Investigating the Hate–Credibility Nexus Across Datasets and Content Formats

The rapid spread of hate speech and low credibility content (LCC) online presents escalating challenges worldwide. While these phenomena have historically been studied separately, this work investigates their relationship across long-form text formats, revealing striking contradictions to existing research focusing on short-form text. While hate speech has been defined as "direct attacks on individuals based on protected characteristics, defining "fake news" remains more contested. Prior works by Ngueajio et al. (ACM Computing Surveys 2025) characterize fake news as fabricated information presented as true, shared to deceive or manipulate public opinion, often designed to trigger emotional reactions and potentially cause reputational harm. In this work, we adopt a pragmatic approach by treating fake news as LCC containing verifiably false or misleading claims, regardless of intent, thus aligning with recent related computational work (DeVerna et al., PLoS One 2024).

Research Question and Motivation: Prior research by Mosleh et al. (PNAS Nexus 2024) suggests that fake headlines contain more hate speech (beta = -0.19). However, their conclusion is based solely on short-form content (mean length = 60 words) from Politifact and Snopes websites, assessing content toxicity with Google Perspective API, which has been shown to exhibit biases (Nakka, ACM WebSci 2025) and using a non-contextual method **(**Davidson et al., AAAI ICWSM 2017**)** for hatefulness detection. Ultimately, a unified hatefulness score is calculated using Principal Component Analysis (PCA), analyzing hate-LCC relationships with linear regression (LR). **Theoretical Framework:** Drawing from framing theory (Entman, Journal of Communication, 1993), which suggests that content format shapes how information is presented. We hypothesize that (i) content format influences how hate speech is expressed (direct vs. reported framing), and (ii) the hate–credibility relationship may not generalize across datasets or formats.

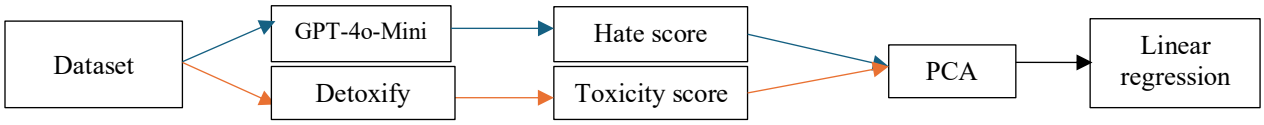

Figure 1: Methodological Pipeline

Method: We replicate Mosleh et al.'s PCA + LR pipeline (see Figure 1) on long-form content (mean length=3229 words) to test format generalizability, but introduce two core improvements: (i) **Hatefulness Scoring:** LLM-based, human-validated 5-point rubric adapted from UC Berkeley D-Lab guidelines.(ii) **Toxicity Scoring:** Detoxify (open-source, interpretable) instead of Perspective API. We apply this approach to both Mosleh's dataset, comprising 14617 short form texts, as well as a balanced (2500 real/fake each) random sample of the WeLFake dataset (Verma et al., IEEE TCSS 2021) sourced from diverse topics. (iii) The assessment of the framing hypothesis was done by analyzing direct vs contextual reports of hate using this prompt:

```
Prompt (temperature = 0 for consistency): You are an expert in online hate speech detection. You will be provided with a text '{text}', and your task is to identify the speech type
DIRECT: The text directly uses hateful language or expresses harmful sentiments
REPORTING: The text quotes, references, or discusses hateful language within an educational, journalistic, or analytical context.
```

Study	beta	R^2
Mosleh(Short)-original	-0.19	7%
Mosleh(Short)-our method	-0.47	4.4%
WelFake (Short)	+0.82	16.9%
WelFake (Long)	+0.84	17.5%

*$p < 0.001$. Negative β: more hate in LCC; Positive β: more hate in real news.

Table 1: Comparison of LR Results, Mosleh's work vs our method on Mosleh and WelFake Dataset

Key Findings: Our replication confirms Mosleh's negative association on short-form content, but with six-fold higher explanatory power. Surprisingly, Welfake shows **reversed relationships**: real news contains substantially more hate in both short and long form, contradicting Mosleh et al.'s findings as shown in **Table 1**. Specifically, from **Table 2,** we observe that (i) real texts account for 59% direct hate expression versus 41% contextual reporting, while (ii) LCC accounts for ed or contextual in our sample set. Cross-format validation using Mosleh's dataset (n=14,617) reveals that short-form content defaults to reported framing regardless of credibility, with the tendency for LCC to maintain plausible deniability by mainly reporting (e.g., "X said Y or Z reports that Y", without any hateful language itself). These findings support our hypothesis and highlight a key limitation in using hatefulness scores alone for moderation, since failing to account for how hate is framed may lead to misclassification. **Code/data available on GitHub upon acceptance of this work.**

	Short Form (Mosleh)		Long form (Ours)	
	Direct	Reporting	Direct	Reporting
All Texts	20.01%	**79.99%**	46.99%	**53.01%**
Real	19.42%	**80.58%**	**58.88%**	41.12%
Fake	20.23%	**79.77%**	1.92%	**98.08%**

Table 2: Direct vs Reported Hate patterns across datasets

Limitations and Next Steps: Our analysis uses a smaller sample (5,000 vs 14,617), which may limit direct comparison; yet the effect size difference (β = +0.84 vs -0.19) exceeds typical sampling variation. Future work will systematically examine potential confounders, such as topical composition, framing style, to disentangle dataset and format effects. We are also developing a theoretical model of the hate–credibility nexus. Some pathways have been identified from previous research (Ngueajio et al., ACM Computing Surveys 2025).

In conclusion, the hate speech-LCC relationship is format and dataset-dependent, not universal. We observe that (i) LCC does not always contain more hate, since (ii) LCC often primarily reports on hateful content through reported speech, an important distinction for effective moderation. This work highlights the significance of format-credibility interaction in hate speech detection, with implications that extend to the design of content moderation systems and computational social science methodologies.