Barkhatova, Elvira (2025) Media manipulation detection: Challenges and perspectives. Journal of Computer-Assisted Linguistic Research 9: 19-42. ISSN: 2530-9455.

Media manipulation detection: Challenges and perspectives

Elvira Barkhatova*
Immanuel Kant Baltic Federal University, Russia
*Corresponding author: elvira.barkhatova@gmail.com
Received: 28 September 2025 / Accepted: 25 October 2025 / Published: 20 November 2025

Abstract
The article deals with the issue of mass media manipulation and presents a prototype of an innovative browser extension specifically designed to identify biases and manipulative techniques across various dimensions, including lexical, syntactic, and pragmatic levels. This tool is designed not only to detect subtle forms of manipulation embedded within media narratives but also to empower users with an insightful understanding of these tactics. The primary objective of this program is to inform users regarding the potential threats linked with consuming media content, such as news articles or analytical political materials. By meticulously scrutinizing the language employed in these texts, the extension aims to uncover the underlying agendas that may distort public perception. The strategy involves the incorporation of an optimal combination of features, such as detection of emotionally charged vocabulary, biased framing, unsourced claims, ambiguities, and more. The article analyzes both the advantages and the drawbacks of the current approach and provides suggestions for further improvement of the future program.

Keywords: mass media, manipulation detection, logical fallacies, ambiguity, biased framing

1. Introduction

The study of manipulation in mass media has become increasingly urgent in the 21st century as digital platforms, political polarization, and technological advancements have transformed the information environment. Media messages no longer simply inform or persuade, but they frequently seek to manipulate, subtly guiding public opinion through covert techniques that exploit emotional, cognitive, and social vulnerabilities. The consequences are far-reaching: manipulation contributes to declining trust in media institutions and the spread of disinformation. For this reason, research into manipulative communication practices is not only a matter of academic interest but also a pressing societal necessity.

DOI: https://doi.org/10.4995/jclr.2025.24727
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

One of the key points to consider in this field is the distinction between persuasion and manipulation. Persuasion is generally understood as a transparent and legitimate form of influence, where arguments are presented openly, supported by evidence, and directed toward rational deliberation. As O'Keefe (2002) argues, persuasion respects the autonomy of the audience, enabling individuals to make informed choices based on reasoned evaluation. Manipulation, by contrast, operates in ways that are deceptive, coercive, or obscured. It often employs emotional appeals, cognitive biases, or the selective omission of information in order to steer individuals toward conclusions they might not otherwise reach (Corner 2007). Persuasion is essential for the exchange of competing ideas, whereas manipulation undermines informed decision-making and distorts public discourse. This makes manipulation not only distinct from persuasion but also more dangerous, particularly when it erodes collective trust in media and institutions.

Given these risks, there is a strong need for practical tools that help users recognize manipulation in real time. We argue that the development of user-friendly, lightweight systems such as browser extensions can make a tangible contribution to both personal media literacy and professional media analysis. The potential target users of such a system can be broadly divided into two groups. First, regular internet users, who might be interested in increasing personal awareness, improving their own media literacy, and making more informed choices. For this group, the system could provide diverse source recommendations, monitor the credibility of consumed content, and highlight manipulative patterns such as emotional or 'absolute' language. Second, media specialists, including journalists and researchers, would benefit from advanced analytical features. These may include tracking the accuracy of scientific claims, monitoring emotional language trends in coverage, analyzing shifts in source credibility, documenting misinformation patterns, comparing the use of fallacies across the web, and assessing the accuracy of fact-checking practices.

By addressing the needs of both audiences, such a tool would function not only as a safeguard against manipulative techniques but also as an educational and research instrument. In this sense, it bridges the gap between theoretical studies of manipulation and their practical applications.

2. Manipulation research landscape

Research into manipulative techniques in mass media discourse has expanded considerably in the XXI century, reflecting the growing complexity of political communication and digital media ecosystems. Media manipulation is not limited to overt propaganda but often manifests in subtle forms of framing, agenda-setting, emotional appeals, and technologically amplified disinformation. Contemporary studies highlight how these techniques shape interpretation and influence collective behavior.

Traditional mass media manipulation frequently operates through framing and agenda-setting. Framing refers to the selective emphasis of certain aspects of reality to guide interpretation,

often privileging specific problem definitions and causal explanations (Entman 1993; Matthes 2012; Karpf 2016). Agenda-setting, while closely related, determines which issues gain prominence in public debate, often privileging elite perspectives and marginalizing alternative views (McCombs 2014; Guo and Vargo 2017). These mechanisms illustrate how manipulative discourse may restrict interpretive freedom without overt deception.

The use of emotional appeals is another central manipulative strategy. Media scholars argue that affective communication, which actively exploits fear, outrage or pride, mobilizes audiences more effectively than rational argumentation (Wahl-Jorgensen 2019). The prevalence of emotionally charged discourse has fueled polarization in contemporary journalism, reinforcing group identities while delegitimizing opponents (Waisbord 2018). These affective dynamics blur the boundary between persuasion and manipulation, particularly when emotions are deliberately employed to interfere with critical judgment. The rise of social media has enabled computational propaganda, where automated bots, trolls, and algorithmic targeting amplify manipulative content. Marwick and Lewis (2017) further document how conspiracy communities and extremist groups weaponize digital means to circulate manipulative narratives that exploit identity politics.

In response, fact-checking initiatives have gained prominence. Manual fact-checking organizations such as *Media Bias/Fact Check*, *Verificado*, *Vera Files*, and *Full Fact* play a crucial role in verifying claims and exposing manipulative discourse. However, the scale of online manipulation has necessitated computational approaches. Preslav Nakov and colleagues have been at the forefront of developing automated propaganda and disinformation detection tools. Their work emphasizes the fine-grained identification of propaganda techniques such as appeal to fear, name-calling, and false dilemmas (Baly et al. 2018; Atanasova et al. 2019; Barrón-Cedeño et al. 2020). Through projects like the *Tanbih News Aggregator*, Zhang et al. (2019) propose methods to diversify and 'debias' media coverage. More recently, Nakov's research has addressed the integration of large language models into fact-checking pipelines (Su, Cardie, and Nakov 2023), claim normalization to support scalable verification (Sundriyal et al. 2023), and multilingual fact-checking frameworks for detecting bias and misinformation (Nakov et al. 2024). Shared tasks such as the CLEF CheckThat! Lab (Piskorski et al. 2024; Alam et al. 2025) have provided annotated corpora and evaluation benchmarks, establishing a robust infrastructure for propaganda and fact-checking research.

Complementing these efforts, new real-time detection systems are being developed to preempt manipulative campaigns. Ford et al. (2022) introduce the *Misinformation Early Warning System (MEWS)*, which is capable of detecting manipulation across text, images, and video in real time. Zhang et al. (2024) propose *FastForensics*, a lightweight media manipulation detection tool optimized for portability, allowing journalists and activists to authenticate content on the move. The need for such detection tools is largely explained by the fact that manipulation frequently involves identity deception. Tsikerdekis (2018) demonstrates how identity falsification on social media can be detected in real time, highlighting the role of impersonation in disinformation campaigns. Agravat et al. (2024) extend this work through machine learning models that

classify fake versus genuine social media profiles using both text and visual data. Similarly, Maathuis et al. (2023) apply deep learning to improve automated disinformation detection, highlighting the effectiveness of neural architectures for large-scale classification.

The literature illustrates that manipulative discourse in mass media encompasses both traditional rhetorical strategies (framing, agenda-setting, emotional appeals) and technologically amplified disinformation (computational propaganda, identity deception, manipulated images and videos). Counter-strategies have evolved accordingly: fact-checking organizations provide corrective interventions, while computational approaches, particularly those developed by Preslav Nakov and collaborators, enable large-scale detection of manipulative techniques. Recent innovations such as MEWS, FastForensics, and machine learning-based profile verification further highlight the hybrid ecosystem of solutions. The ongoing challenge lies in maintaining the balance between protecting audiences from manipulation while preserving legitimate persuasive communication.

3. Methods

The Media Bias Detector operates as a multi-layered Chrome extension,written in JavaScript, built on Manifest V3 architecture, coordinating several specialized components to analyze web content in real-time. At the system's core, a background service worker manages the extension's lifecycle and facilitates inter-component communication, while persisting user preferences through Chrome's synchronized storage API. The popup interface functions as the primary control panel, presenting analysis results and enabling settings configuration. Upon activation, the extension injects a content script that orchestrates the analysis pipeline by traversing the document object model and extracting textual content. This content script delegates specialized analytical tasks to three distinct modules: a sentiment dictionary that employs lexicon-based categorization with Set data structures to achieve O(1) lookup performance; an ambiguity detector that implements formal linguistic tests derived from Zwicky and Sadock's (1975) framework to identify lexical, structural, and contextual ambiguities; and a fact-checker that utilizes a multi-stage claim detection pipeline with weighted confidence scoring algorithms. The system presents its findings through color-coded text highlighting and interactive tooltips. To maintain browser responsiveness, the architecture implements chunked processing that handles content progressively, ensuring thorough analytical coverage across multiple dimensions of potential media bias even on content-intensive pages.

3.1. Test material

We start with a brief description of the material that was used for testing the individual features. The first step included creating artificial (or “mock”) texts in English that were designed to resemble a news article, but contain a high density of the features under the test. It was achieved by specifically asking DeepSeek to generate HTML files containing particular manipulative features that we aimed to test. The prompts used to generate the texts included the definitions and examples of the biases under test. These files were then opened in a browser

window where the extension was run. The number of files corresponds to the number of features tested (i.e. 7) in order to test the features separately. The idea behind this phase is to make sure the program functions in the first place.

The next step was testing the combination of features on real media texts. This part was more challenging because usually the manipulative techniques are more or less subtle, thus making it more complicated to find a highly biased text that would contain all the manipulative features built into our extension. So, we searched the web for the latest news and compiled a list of 12 potentially manipulative web pages (e.g., *The Federalist, Daily Caller, Breitbart*, and others).

3.2. Features: emotional language

As we have already found out, the use of emotionally charged vocabulary is a prominent strategy in manipulative communication. Such language relies on the affective dimension of discourse, appealing to pathos rather than logos. Terms like *"shocking," "outrageous," "heroic,"* or *"disastrous"* do not merely convey descriptive content; rather, they predispose the audience to evaluate the referent in a particular way. In the context of media texts, emotional vocabulary often functions to dramatize events, elevate ordinary occurrences, or intensify perceived threats.

To implement this feature in the current project, a frequency-based lexicon of emotional terms was extracted from the *News on the Web* (NOW) corpus. This source was selected due to its focus on contemporary media discourse, thereby ensuring contextual relevance for the detector. The vocabulary was organized into several functional groups, reflecting both frequency and semantic scope. For now the total number of lexical items is 180, and include the following subcategories: Core Power Emotions (18, e.g., *love* [150.2 per million], *happy* [82.7], *hope* [68.4], *fear* [59.1]); Dominant Emotional States (20, e.g., *sad* [48.9], *joy* [42.5], *tears* [45.2], *smile* [62.4]); Strong Emotional Reactions (37, e.g., *scream* [22.3], *devastated* [21.3], *excited* [19.8]); Common Emotional Descriptors (40, e.g., *furious* [14.8], *horrible* [14.2]); Specialized Emotional Vocabulary (20, e.g., *bittersweet* [1.1], *nerve-wracking* [1.0]); Physical Manifestations (15, e.g., *grin* [9.4], *sob* [7.1]); Complex Emotional States (15, e.g., *nostalgic* [3.6], *vulnerable* [5.4]); and Emotional Body Language (15, e.g., *eye-rolling* [1.8], *fist-clenching* [0.9]). We thought it reasonable to include these lexical items in the detector, but we don't deny the possibility of expanding it, if the speed of analysis is preserved, and the extension remains lightweight

So, the emotional vocabulary is identified through lexicon-based matching. A predefined set of emotional tokens drawn from both positive and negative domains is compared against normalized text segments. The system tokenizes sentences, applies word-boundary checks via regular expressions, and flags sentences containing one or more matches. Each occurrence contributes to an overall emotionality score, which is stored in structured output for subsequent visualization in the browser extension interface (Fig.1). This approach permits sentence-level highlighting of emotionally charged words (Fig.2), thereby enabling readers to distinguish between factual content and rhetorical amplification.

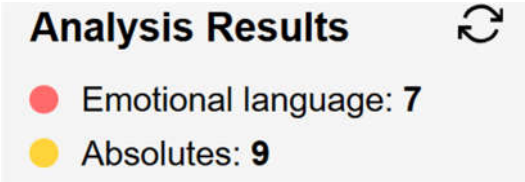

FIGURE 1. A SCREENSHOT OF AN EXTENSION INTERFACE CONTAINING EMOTIONAL LANGUAGE AND ABSOLUTES FEATURES

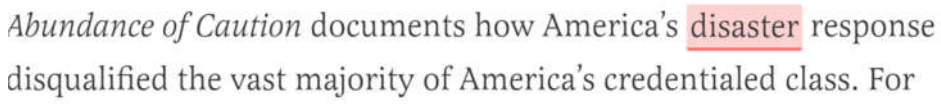

FIGURE 2. A SCREENSHOT OF A DETECTED 'EMOTIONAL' LEXICAL ITEM

3.3. Features: absolutes

Absolute expressions represent another salient mechanism of manipulation. These terms, such as *"always," "never," "everyone," "nobody," "all,"* and *"none,"* convey categorical certainty and exclude nuance. Their rhetorical function is to amplify claims by presenting them as universal truths, irrespective of empirical variation. In argumentative discourse, absolutes often simplify complex issues, discouraging critical evaluation by implying that counterexamples cannot exist.

The current version of the extension detects absolutes through a targeted lexical list of 26 items encoded within the content script. During analysis, each sentence is scanned for absolute markers using regular expressions that ensure word-boundary precision. Matches increment a per-sentence absolute count, which is subsequently stored in the annotation output. This process enables the extension to highlight sentences that rely on categorical generalization, thereby alerting users to potential oversimplification in the source material (Fig. 3).

As with emotional vocabulary, the approach is not without constraints. Idiomatic uses of absolutes (e.g., *"never mind"*) may yield false positives, while negated constructions (*"not everyone"*) can reduce universality but still be flagged. One of the further improvements may involve the incorporation of part-of-speech tagging or syntactic heuristics to improve discrimination between literal and idiomatic usage. Another possibility is contextual weighting (e.g., proximity to evidence or citations) that could reduce over-flagging in legitimate contexts.

FIGURE 3. A SCREENSHOT OF A DETECTED 'ABSOLUTE' LEXICAL ITEM

3.4. Features: unsourced claims

The increasing reliance on digital content for information has brought with it the widespread issue of unsourced claims, which can undermine the trustworthiness of information. Unsourced

claims are those statements that assert factual information without citing the source of that information, leaving readers unable to verify its validity. Common examples include broad declarations like "Studies show that coffee prevents cancer" or "Experts believe the economy will crash," where the absence of specific references such as study names, dates, or expert credentials makes it difficult to assess the reliability of the claims (Table 1). These types of claims are often vague and lack crucial details that could help verify their authenticity, such as the sample size of a study, the methodology used, or the qualifications of the purported experts.

Unsourced claim	**Properly sourced claim**
"Studies show that coffee prevents cancer"	"A 2023 Harvard study of 50,000 participants found coffee may reduce cancer risk by 15%"
Problem: Which studies? When? How many participants?	Good: Specific institution, date, sample size, precise finding
"Experts believe the economy will crash"	"Dr. Jane Smith, Chief Economist at Goldman Sachs, predicts economic downturn"
Problem: Which experts? What's their expertise?	Good: Named expert, credentials, specific prediction
"Research indicates this diet works"	"A peer-reviewed study published in JAMA showed this diet effective for 78% of participants"
Problem: What research? Peer-reviewed? Sample size?	Good: Publication type, journal name, specific results

TABLE 1. COMPARATIVE CLAIM SOURCING EXAMPLES

In contrast, properly sourced claims provide clear, verifiable information that enhances their credibility. For example, a claim like "A 2023 Harvard study of 50,000 participants found coffee may reduce cancer risk by 15%" offers a specific institution (Harvard), a date (2023), a sample size (50,000 participants), and a precise finding (15% reduction in cancer risk). Similarly, a properly sourced claim such as "Dr. Jane Smith, Chief Economist at Goldman Sachs, predicts economic downturn" includes the expert's name, professional title, and institutional affiliation, allowing readers to verify both the claim and the authority behind it.

To tackle the problem of identifying unsourced claims, our tool uses regular expressions (regex) to scan textual content for phrases commonly associated with unsourced claims. These include, but are not limited to, vague authority appeals such as "experts say," "studies show," or "officials report," which suggest that the claim is based on an authoritative source without specifying who those authorities are. Other patterns the system looks for include vague statistical claims, such as "most people," "recent data," or "several reports," which can imply broad consensus without providing actual data or references. Additionally, the system identifies passive voice constructions (e.g., "it is believed," "it is shown") that obscure the source

of the information, and definitive unsourced claims like "the fact is" or "it is a fact that," which present the information as irrefutable without any supporting evidence. Overall, the number of patterns is 26: 17 for the vague authority appeals, and 9 for vague statistical claims.

In addition to its pattern-matching capabilities, the script includes test cases generated by AI that simulate different types of claims, ensuring that the system can handle a variety of real-world scenarios. These test cases include examples of properly sourced claims and unsourced claims. Additionally, it evaluates passive voice claims, which are especially difficult to identify because they obscure the agent of the action, thus making it harder to pinpoint the source.

The modular structure of the unsourced claim detector also allows for easy customization and expansion. The detection logic can be adjusted or extended by adding more regular expression patterns to capture new types of unsourced claims as they emerge.

3.5. Features: biased framing

Biased framing represents a subtler yet highly consequential form of media manipulation. Whereas emotionally charged vocabulary operates at the level of lexical intensity and quantification, biased framing restructures discourse at the level of evaluation and perspective, predisposing audiences toward particular interpretations. This strategy embeds ideological judgments within descriptive language, guiding readers to perceive actors, policies, or events in normative terms while preserving the surface appearance of objective reporting. Phrases such as *"radical agenda," "so-called expert,"* or *"common-sense reform"* illustrate how evaluative framing introduces implicit value judgments that shape perception without overt argumentation.

Within our detector, the lexicon was divided into two broad categories: negative or pejorative framing (60 lexical items) and positive or laudatory framing (40 lexical items). The first category is commonly used to criticize, vilify, or delegitimize actors, employing terms such as *radical, extremist, fanatic, zealot, militant, terrorist, bigot, racist,* and *xenophobe*. These lexical items cast subjects as dangerous, illegitimate, or morally corrupt, thus preventing any sympathetic interpretation. By contrast, the second category confers legitimacy and prestige, praising or glorifying subjects with terms such as *hero, patriot, defender, savior, martyr, visionary, reformer, freedom fighter, icon, legend,* and *genius*. In this case, framing elevates individuals or groups by appealing to cultural archetypes and valorizing narratives, often obscuring the complexities of their actions or policies. Together, these contrasting vocabularies reveal the dual nature of biased framing: it functions simultaneously to demonize opponents and to glorify allies, thereby reinforcing ideological polarization.

The implementation of biased framing detection within our extension focuses on phrase-level pattern recognition. A curated set of framing expressions, supported by regular expression templates, is applied to text segments. The detector specifically searches for collocations of evaluative adjectives with institutional or policy-related nouns (e.g., *"failed policy," "corrupt leader"*), as well as templated constructions such as *"so-called"* followed by a profession or title. When such patterns are detected, the system flags them.

By making biased framing explicit, the system not only exposes ideological manipulation but also equips users to critically evaluate how language shapes perception.

3.6. Features: logical fallacies

Logical fallacies represent another critical feature of unreliable or manipulative argumentation. When initialized, the detector loads two primary resources: a library of fallacy patterns (Fig. 4), which is a dictionary of 15 fallacy categories (such as ad hominem or strawman), and 27 contextual indicators, which serve as secondary cues to strengthen detection. Each fallacy category contains carefully crafted regular expressions designed to recognize linguistic markers of specific fallacies. For example, the *ad hominem* category searches for patterns that attack the individual rather than their argument, while the *strawman* category identifies instances where a position is misrepresented to make it easier to refute.

```
// Bandwagon Fallacy - Arguing something is true because many people believe it
bandwagon: {
  patterns: [
    /\beveryone\s+(knows|believes|thinks|agrees)\s+that\b/gi,
    /\bmost\s+people\s+(would\s+agree|believe|think)\s+that\b/gi,
    /\ball\s+the\s+(smart|intelligent|reasonable)\s+people\b/gi,
    /\bif\s+everyone\s+else\s+is\s+doing\s+it\b/gi,
    /\bthe\s+majority\s+of\s+(people|Americans|voters)\s+(believe|support)\b/gi,
  ],
  keywords: new Set(["popular", "mainstream", "consensus", "majority", "widespread"]),
  description: "Arguing that something is true or right because many people believe it",
},
```

FIGURE 4. A CODE SNIPPET ILLUSTRATING PATTERNS OF BANDWAGON FALLACY

Once the text under analysis is opened, the detector begins by scanning it against the predefined regex patterns. For every fallacy type, the algorithm loops through the associated patterns and runs searches within the text. When a match is found, that portion of the content is flagged as a potential fallacy. To enhance the accuracy of this process, the detector includes a contextual analysis layer (Fig. 5). This secondary step checks for additional textual cues that often signal fallacious reasoning, such as intensifiers ("obviously," "clearly"), sarcasm markers, or rhetorical framing phrases ("as everyone knows"). These contextual signals do not replace the initial regex detection but instead add weight to the likelihood that a fallacy has been correctly identified.

```
// Qualifier words that may indicate weak reasoning
qualifiers: new Set([
  "obviously",
  "clearly",
  "certainly",
  "undoubtedly",
  "without question",
  "everyone knows",
  "it's common sense",
  "any fool can see",
```

FIGURE 5. A CODE SNIPPET ILLUSTRATING SOME TEXTUAL CUES

The output of the system is both diagnostic and explanatory. Detected fallacies are labeled with their category and accompanied by short educational descriptions (Fig. 6). In total, the detector covers a broad range of well-documented fallacies, including appeal to authority (e.g., *Top economists say this plan will work, so it must be right*), appeal to emotion (e.g., *Think of the children - how could you oppose this policy?*), circular reasoning (e.g. *The candidate is trustworthy because she is honest, and we know she is honest because she is trustworthy*), red herring (e.g., *Why worry about climate change when we still have unemployment to fix?*), bandwagon appeals (e.g., *Everyone knows this policy is the best solution*), hasty generalization (e.g., *I met some politicians from that party - they are all corrupt*), false cause (e.g., *Ever since the new mayor took office, crime has dropped dramatically, so she must be the reason*), appeal to ignorance (e.g., *No one has proven that aliens do not exist*), tu quoque (whataboutism) (e.g., *You accuse our country of human rights violations, but your own record on immigration is terrible*), no true Scotsman (e.g., *No true patriot would question the decisions of the government*), and loaded questions (e.g., *When did you stop misleading the public about your finances?*) Each detection can be paired with concrete examples to illustrate how the fallacy operates in practice, thereby reinforcing the educational aspect of the tool.

The design rests on three complementary pillars: pattern recognition (via regex), contextual analysis (via surrounding text cues), and explanations (via tooltips or annotations). Together, these ensure the system is not only capable of identifying fallacies with reasonable precision but also of functioning as a teaching aid for users seeking to improve their critical thinking skills. By flagging flawed reasoning in real time and providing accessible explanations, the logical fallacy detecting feature serves both as a credibility filter for argumentative text and as an instructional resource in logic and rhetoric.

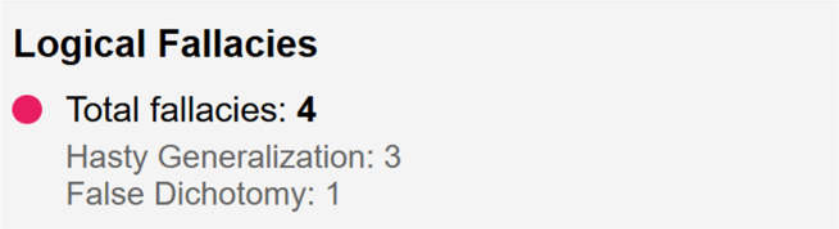

FIGURE 6. A SCREENSHOT OF AN EXTENSION INTERFACE CONTAINING LOGICAL FALLACIES

3.7. Features: ambiguity analysis

Here we come to the most troublesome features of our detector - ambiguity analysis and fact-checking (the subsequent section). Since ambiguity detection is one of the most complicated tasks within our project, we thought it reasonable to start with a quick overview of the phenomenon.

Beyond its theoretical significance, ambiguity plays a crucial role in media and political discourse, where it is frequently deployed as a manipulative device. As Sennet (2023) emphasizes, ambiguity arises not merely from linguistic indeterminacy but from the interpretive latitude available to hearers and readers. This very openness makes it an attractive tool for strategic communication in news and political commentary. Ambiguous constructions

allow communicators to maintain plausible deniability, appeal simultaneously to multiple audiences, or obscure accountability for controversial claims. Within philosophy of language and formal semantics, *ambiguity* designates a lexical property whereby a single expression encodes multiple, distinct and conventionally established meanings. According to Stanford Encyclopedia of Philosophy, ambiguity must be sharply distinguished from several superficially similar but theoretically divergent sources of interpretive multiplicity.

Vagueness arises from semantic indeterminacy, typically manifested in borderline cases or fuzzy boundaries (Gillon 1990). For instance, *bald* exhibits vagueness without ambiguity, while *bat* exhibits ambiguity without vagueness. Context sensitivity involves systematic variation in semantic content as a function of contextual parameters, absent any change in lexical convention. Indexicals such as *I* are paradigmatic: the referent shifts across utterances, yet the expression retains a single meaning. In contrast, lexical items such as *bank* are ambiguous, not indexical. While contextual information may serve as a disambiguator, the defining property of ambiguity is not context dependence but the existence of multiple entrenched meanings (Sennet 2023).

Under-specification and sense generality arise when an expression leaves certain features of its interpretation unspecified while retaining a unitary sense. For instance, *one of my sisters* fails to determine a unique referent but is not ambiguous. Similarly, *filet mignon* does not lexically encode a distinction between raw and cooked, though the extension of the term may diverge across pragmatic settings. Although sense generality is often conflated with vagueness, its theoretical profile is distinct: it involves under-determination rather than indeterminacy or lexical multiplicity.

Sense and reference transfer introduces further complications: expressions such as *I am parked on G St.* shift reference from speaker to associated vehicle. These cases resist straightforward classification, as the interpretive mechanism is not lexically encoded ambiguity but pragmatic transfer of sense or reference, interacting with syntactic and semantic constraints in non-trivial ways (Sennet 2023).

Notably, these phenomena (vagueness, context sensitivity, under-specification, and reference transfer) may co-occur in a single utterance, but their theoretical diagnostics, semantic mechanisms, and explanatory demands diverge substantially. Maintaining clear distinctions among them is therefore essential for a principled account of ambiguity.

Ambiguity is usually divided into three broad groups: lexical, syntactic, and pragmatic. Lexical ambiguity is often harnessed in journalistic headlines through polyvalent terms such as *crisis, reform,* or *radical*, which can admit both neutral and evaluative readings depending on the reader's predispositions. This aligns with observations in framing research, where evaluative vocabulary operates to pre-structure interpretation under the guise of objectivity. Likewise, syntactic ambiguities in political statements, such as deliberate scope vagueness (e.g. "*Not every*

regulation hurts business"), permit speakers to reinterpret their words retroactively, shifting their stance as circumstances evolve (cf. Jurafsky & Martin 2024; Wasow 2002).

Pragmatic ambiguity is particularly salient in manipulative discourse. Politicians and media outlets may formulate statements whose intended force is left ambiguous, enabling sympathetic audiences to infer one interpretation while opponents register another. This phenomenon has been studied in works on political discourse (Kenzhekanova and Zhanabekova 2015). Presuppositional ambiguity further amplifies manipulative potential: phrases such as *"the government's failed policy"* not only assert failure but also presuppose that the initiative was indeed a government policy. As Entman's (1993) framing theory underlines, such linguistic packaging embeds evaluative judgments, thereby guiding audience interpretation without overt persuasion.

Recent computational research on propaganda and fact-checking highlights that ambiguity often co-occurs with other manipulative techniques. Nakov et al. (2024), for instance, note that fine-grained propaganda detection must account for ambiguous phrasing that conceals intent, since automated systems otherwise risk missing covert manipulative strategies.

Across these domains, the central challenge lies in differentiating genuine ambiguity (multiple entrenched semantic values) from under-specification, context-sensitivity, or pragmatic enrichment. For analysts and media-literacy tools, detecting and flagging ambiguous constructions is essential, as it reveals how linguistic indeterminacy can be employed to distort public opinions while maintaining the facade of neutrality.

The detection of ambiguity in natural language has long relied on diagnostic traditions that attempt to distinguish genuine semantic 'bifurcation' from mere vagueness or underspecification. Among the most influential contributions is Zwicky and Sadock's *Ambiguity Tests and How to Fail Them* (1975), which argues that ambiguity must be established through systematic interpretive contrasts rather than introspection alone. Central to their framework is the observation that judgments of absurdity, such as zeugma or syllepsis, often expose competing readings that cannot be simultaneously sustained. While powerful, these diagnostics remain challenging in borderline or philosophical cases, and any computational implementation must therefore balance linguistic sensitivity with robustness against interpretive variability.

Now let us turn to the current architecture of our ambiguity analysis feature. Our initial intention was to translate canonical ambiguity diagnostics into computationally tractable procedures. Conjunction reduction tests, for instance, expose zeugma when a word cannot simultaneously sustain multiple readings (e.g. *She went to the bank and deposited a check, and so did he*, well-suited under "financial institution" but not "riverbank"). Ellipsis tests similarly detect ambiguity when elliptical constructions cannot be resolved under a single sense (e.g. *John saw her duck and Bill did too*). Contradiction tests are also operationalized and are aimed at revealing apparent contradictions that dissolve under distinct readings (e.g. *The chicken is ready to eat, but*

the chicken is not ready to eat). Finally, definitional tests evaluate whether a single unified lexical entry suffices; their failure, as in *bank*, indicates true ambiguity rather than contextual flexibility.

Building on these diagnostics, the detector classifies ambiguity into several linguistically motivated categories (Fig. 7). Lexical ambiguity encompasses homonymy (e.g. *bank, bat, bark*) and polysemy (e.g. *head, run, play*), including idiomatic expressions such as *kick the bucket*, which defy compositional analysis. Structural ambiguity emerges from multiple possible logical forms for a single surface string, including attachment ambiguities (e.g. *John floated the boat between the rocks*), scope alternations (e.g. *Every woman squeezed a man*), and quantifier interactions (e.g. *Some politician admires every journalist*). Contextual ambiguity is captured by expressions whose interpretation depends on discourse or modal context, as with *any* (i.e. free-choice vs. universal) or *John must be at home* (i.e. epistemic vs. deontic). Special attention is devoted to zeugma (e.g. *She opened her mind and her wallet*) and contradiction patterns (e.g. *All that glitters is not gold*, ambiguous between "not all" and "none").

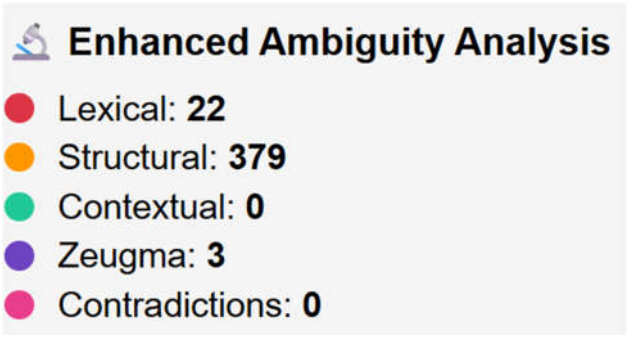

FIGURE 7. A SCREENSHOT OF AN AMBIGUITY DETECTOR INTERFACE

Results are presented with interpretability in mind. Each flagged instance is accompanied by explanatory tooltips specifying the ambiguity type (Fig. 8 and 9). In addition, the system enumerates all attested readings, e.g., the four interpretations of *I love you too* identified by Bach (1982), enabling users to visualize the interpretive space rather than simply receiving a binary flag.

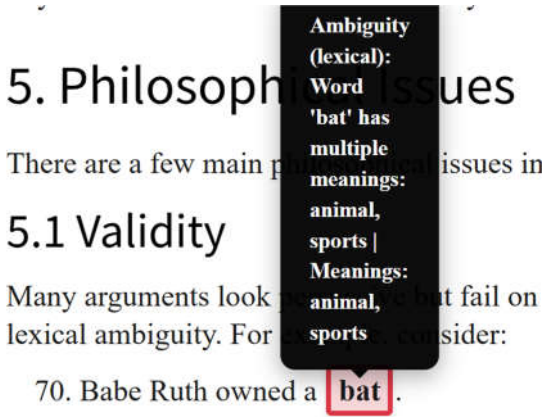

FIGURE 8. A SCREENSHOT OF AN EXAMPLE OF LEXICAL AMBIGUITY DETECTED IN A TEXT

Unlike black-box statistical classifiers, the architecture is explicitly linguistically principled. By embedding Zwicky and Sadock's diagnostic tradition into a computational environment, the detector is supposed to distinguish genuine ambiguity from phenomena such as vagueness or sense generality. The ambiguity detection module operates through direct pattern-based identification of lexical and structural overlaps that signal potential ambiguity. Specifically, the system iterates through the input text to locate predefined surface strings known to trigger polysemy or zeugmatic constructions (e.g., *bank, light*), comparing each candidate phrase to a lexicon of ambiguous collocations. Upon detection, the algorithm records contextual information such as the ambiguous token, its surrounding phrase, and positional index within the text. This rule-based approach prioritizes precision through explicit string matching rather than probabilistic inference. In this way, the system should advance both the theoretical clarity and the practical detectability of ambiguity. However, as we will see in the Results and Discussion section, not everything runs smoothly yet.

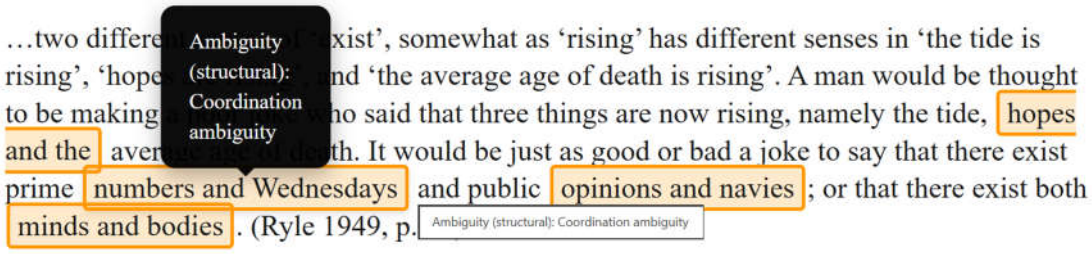

FIGURE 9. A SCREENSHOT OF AN EXAMPLE OF STRUCTURAL AMBIGUITIES DETECTED IN A TEXT

3.8. Features: fact-checking

Another complex issue is implementing a properly working fact-checking system. While our ambiguity detection module diagnoses interpretive multiplicity at the lexical, syntactic, and pragmatic levels, linguistic ambiguity is only one vector by which natural language contributes to misinformation or misunderstanding. Another crucial dimension is the factual status of utterances: a sentence may be structurally unambiguous and semantically clear, yet nevertheless encode a false or misleading claim. To address this challenge, we devised a fact-checking subsystem that operates in parallel with ambiguity detection.

The fact-checking component of the system is designed to complement the ambiguity detection module by providing external validation of claims encountered in media discourse. Its guiding principle is methodological transparency: all data sources are freely accessible, and no reliance is placed on proprietary datasets or commercial APIs. This ensures that the extension remains lightweight and accessible to both everyday users and research communities.

This module is still under construction, so the current logic behind the fact-checking component of our browser extension has the following look. The system integrates curated repositories of verified claims via RSS feeds from established fact-checking organizations. Primary sources include FactCheck.org (specializing in U.S. political discourse), Snopes (covering health

misinformation, urban legends, and general claims), and PolitiFact (annotated political claims with "Truth-O-Meter" ratings). These sources collectively guarantee broad topical coverage across politics, health, economics, and science. Optional integration with NewsAPI (free tier: 1,000 requests/day) might further extend cross-verification capacity. This commitment to open and reproducible resources parallels recent calls for transparency in computational fact-checking.

The fact-checking pipeline is sentence-based, mirroring the granularity of the ambiguity detection framework. Each input text is segmented into sentences, which are classified as potential factual claims. Content words are extracted, lemmatized, and normalized (i.e., standardized in terms of spelling variants, lowercasing, etc.), while stopwords are filtered out to improve robustness. Matching is then performed via fuzzy string alignment, weighted term overlap, and contextual filtering, which prevents false positives caused by homonymy (e.g., *bank* as financial vs. riverbank). Once candidate alignments are identified, claims are mapped to labels drawn from the fact-checking repositories. This approach is supposed to ensure that factual verification is methodologically aligned with interpretive testing: just as ambiguity detection distinguishes competing readings via formal diagnostics, fact-checking distinguishes factual outcomes via structured external evidence.

Early iterations of the system faced practical limitations, most notably the tendency to default all sentences to "unverified." This was primarily caused by CORS (Cross-Origin Resource Sharing) restrictions blocking RSS feed access and by simplistic keyword matching. We attempted to introduce the following design improvements: 1) CORS proxy integration via *api.allorigins.win* to enable seamless in-browser RSS retrieval; 2) enhanced matching logic incorporating fuzzy alignment and weighted overlap; 3) local mock database fallback, consisting of entries spanning politics, health, economics, and science. This fallback ensures that classification remains available even when live data sources are inaccessible, much as the ambiguity module provides principled defaults for borderline cases.

The fact-checking results are communicated through intuitive visual coding. Paragraphs are put into frames and highlighted in green (verified), red (false/disputed), yellow (mixed/partially true), or gray (unverified) (Fig. 10). The final results are presented in percentages, and the approximate credibility score is elicited (Fig. 11).

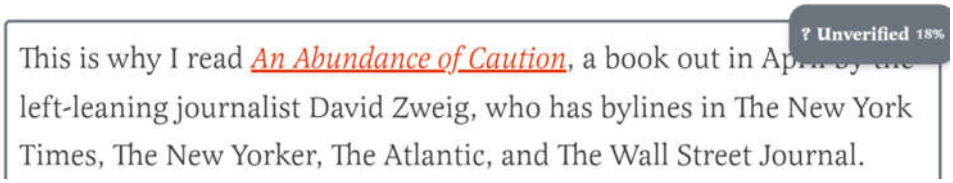

FIGURE 10. A SCREENSHOT OF AN EXAMPLE OF AN UNVERIFIED CLAIM

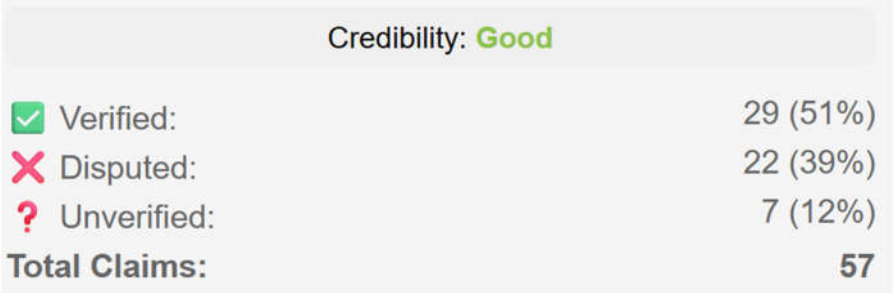

FIGURE 11. A SCREENSHOT OF THE CREDIBILITY SCORE INTERFACE

To ensure usability, the module is being designed to be responsive and scalable. Parallel API querying is supposed to reduce latency, smart caching - to minimize redundant lookups, and rate limiting - to ensure compliance with free-tier constraints. Ideally, fallback mechanisms should guarantee that every sentence receives a classification, thereby preserving user trust and interface consistency.

The fact-checking module is conceptualized as a complement to ambiguity detection, with integration occurring at two levels. First, both modules operate at the sentence level, allowing unified annotation of interpretive ambiguity and factual validity for the same textual unit. Second, they support complementary disambiguation: where ambiguity yields multiple possible readings, fact-checking tests the truth of each interpretation against external knowledge bases; where claims are unambiguous, fact-checking alone suffices. For example, the sentence *"The bank is collapsing"* is initially ambiguous between "financial institution" and "riverbank." Once the ambiguity module disambiguates to the financial sense, the fact-checking pipeline verifies whether such an event has been reported. Conversely, *"Vaccines cause autism"* is semantically unambiguous but factually false, a judgment made possible by external repositories. This dual framework (interpretive precision plus factual verification) is aimed at ensuring that the system not only highlights how media discourse can mislead but also grounds those insights in independently verifiable evidence.

4. Results and Discussion

This section analyzes the performance of the proposed media bias detection prototype and discusses the implications of the results, highlighting both the effective aspects of the current design as well as the areas requiring future improvement.

4.1. Emotional vocabulary, absolutes, and biased framing

The emotional vocabulary detector, built on a frequency-based list derived from the NOW corpus, demonstrated strong suitability for media analysis, outperforming generic sentiment lexicons.

Although effective in detecting overt emotional markers, the method has limitations. Lexicon-based approaches are vulnerable to false positives (e.g., idiomatic expressions or quoted

material) and false negatives (e.g., multiword evaluative phrases not included in the dictionary, like *The speech left many on edge*). For instance, polysemy can lead to false positives when emotionally charged words are used in neutral contexts (e.g., *The minister spoke about the pain points of the supply chain*). The solution might be to check nearby attribution (e.g. presence of a citation/source) and lower the score when a reliable source token appears within the same sentence. In addition, to enhance accuracy the system could be augmented with n-gram matching for phrases, and eventually, lightweight sentiment analysis or transformer-based embeddings to better capture contextual sensitivity.

Similarly, the detector for absolute terms (e.g., *always, never*) functioned reliably, largely because the finite inventory of such terms allows for high-precision pattern matching. However, there are some weak aspects as well, for instance, false positives with idioms and non-propositional uses, and negation interplay (e.g., double negatives*: There is nothing not worth discussing*). These issues might be tackled by employing small context windows and POS heuristics (e.g. checking for the nearest verb phrase) and making sure the algorithm correctly handles contexts like "not everyone" (which actually weakens universality, rather than asserting it). Also, flagging alone is coarse, so it is useful to pair absolute detection with a request for supporting evidence (e.g., look for citations nearby) to better prioritize user attention.

The biased framing detector, which also relies on a database of phrases, performed adequately as well. However, this method revealed inherent limitations common to lexicon-based approaches. Novel or subtle framings that deviate from known patterns are likely to be missed (e.g., *Experts close to the administration claim the numbers are accurate*). Furthermore, some evaluative terms are context-dependent and may not constitute bias in all instances (e.g., *He calls for a progressive tax structure*). Enhancements might include integrating distributional semantics to identify novel evaluative collocations or applying supervised models trained on annotated corpora of framing devices. Such extensions would enhance the robustness and adaptability of the system to evolving media language.

4.2. Unsourced claims and logical fallacies

The use of regex patterns proved to be a simple yet effective method for identifying phrases indicative of unsourced claims. The modular class structure of the unsourced claim detector is well-suited for integration into a browser extension, facilitating the processing of web-based content like news articles and blog posts.

Despite its effectiveness, the current rule-based approach may miss claims presented through more complex or indirect sentence structures (e.g. *The rumors of corruption have been widely spread through the media*). To increase recall, the pattern-matching logic could be refined, and machine learning techniques could be incorporated to handle a wider variety of phrasings while maintaining precision. It could also be useful to make the detector generate a comprehensive report that would include such elements as: 1) credibility level (e.g., low, high) - a qualitative assessment of the overall reliability of the text based on the balance of sourced and unsourced

claims; 2) overall score - numerical score that reflects the proportion of sourced versus unsourced claims within the text; 3) list of unsourced claims - a detailed list of all unsourced claims identified during the analysis, including an explanation of why each claim was flagged as unsourced.

As for logical fallacies, the prototype successfully identifies them by pattern-matching common linguistic markers. For example, the system flags direct ad hominem attacks like "you are stupid" using predefined regex patterns. This rule-based method provides a transparent and computationally lightweight detection mechanism.

However, its effectiveness is directly tied to the coverage of its pattern library. The system is vulnerable to paraphrasing; an argument with the same fallacious intent but different phrasing (e.g., "your argument reflects poor reasoning") may not be detected because the exact regex isn't matched. In contrast, understanding meaning would require some form of semantic analysis (like natural language models do), where the system can infer intent even if the words are different. To solve this problem, the pattern library could be expanded using synonym sets from lexical databases like WordNet or through embedding-based similarity to capture a broader range of expressions. For example, instead of hardcoding "stupid|dumb", we could include synonyms from a lexical database like WordNet or embeddings-based similarity. Hopefully, in this way, synonymous lexical items (e.g., "foolhardy", "clueless", "uninformed") could also be caught. Incorporating contextual rules (e.g., detect "you're wrong because you're X" where X is derogatory) that analyze surrounding words could also improve detection robustness.

4.3. Ambiguity detection and fact-checking

As a prototype, the system has several areas that require significant refinement. Ambiguity detection and fact-checking proved to be the most problematic components so far.

The current method for calculating an ambiguity score generates a high rate of false positives and is not yet reliable. For example, in terms of false positives, the most vulnerable aspect so far is detecting structural ambiguity. It is obvious that the system flags constructions that are not even close to being truly ambiguous. This substantially misleads the user. So, the underlying algorithm for structural ambiguity risk analysis requires a fundamental redesign to become practically useful. Moreover, for some reason, there are issues with detecting contextual ambiguity and contradictions, so we definitely need to inspect the errors more thoroughly and potentially revise the detection rules. In addition, zeugma detection needs improvement (e.g. expanding the pattern library) as not all types of zeugma are identified properly.

Another issue lies in the visualization of ambiguity analysis results. It seems like the summary is either not well-connected to the scanning procedures, or the final result is not being calculated properly. We argue that parameters such as the ambiguity score and risk level (Fig.

12) should prove particularly useful for the extension users, so we will attempt to preserve them. Therefore, a further inspection is necessary in this case.

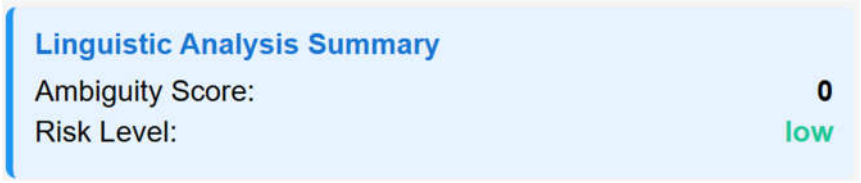

FIGURE 12. A SCREENSHOT OF A PROBLEMATIC RESULTS OUTPUT

The fact-checking component remains rather superficial, producing both false positives and false negatives that undermine its objectivity. For example, as it is visible from Figures 13 and 14, the distinction between "mixed" and "half-true" claims is still obscure and needs a fix.

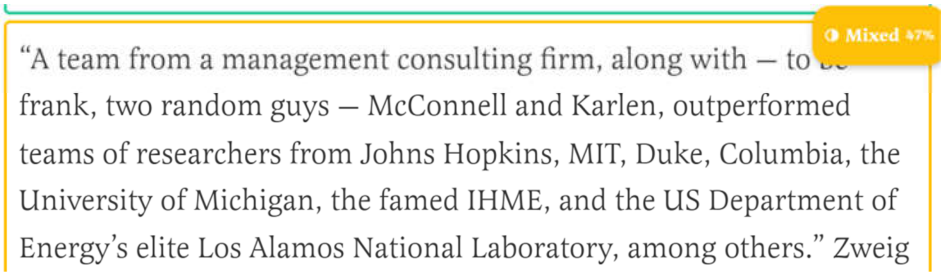

FIGURE 13. A SCREENSHOT OF A CLAIM FLAGGED AS MIXED

FIGURE 14. A SCREENSHOT OF A CLAIM FLAGGED AS HALF-TRUE

Also, the current design lacks the desired level of transparency and proper visualization. For example, it would be ideal to enable the user to trace classifications back to their sources by revealing detailed metadata, including the original fact-check URLs. Explanational tooltips, already successfully employed for logical fallacies detection, could be implemented in the fact-checking component as well. An architectural redesign is likely necessary to improve the accuracy of this feature.

The general question is how many analysis results the regular user will ultimately need. On the one hand, if the number of features is low, the detector is too superficial and doesn't really make any sense. On the other hand, if the features list is too extensive, how should the user navigate it and make the final conclusion? Which features can be summarized into some kind of

"verdict", and which ones should be left for the user to discover? These questions remain open for now.

5. Conclusion

The given study presented and evaluated a prototype browser extension designed to detect manipulative strategies and biases in mass media texts. By integrating multiple analytical features ranging from emotional vocabulary and absolutes to biased framing, unsourced claims, logical fallacies, ambiguity, and fact-checking, the system is aimed at bridging the gap between theoretical research on manipulation and practical tools for enhancing media literacy. The results indicate that several modules, particularly those based on lexicon-driven and regex-based approaches (e.g., emotional vocabulary, absolutes, biased framing, and logical fallacies), already demonstrate substantial utility for both everyday users and professional analysts. At the same time, more complex components, in particular, ambiguity detection and fact-checking, highlight the inherent challenges of modeling subtle semantic phenomena in computational systems.

As we have already shown in the previous sections, not everything runs smoothly yet. Emotional vocabulary, absolutes, biased framing, and logical fallacies operate reliably, though they could benefit from minor refinements. The unsourced claims detector is functional but requires broader coverage and moderate enhancements, i.e., some "medium-level" improvements. By contrast, ambiguity analysis and fact-checking remain underdeveloped and require significant redesign and tighter integration with the system as a whole. Although the prototype remains at an early stage, we are hopeful that it has potential for future refinement and expansion.

The contribution of this work is two-sided. First, it demonstrates the feasibility of a lightweight, user-friendly extension capable of providing real-time analysis of manipulative language strategies. Second, it exposes critical limitations of rule-based and lexicon-dependent methods when confronted with context-sensitive, synonym-rich, or structurally complex discourse. These findings underline the necessity of hybrid approaches that combine linguistically principled diagnostics with advances in natural language processing, such as embeddings-based similarity and transformer-based models, while maintaining transparency and interpretability for regular users.

We suggest that future research should concentrate on refining the accuracy of ambiguity and fact-checking modules, improving visualization and explainability, and conducting user-centered evaluations to balance analytical depth with interface usability. Ultimately, tools of this kind can contribute not only to individual critical media engagement but also to broader societal resilience against manipulation and disinformation in public discourse.

Acknowledgements

The publication was prepared within the framework of the RSF (Russian Science Foundation) project 22-18-00594 "Cognitive Models of Identification and Counteraction to Manipulation in the Media Space". I express my gratitude to all the organizers and participants of the 12th International Conference on Meaning and Knowledge Representation (MKR 2025) for inspiration.

References

Agravat, Aniket, Umang Makwana, Sahil Mehta, Devashish Mondal, and Sushant Gawade. 2024. "Fake Social Media Profile Detection and Reporting Using Machine Learning." *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)* 4 (5): 1103–1112. https://www.scribd.com/document/752210797/Fake-Social-Media-Profile-Detection-and-Reporting

Alam, Firoj, Julia Maria Struß, Tanmoy Chakraborty, Stefan Dietze, Salim Hafid, Katerina Korre, Arianna Muti, Preslav Nakov, Federico Ruggeri, Sebastian Schellhammer, Vinay Setty, Megha Sundriyal, Konstantin Todorov, and Venktesh V. 2025. "The CLEF-2025 CheckThat! Lab: Subjectivity, Fact-Checking, Claim Normalization, and Retrieval." In *Advances in Information Retrieval*. Springer. https://arxiv.org/pdf/2503.14828

Atanasova, Pepa, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. "Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims." In *CLEF 2019 Working Notes*. CEUR Workshop Proceedings 2380. https://ceur-ws.org/Vol-2380/paper_269.pdf

Bach, Kent, and Robert M. Harnish. 1982. Linguistic Communication and Speech Acts. Cambridge, MA: MIT Press.

Baly, Ramy, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. "Predicting Factuality of Reporting and Bias of News Media Sources." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3528–3539. Brussels: Association for Computational Linguistics. https://aclanthology.org/D18-1389/

Barrón-Cedeño, Alberto, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. "CheckThat! at CLEF 2020: Enabling the Automatic Identification and Verification of Claims in Social Media." In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of CLEF 2020*, 499–507. Lecture Notes in Computer Science 12260. Cham: Springer. https://doi.org/10.1007/978-3-030-45442-5_65

Corner, John. 2007. "Mediated Politics, Promotional Culture and the Idea of 'Propaganda.'" *Media, Culture & Society* 29 (4): 669–677. https://doi.org/10.1177/0163443707078428

Entman, Robert M. 1993. “Framing: Toward Clarification of a Fractured Paradigm.” *Journal of Communication* 43 (4): 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

FactCheck.org. 2025. Accessed September 27, 2025. https://www.factcheck.org/

Ford, Trenton W., Michael Yankoski, William Theisen, Tom Henry, Farah Khashman, Katherine Dearstyne, Tim Weninger, and Pamela Bilo Thomas. 2022. “MEWS: Real-time Social Media Manipulation Detection and Analysis.” In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Trac*k, 325–329. Proceedings of Machine Learning Research 176. PMLR. https://proceedings.mlr.press/v176/ford22a.html

Gillon, Brendan S. 1990. “Ambiguity, Generality, and Indeterminacy: Tests and Definitions.” *Synthese* 85 (3): 391–416. https://doi.org/10.1007/BF00484835

Guo, Lei, and Chris J. Vargo. 2017. “Global Intermedia Agenda Setting: A Big-Data Analysis of International News Flow.” *Journal of Communication* 67 (4): 499–520. https://doi.org/10.1111/jcom.12311

Jurafsky, Daniel, and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd ed. Draft. Upper Saddle River, NJ: Prentice Hall. https://web.stanford.edu/~jurafsky/slp3/

Karpf, David. 2016. *Analytic Activism: Digital Listening and the New Political Strategy*. New York: Oxford University Press.

Kenzhekanova, Kuralay, Magulsim Zhanabekova, and Tolkyn Konyrbekova. 2015. “Manipulation in Political Discourse of Mass Media.” *Mediterranean Journal of Social Sciences* 6(3): 245–251. https://www.richtmann.org/journal/index.php/mjss/article/view/7013

Maathuis, Clara, Iddo Kerkhof, Rik Godschalk, and Harrie Passier. 2023. “Design Lessons from Building Deep Learning Disinformation Generation and Detection Solutions.” In *Proceedings of the 22nd European Conference on Cyber Warfare and Security (ECCWS 2023)*, 285–293. Reading, UK: Academic Conferences International Limited. https://doi.org/10.34190/eccws.22.1.1071

Marwick, Alice E., and Rebecca Lewis. 2017. *Media Manipulation and Disinformation Online*. New York: Data & Society Research Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online/

Matthes, Jörg. 2012. “Framing Politics: An Integrative Approach.” *American Behavioral Scientist* 56 (3): 247–259. https://doi.org/10.1177/0002764211426324

McCombs, Maxwell E. 2014. *Setting the Agenda: The Mass Media and Public Opinion*. 2nd ed. Cambridge: Polity Press.

Nakov, Preslav, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Muhammad Mujahid, and Husrev Taha Sencar. 2024. "A Survey on Predicting the Factuality and the Bias of News Media." In *Findings of the Association for Computational Linguistics*: ACL 2024, 15947–15962. Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.944

NewsAPI. 2025. Accessed September 27, 2025. https://newsapi.org/

O'Keefe, Daniel J. 2002. *Persuasion: Theory and Research*. 2nd ed. Thousand Oaks, CA: Sage Publications.

Piskorski, Jakub, Preslav Nakov, Tamer Elsayed, Giovanni Da San Martino, Maram Hasanain, Georgi Karadzhov, Alexey Nedoluzhko, et al. 2024. "Overview of the CLEF-2024 CheckThat! Lab: Detecting Persuasion Techniques and Related Tasks." In *CLEF 2024 Working Notes*. CEUR Workshop Proceedings 3740: 356–381. http://ceur-ws.org/Vol-3740/paper-26.pdf

PolitiFact. 2025. Accessed September 27, 2025. https://www.politifact.com/

Sennet, Adam. 2023. "Ambiguity." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Summer 2023 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/ambiguity/

Snopes. 2025. Accessed September 27, 2025. https://www.snopes.com/

Su, Jinyan, Claire Cardie, and Preslav Nakov. 2023. "Adapting Fake News Detection to the Era of Large Language Models." arXiv preprint arXiv:2311.04917. https://arxiv.org/abs/2311.04917

Sundriyal, Megha, Tanmoy Chakraborty, and Preslav Nakov. 2023. "From Chaos to Clarity: Claim Normalization to Empower Fact-Checking." In *Findings of the Association for Computational Linguistics*: EMNLP 2023, 5891–5903. Singapore: Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.439/

Tsikerdekis, Michail. 2018. "Real-Time Identity-Deception Detection Techniques for Social Media: Optimizations and Challenges." *IEEE Internet Computing* 22 (5): 35–45. https://doi.org/10.1109/MIC.2017.265102442

Wahl-Jorgensen, Karin. 2019. *Emotions, Media and Politics*. Cambridge: Polity Press.

Waisbord, Silvio. 2018. "Truth Is What Happens to News: On Journalism, Fake News, and Post-Truth." *Journalism Studies* 19 (13): 1866–1878. https://doi.org/10.1080/1461670X.2018.1492881

Wasow, Thomas. 2002. *Postverbal Behavior*. Stanford, CA: CSLI Publications.

Zhang, Yangxiang, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. 2024. “FastForensics: Efficient Two-Stream Design for Real-Time Image Manipulation Detection.” arXiv preprint arXiv:2408.16582. https://arxiv.org/abs/2408.16582

Zhang, Yifan, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. “Tanbih: Get to Know What You Are Reading.” arXiv preprint arXiv:1910.02028. https://arxiv.org/abs/1910.02028

Zwicky, Arnold M., and Jerrold M. Sadock. 1975. “Ambiguity Tests and How to Fail Them.” In *Syntax and Semantics*, Vol. 4, edited by John P. Kimball, 1–36. New York: Academic Press. https://web.stanford.edu/~zwicky/ambiguity-tests-and-how-to-fail-them.pdf