (34%), maintaining a balanced representation of different campaigns across both subsets. Then, we evaluated TF-IDF voting by conducting multiple experiments to determine the optimal similarity threshold. We systematically increased the threshold from 0.1 to 0.9 and recorded the corresponding accuracy values. Figure 6 illustrates the results of the evaluation for different threshold values. It is possible to notice that we achieved the highest accuracy (56%) with a threshold of 0.25. When the threshold was lower, more articles were classified, but the inclusion of loosely related tuples led to an increase in false positives, reducing overall precision. On the other hand, setting the threshold too high resulted in underfitting, where valid attributions were missed, leading to a drop in recall. Beyond 0.25, accuracy declined as the stricter similarity requirements prevented many articles from receiving a campaign label. These findings highlight that while TF-IDF voting benefits from flexible attribution, it remains highly dependent on threshold tuning for optimal performance.

TF-IDF thresholding. Unlike TF-IDF voting, which assigns a campaign based on the cumulative count of lexically similar tuples, *TF-IDF thresholding* introduces a stricter requirement by imposing a minimum number of matching tuples before an article can be attributed to a disinformation campaign. The pro-

cess begins with grouping tuples by article and converted into TF-IDF vectors to compute their lexical similarity with known campaign data, as in the voting approach. However, rather than relying on a majority voting mechanism, this method sets campaign-specific thresholds, which define the minimum number of tuples that must surpass a similarity threshold for an article to be classified under a particular campaign. If none of the campaigns reach this requirement, the article remains unclassified, ensuring that only articles with a strong enough lexical resemblance to an existing disinformation campaign receive an attribution.

This approach differs from TF-IDF voting because it prioritizes precision over recall, avoiding misclassifications due to weak or incidental lexical similarities. However, it also introduces a significant limitation: since the number of tuples extracted per article varies, some articles may contain too few tuples to surpass the required threshold, even if they belong to that specific campaign.

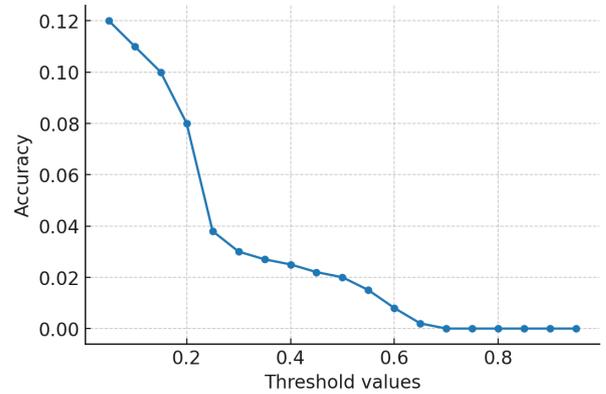


Figure 7: Evaluation of TF-IDF thresholding.

As in the previous evaluation, we tested TF-IDF thresholding with different threshold values to determine the most effective setting. However, the results, as shown in Figure 7, showed that TF-IDF thresholding performed significantly worse than voting, with a maximum accuracy of only 12% at a threshold around 0.1, which is too low to effectively determine if an article belongs to a specific campaign. The main reason for this poor performance is that many articles do not contain enough extracted tuples to surpass the required minimum similarity threshold, resulting in a high number of unclassified samples. As the threshold increased, accuracy dropped further, highlighting the method’s inability to effectively handle variation in article content. Unlike TF-IDF voting, which provides attribution based on cumulative votes, this approach over-prioritizes precision, leading to a sharp decline in recall and making it impractical for real-world disinformation campaign attribution.

When comparing both methods, TF-IDF voting emerges as the better one for attributing fake news articles to disinformation campaigns. It offers higher accuracy (56%) and greater adaptability, allowing it to classify more articles effectively. By contrast, TF-IDF thresholding’s rigid minimum similarity requirement results in a substantial number of unclassified sam-

ples, making it far less effective, with a peak accuracy of only 12%. The primary advantage of TF-IDF voting is its flexibility, as it can accommodate linguistic variations in fake news articles while still maintaining a degree of precision. However, it remains highly sensitive to threshold tuning, requiring careful parameter selection to avoid misclassifications. In contrast, TF-IDF thresholding, while theoretically more precise, proves to be too restrictive in practical applications, as the number of extracted triples per article varies, preventing many valid attributions.

6.2.2. Semantic similarity

This analysis builds upon the lexical similarity evaluation by introducing a semantic approach to fake news attribution. By relying on direct word overlap, TF-IDF-based methods often fail to capture meaningful connections between reworded but conceptually equivalent disinformation content. To address this limitation, we employed *Sentence-BERT* (SBERT), which enhances attribution accuracy by detecting semantic equivalence between extracted triples and campaign reference data. Leveraging word embeddings, SBERT enables a context-aware comparison, making it more resilient to textual variations introduced by disinformation actors attempting to evade detection.

SBERT is a modified version of BERT optimized for sentence-level similarity tasks. Unlike standard BERT, which processes input tokens independently, SBERT generates dense vector representations, i.e., embeddings, that can be compared directly using similarity metrics. An embedding is a mathematical representation of a word or sentence in a high-dimensional space, where semantically similar words or phrases are mapped closer together. This structure allows SBERT to detect reworded narratives, even when they do not share direct lexical overlap. For instance, the model can recognize that the triples `<Organization X, promotes, misinformation>` and `<Group Y, spreads, false narratives>` convey the same message despite using different terminology. This capability is critical for disinformation analysis, as fake news articles often undergo rephrasing and paraphrasing while maintaining their core misleading message.

To assess the effectiveness of SBERT in attributing fake news articles to disinformation campaigns, we performed an evaluation that mirrors the TF-IDF evaluation but incorporates semantic similarity instead of lexical matching. The dataset was split into a training set containing campaign-labeled triples and a test set comprising unlabeled triples to be classified. First, we divided the dataset into a training set (66%) and a test set (34%), maintaining a balanced representation of different campaigns across both subsets. Then, we transformed each triple in the test set into an embedding, creating a numerical vector representation. For this purpose, we leveraged the embedding model in the `sentence-transformers` Python library [57, 58]. We applied the same process to all triples in the training set, ensuring each campaign had an associated set of embeddings. Once the embeddings were generated, we computed cosine similarity scores between each test triple and all training triples. We adopted a voting-based approach to determine campaign attribution, assigning a test tuple to the campaign whose training

tuples had the highest similarity scores. Again, threshold optimization was a critical aspect of the evaluation. We experimented with multiple similarity thresholds to determine the best cutoff for confident attribution.

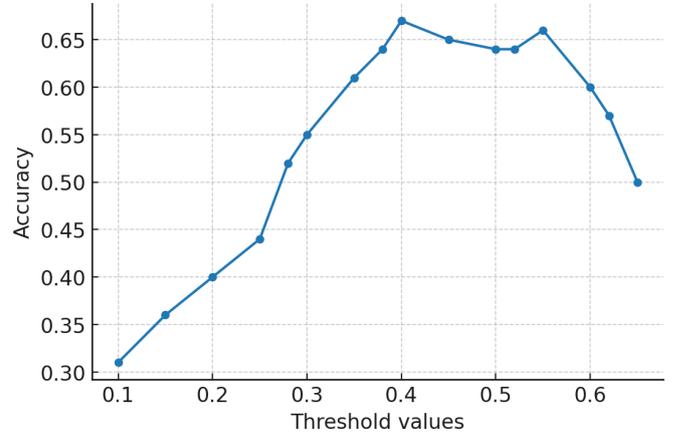


Figure 8: Evaluation of SBERT.

To assess the effectiveness of SBERT in attributing fake news to disinformation campaigns, we employed the same accuracy metric defined in the lexical similarity analysis, calculated as the ratio between the number of correctly classified articles and the total number of fake news articles. Figure 8 illustrates the results of the evaluation for different threshold values. These results demonstrate that SBERT outperformed TF-IDF-based approaches, achieving an accuracy of 67.5% with a similarity threshold of 0.4. Again, the results indicate a trade-off between precision and recall depending on the threshold selection. Increasing the threshold imposes stricter similarity criteria, thereby improving precision but reducing recall, as fewer triples are considered sufficiently similar to match with a campaign. Conversely, lowering the threshold leads to a higher recall by permitting a broader set of matches; however, this comes at the cost of reduced precision due to increasing false positives. The superior performance of SBERT over TF-IDF confirms the importance of capturing semantic rather than purely lexical similarity in fake news attribution. By leveraging word embeddings, SBERT effectively identifies connections between different phrasings of the same narrative, making it particularly robust against minor textual variations and paraphrasing strategies commonly used in disinformation campaigns. SBERT demonstrated several key advantages over lexical similarity methods. Its ability to capture meaning beyond direct word matching made it highly effective in detecting reworded but semantically identical disinformation content. By leveraging word embeddings, SBERT provided a context-aware approach, reducing the risk of missing campaign attributions due to superficial linguistic changes. Additionally, its performance gains in accuracy highlighted its suitability for fake news analysis, where threat actors frequently paraphrase misleading narratives.

RQ2.1: *How effectively can traditional NLP techniques associate extracted tuples with known disinformation campaigns?*

Traditional NLP techniques such as TF-IDF- and SBERT-based similarity offer two distinct approaches to associating extracted tuples with disinformation campaigns. TF-IDF relies on a lexical representation of text, where the importance of words is weighted based on their frequency in a document relative to the entire dataset. This approach is effective when disinformation campaigns rely on repetitive language patterns, making lexical overlap a strong attribution indicator. However, it struggles with paraphrasing and variations in wording, leading to lower overall accuracy. Conversely, SBERT enhances this process by leveraging semantic embeddings. This allows SBERT to capture deeper relationships between tuples, improving its ability to generalize across variations in wording and rephrasings commonly found in fake news articles. Ultimately, while both methods provide valuable insights, semantic similarity analysis proves more effective for this attribution task, as it accounts for textual variations that traditional keyword-based methods cannot fully address. However, the performance of both approaches remains suboptimal, suggesting that more advanced techniques, such as fine-tuned large language models, could further enhance fake news attribution.

6.2.3. Large Language Models

To further improve the attribution of fake news articles to disinformation campaigns, we investigate the use of fine-tuned Large Language Models (LLMs). While previous approaches, such as TF-IDF-based lexical similarity and SBERT-based semantic similarity, provided valuable insights, they remained limited in their ability to fully capture nuanced campaign patterns. Fine-tuning a transformer-based LLM allows the model to learn richer contextual representations from labeled data, potentially yielding a more robust attribution mechanism. As introduced in Section 3.3, we leverage DistilBERT, a lightweight and computationally efficient variant of BERT. The objective of this evaluation is to determine how effectively an LLM, specifically DistilBERT, can be adapted to the task of disinformation campaign attribution based on the `<subject, relation, object>` triples. Unlike traditional NLP techniques such as TF-IDF and SBERT, which rely on predefined similarity measures, fine-tuning allows the model to learn contextual representations from labeled examples, potentially improving classification accuracy. Through this evaluation, we assess whether fine-tuning on our dataset enhances the model’s ability to generalize across unseen articles, providing a robust solution for automated fake news attribution.

To ensure a structured evaluation, we divided our dataset into three distinct subsets:

- *Training set (80%)*: used to fine-tune DistilBERT, allowing it to learn patterns and associations between extracted

tuples and their corresponding disinformation campaigns.

- *Validation set (10%)*: used to optimize hyperparameters and monitor performance during training, preventing overfitting by adjusting model parameters accordingly.
- *Test set (10%)*: reserved for final evaluation, this subset consists of unseen data to assess the generalization capabilities of the trained model.

The dataset was split using a stratified sampling approach to ensure that each disinformation campaign was proportionally represented in all three subsets. Additionally, we applied label encoding to map campaign names to numerical values and tokenized the tuples using the WordPiece tokenizer to prepare them for input into the transformer model.

The fine-tuning process was performed using a pre-trained DistilBERT model [59], which was adapted to our specific task by adding a classification layer. The training was conducted using cross-entropy loss, with a learning rate of $2e-5$ and a batch size of 32. The model was fine-tuned over multiple epochs, with an early stopping strategy to prevent overfitting.

Each extracted tuple was processed independently, producing a probability distribution over the set of known disinformation campaigns. Since each article contains multiple tuples, we applied a majority voting mechanism to aggregate individual tuple predictions into a final article classification:

$$C^*(a_i) = \arg \max_{C_j} V_{C_j}$$

where $C^*(a_i)$ is the most likely campaign for article a_i , and V_{C_j} represents the number of tuples assigned to campaign C_j .

We assessed the performance of DistilBERT using the following accuracy metric, which measures the proportion of correctly classified articles:

$$Accuracy = \frac{N_{CA}}{N_A}$$

where N_{CA} is the number of correctly classified articles, and N_A is the total number of articles in the test set.

Figure 9 presents the training and validation accuracy trends over multiple epochs, alongside the loss curves that indicate the model’s learning progress. Initially, both training and validation accuracy increased steadily, demonstrating that the model was effectively learning meaningful patterns from the dataset. However, after epoch 14, the validation accuracy stabilized while the training accuracy continued to improve, signaling the onset of overfitting, a common issue where the model becomes too specialized to the training data and fails to generalize effectively to unseen examples. A similar trend was observed in the loss curves, where the training loss kept decreasing, while the validation loss reached its minimum around epoch 14 before beginning to rise. This divergence between training and validation loss further confirmed that the model was memorizing training instances rather than learning generalizable features. To address this issue, we implemented an early stopping mechanism, selecting the best-performing model checkpoint at epoch 14 to

prevent further overfitting and ensure better generalization to unseen fake news articles. The stabilized validation loss and accuracy at this checkpoint suggest that the model had captured the optimal balance between learning and generalization.

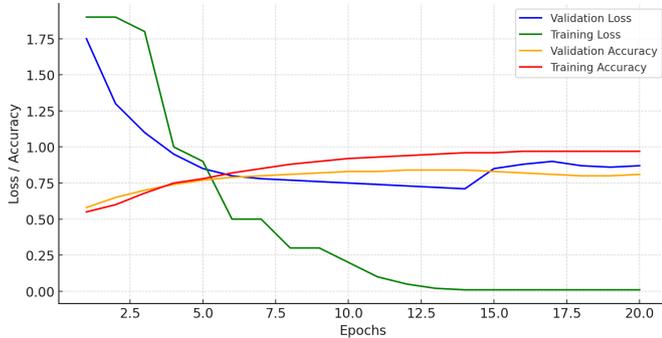


Figure 9: Training and validation accuracy during fine-tuning of DistilBERT.

Table 7 compares accuracy results across different attribution method. The DistilBERT model achieved an accuracy of 94%, significantly outperforming traditional NLP techniques such as TF-IDF (56.0%) and SBERT (67.5%).

The primary advantage of DistilBERT lies in its contextual understanding, allowing it to learn deep representations of disinformation narratives rather than relying on strict word-matching (as in TF-IDF) or predefined semantic similarity metrics (as in SBERT). Furthermore, DistilBERT is highly robust against variations in wording and phrasing, making it more resilient to the paraphrasing strategies often employed in fake news campaigns. Unlike the other approaches, which require manually defining similarity measures or thresholds, DistilBERT automatically learns the most relevant features from labeled data, leading to higher generalization across different campaigns. This adaptability makes fine-tuned DistilBERT the most effective method tested, providing a more accurate and scalable solution for automated fake news attribution.

Method	Accuracy
TF-IDF Voting	56.0% ■■
TF-IDF Thresholding	12.0% ■
SBERT Voting	67.5% ■■
DistilBERT	94.0% ■■■■

Table 7: Comparison of accuracy scores across different fake news attribution methods.

RQ2.2: *Can fine-tuned LLMs improve the accuracy and reliability of fake news attribution compared to traditional NLP-based approaches?*

The experimental analysis of DistilBERT demonstrates that fine-tuned LLMs significantly enhance the accuracy and reliability of fake news attribution compared to traditional NLP-based approaches. While lexical simi-

larity and semantic similarity provided useful insights into the association between extracted tuples and disinformation campaigns, they were inherently limited by their reliance on predefined similarity metrics. In contrast, fine-tuning a transformer-based model such as DistilBERT allows it to learn campaign-specific features from labeled examples, improving its ability to generalize across unseen articles. Our evaluation results confirm that DistilBERT outperforms all previous approaches, thanks to its ability to capture contextual meaning, rather than relying solely on lexical overlap or fixed similarity thresholds. Unlike TF-IDF, which struggles with wording variations and paraphrasing, and SBERT, which relies on static pre-trained embeddings, fine-tuning enables DistilBERT to adapt dynamically to the patterns and nuances of disinformation campaigns. Another crucial advantage of fine-tuned LLMs is their ability to automate feature learning, removing the need for manually crafted similarity measures. While traditional NLP techniques require threshold tuning and manual similarity criteria, DistilBERT learns the most relevant attribution patterns directly from the dataset. This leads to higher generalization capabilities, ensuring the model remains effective even when encountering previously unseen campaigns.

7. Conclusion

In this study, we introduced *concept-based Cyber Threat Intelligence (CTI) indicators*, a novel approach to enhancing the detection and attribution of disinformation campaigns. Unlike traditional CTI indicators, which primarily rely on low-level artifacts such as domain names and social media handles, concept-based CTI captures the underlying narratives and relationships within disinformation content. These indicators focus on semantic and structural characteristics, ensuring greater resilience against adversarial modifications. By leveraging structured `<subject, relation, object>` tuples, we represent disinformation in a format that preserves key entities, their interactions, and thematic consistency, enabling robust analysis across different campaigns. To support this approach, we developed *FakeCTI*, the first dataset that systematically links fake news articles to known disinformation campaigns and threat actors. FakeCTI comprises 12,155 articles spanning 43 disinformation campaigns, annotated with metadata specifying their campaign origin, associated threat actors, and dissemination platforms. FakeCTI aims to facilitate research in automated disinformation detection and attribution, representing a key resource for the cybersecurity and disinformation research communities. We employed Large Language Models (LLMs) to extract structured intelligence from fake news articles. Furthermore, we investigated multiple attribution techniques, comparing lexical similarity, semantic similarity, and fine-tuned LLMs. The results highlight that fine-tuned LLMs achieved the highest attribution accuracy (up to 94%), significantly outperforming previous approaches. These findings underscore the

potential of machine learning-enhanced CTI in improving the tracking and attribution of disinformation campaigns.

Acknowledgments

This work has been partially supported by the *IDA—Information Disorder Awareness* Project funded by the European Union-Next Generation EU within the SERICS Program through the MUR National Recovery and Resilience Plan under Grant PE00000014.

References

- [1] K. Shu, H. R. Bernard, H. Liu, Studying fake news via network analysis: detection and mitigation, *Emerging research challenges and opportunities in computational social network analysis and mining* (2019) 43–65.
- [2] P. Meel, D. K. Vishwakarma, Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities, *Expert Systems with Applications* 153 (2020) 112986.
- [3] Politico, Up close and personal with ai ‘fake news’, <https://www.politico.com/newsletters/digital-future-daily/2024/04/16/up-close-and-personal-with-ai-fake-news-00152573>.
- [4] E. Parliament, The impact of disinformation on democratic processes and human rights in the world, [https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU\(2021\)653635_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf).
- [5] UNESCO, Journalism, fake news & disinformation: handbook for journalism education and training, <https://unesdoc.unesco.org/ark:/48223/pf0000265552>.
- [6] E. Aïmeur, S. Amri, G. Brassard, Fake news, disinformation and misinformation in social media: a review, *Social Network Analysis and Mining* 13 (1) (2023) 30.
- [7] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, L. De Alfaro, Some like it hoax: Automated fake news detection in social networks, *arXiv preprint arXiv:1704.07506* (2017).
- [8] D. J. Bianco, The pyramid of pain, <http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html>.
- [9] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, Combating fake news: A survey on identification and mitigation techniques, *ACM transactions on intelligent systems and technology (TIST)* 10 (3) (2019) 1–42.
- [10] M. Mahyoob, J. Al-Garaady, M. Alrahaili, Linguistic-based detection of fake news in social media, *Forthcoming, International Journal of English Linguistics* 11 (1) (2020).
- [11] A. Choudhary, A. Arora, Linguistic feature based learning model for fake news detection and classification, *Expert Systems with Applications* 169 (2021) 114171.
- [12] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, J. Liu, Content based fake news detection using knowledge graphs, in: *The Semantic Web—ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I 17*, Springer, 2018, pp. 669–683.
- [13] Y. Liu, Y.-F. Wu, Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32, 2018.
- [14] F. Monti, F. Frasca, D. Eynard, D. Mannion, M. M. Bronstein, Fake news detection on social media using geometric deep learning, *arXiv preprint arXiv:1902.06673* (2019).
- [15] N. Ruchansky, S. Seo, Y. Liu, Csi: A hybrid deep model for fake news detection, in: *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, 2017, pp. 797–806.
- [16] J. B. Bak-Coleman, I. Kennedy, M. Wack, A. Beers, J. S. Schafer, E. S. Spiro, K. Starbird, J. D. West, Combining interventions to reduce the spread of viral misinformation, *Nature Human Behaviour* 6 (10) (2022) 1372–1380.
- [17] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, A. Flammini, Computational fact checking from knowledge networks, *PLoS one* 10 (6) (2015) e0128193.
- [18] K. Popat, S. Mukherjee, A. Yates, G. Weikum, Declare: Debunking fake news and false claims using evidence-aware deep learning, *arXiv preprint arXiv:1809.06416* (2018).
- [19] OASIS Open, STIX Project, <https://oasis-open.github.io/cti-documentation/stix/intro>.
- [20] MITRE, Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression, <https://www.mitre.org/sites/default/files/publications/stix.pdf>.
- [21] MITRE, The Trusted Automated eXchange of Indicator Information, https://taxii.mitre.org/about/documents/Introduction_to_TAXII_White_Paper_May_2014.pdf.
- [22] MISP Project, MISP Threat Sharing, <https://www.misp-project.org/>.
- [23] DISARM Foundation, DISARM, <https://www.disarm.foundation/framework>.
- [24] M. Corporation, MITRE ATT&CK, <https://attack.mitre.org/>.
- [25] F. S. González, J. Pastor-Galindo, J. A. Ruipérez-Valiente, Toward interoperable representation and sharing of disinformation incidents in cyber threat intelligence, *arXiv preprint arXiv:2502.20997* (2025).
- [26] ANSSI, OpenCTI, <https://github.com/OpenCTI-Platform/opencti>.
- [27] Forum of Incident Response and Security Teams (FIRST), Common Vulnerability Scoring System, <https://www.first.org/cvss/>.
- [28] K. Nakamura, S. Levy, W. Y. Wang, Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, in: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6149–6157.
- [29] D. K. Sharma, S. Garg, Ifnd: a benchmark dataset for fake news detection, *Complex & intelligent systems* 9 (3) (2023) 2843–2863.
- [30] W. Y. Wang, “liar, liar pants on fire”: A new benchmark dataset for fake news detection, *arXiv preprint arXiv:1705.00648* (2017).
- [31] T. P. Institute, Politifact, <https://www.politifact.com/>.
- [32] J. Baumgartner, S. Zannettou, M. Squire, J. Blackburn, The pushshift telegram dataset, in: *Proceedings of the international AAAI conference on web and social media*, Vol. 14, 2020, pp. 840–847.
- [33] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The pushshift reddit dataset, in: *Proceedings of the international AAAI conference on web and social media*, Vol. 14, 2020, pp. 830–839.
- [34] G. Santia, J. Williams, 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, 2018, pp. 531–540.
- [35] Health Nut News, Top Leading Cancer Scientist Found Dead, <https://healthnutnews.com/top-leading-cancer-scientist-found-dead-in-found-dead-in-woods-in-rubber-fetish-suit/>.
- [36] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, E. Chen, Large language models for generative information extraction: A survey, *Frontiers of Computer Science* 18 (6) (2024) 186357.
- [37] G. Li, P. Wang, W. Ke, Revisiting large language models as zero-shot relation extractors, *arXiv preprint arXiv:2310.05028* (2023).
- [38] Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, S. Kurohashi, Gpt-re: In-context learning for relation extraction using large language models, in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 3534–3547.
- [39] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the potential of prompt engineering in large language models: a comprehensive review, *arXiv preprint arXiv:2310.14735* (2023).
- [40] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. El-nashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, *arXiv preprint arXiv:2302.11382* (2023).
- [41] J. Kim, S. Park, K. Jeong, S. Lee, S. H. Han, J. Lee, P. Kang, Which is better? exploring prompting strategy for llm-based metrics, *arXiv preprint arXiv:2311.03754* (2023).
- [42] M. Chen, K. Zhu, B. Lu, D. Li, Q. Yuan, Y. Zhu, Aecr: Automatic attack technique intelligence extraction based on fine-tuned large language model, *Computers & Security* 150 (2025) 104213.
- [43] S. Qaiser, R. Ali, Text mining: use of tf-idf to examine the relevance of words to documents, *International journal of computer applications*

- 181 (1) (2018) 25–29.
- [44] J. Ramos, et al., Using tf-idf to determine word relevance in document queries, in: Proceedings of the first instructional conference on machine learning, Vol. 242, Citeseer, 2003, pp. 29–48.
 - [45] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019). URL <https://api.semanticscholar.org/CorpusID:203626972>
 - [46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
 - [47] Wikipedia, List of political disinformation website campaigns in the united states, https://en.m.wikipedia.org/w/index.php?title=List_of_political_disinformation_website_campaigns_in_the_United_States.
 - [48] Wikipedia, List of fake news websites, https://en.wikipedia.org/w/index.php?title=List_of_fake_news_websites.
 - [49] Wikipedia, List of fake news troll farms, https://en.wikipedia.org/w/index.php?title=List_of_fake_news_troll_farms.
 - [50] Wikipedia, List of corporate disinformation website campaigns, https://en.wikipedia.org/w/index.php?title=List_of_corporate_disinformation_website_campaigns.
 - [51] C. News, Faking it: Unravelling a fake news story involving stephen hawking, <https://www.cambridge-news.co.uk/business/business-news/faking-it-unravelling-fake-news-12468676>.
 - [52] B. Computer, Content farm impersonates 60+ major news outlets, like bbc, cnn, cnbc, <https://web.archive.org/web/20240305204846/https://www.bleepingcomputer.com/news/security/content-farm-impersonates-60-plus-major-news-outlets-like-bbc-cnn-cnbc/>.
 - [53] F. T. Commission, Ftc permanently stops fake news website operator that allegedly deceived consumers about acai berry weight-loss products, <https://web.archive.org/web/20230620213056/https://www.ftc.gov/news-events/news/press-releases/2013/02/ftc-permanently-stops-fake-news-website-operator-allegedly-deceived-consumers-about-acai-berry>.
 - [54] M. Panahi, Llama-3-8b-instruct-v0.10, <https://huggingface.co/MaziyarPanahi/Llama-3-8B-Instruct-v0.10>.
 - [55] J. Lang, Z. Guo, S. Huang, A comprehensive study on quantization techniques for large language models, arXiv preprint arXiv:2411.02530 (2024).
 - [56] H. Wu, P. Judd, X. Zhang, M. Isaev, P. Micikevicius, Integer quantization for deep learning inference: Principles and empirical evaluation, arXiv preprint arXiv:2004.09602 (2020).
 - [57] UKPLab, Sentencetransformers, <https://github.com/UKPLab/sentence-transformers>.
 - [58] UKPLab, Sentencetransformers documentation, <https://sbnet.net/>.
 - [59] HuggingFace, Distilbert base uncased, <https://huggingface.co/distilbert/distilbert-base-uncased>.