LLM-Based Frame and Stance Annotation for 19th-Century Rumour Discourse in US and UK Newspapers: A Digital Humanities Resource

Wanshu Zhang

University of Luxembourg
2, Av. de l'Universite
wanshu.zhang@uni.lu

Abstract

Digital humanities scholars increasingly consult digitized historical newspapers to study how rumours travel, how institutions respond, and how everyday publics negotiate credibility. Yet this interpretive work is slowed by two bottlenecks: noisy OCR text that obscures variant spellings and layouts, and the absence of scalable semantic annotations for what a rumour is "about" and how a text positions itself toward it (assertion, denial, attribution, correction). I present a resource-building pipeline that links a previously validated retrieval workflow for rumour discourse with a new evaluation setup for large language models as semantic annotators. From public-domain US and UK newspapers (1804–1896), I derive sentence-level rumour instances and construct an 800-instance balanced benchmark for (i) topical framing (7-way), (ii) evidential stance (4-way), and (iii) an optional audit flag for temporal anachronism in model rationales. I also report a preliminary pilot with Gemini 2.5 Flash-Lite on a 200-instance singly annotated subset, showing that structured JSON output is operationally stable and that evidential stance appears more tractable than topical framing under the current prompt design. The outcome is a transparent benchmark design and annotation protocol that can be extended to other periods and languages.

Keywords: digital humanities; historical newspapers; rumour discourse; semantic frames; evidentiality; large language models; evaluation

1. Motivation and digital humanities relevance

Rumours are not only false information; they are a historically situated genre through which communities register uncertainty, manage risk, and negotiate authority. In nineteenth-century newspapers, rumours cluster around wars and diplomacy, financial panics, epidemics, crime, and moral anxieties—topics that also structure the archival record that historians and literary scholars read today. Large-scale collections make it possible to trace these dynamics across decades, but the core DH challenge is interpretability: scholars need evidence for what is being talked about and how the text positions itself toward the circulating claim.

The project also speaks directly to lexical change and semantic evolution, since rumor-marking and evidential formulas such as it is said, we learn, and a correspondent writes shift in frequency and function across the nineteenth century.

My project contributes by treating LLMs as assistive annotators for rumour discourse: I aim to speed up the first pass of corpus exploration while preserving the ability to audit and contest model decisions. This paper describes (i) a benchmark derived from public-domain US and UK newspapers (1804–1896), (ii) an annotation framework for topical framing and evidential stance, and (iii) a pilot evaluation setup for LLM-assisted analysis that remains legible to humanists.

I align the annotation task with practical DH questions. For example: When do newspapers attribute rumours to named institutions versus anonymous correspondents? Do denials and corrections concentrate around specific topics? Are gossip-like rumours framed differently across regions and decades? Answering such questions requires structured semantics, but it also requires transparency about uncertainty and historical context.

2. Related work across DH and NLP

Historical newspaper analysis faces persistent obstacles: OCR errors, layout artefacts, non-standard orthography, and domain shift across time and publication venues. These issues motivate robust retrieval and cleaning methods, as well as evaluation practices that report uncertainty rather than hiding it. DH work has emphasized collaborative curation and interpretive workflows that combine computational signals with expert reading, rather than one-shot automation.

In rumour studies, historians have treated false or unverified news as a window into social psychology and information ecologies, from

Proceedings of NEOLOGY AND LARGE LANGUAGE MODELS, pages 66–70

wartime “false news” to moral panics. In NLP, stance detection and framing analysis are well studied for modern sources, but label sets and model behaviors often assume present-day discourse conventions and do not account for historically specific evidential formulas such as we learn, it is rumored, or a correspondent writes.

Modern stance-detection work typically targets contemporary discourse and assumes relatively stable present-day evidential conventions. By contrast, historical NLP must contend with OCR noise, orthographic variation, and temporal domain shift, while nineteenth-century newspapers also rely on period-specific formulas of attribution, hedging, and correction. My benchmark adapts stance-style annotation to this setting by combining a retrieval workflow validated in a DH venue with prompts that explicitly discourage presentist reasoning and require textual grounding.

3. Data and derived units

Source corpora include public-domain newspaper datasets for the US and the UK. The material spans 1804–1896 and covers diverse genres, including local news, international dispatches, editorials, and advertisements embedded in text streams. I treat the rumour instance as the core unit: a sentence, or short sentence group, that asserts, attributes, denies, or corrects a circulating claim.

Using an established two-phase workflow (structural OCR cleaning and orthography-robust retrieval), I identify candidate rumour sentences in the two corpora. Dependency patterns are used only at the retrieval stage to recover proposition-like instances despite historical variation and OCR noise. After light filtering, I remove severely corrupted OCR spans, exact or near-duplicates, and non-informative fragments lacking a recoverable rumour proposition. The benchmark is built through retrieval, sampling, and human annotation; LLMs do not create gold labels.

From this candidate pool, I construct a balanced benchmark sample for annotation and subsequent LLM evaluation. The current sample contains 800 instances, evenly divided by region (400 US, 400 UK) and distributed across mid-century time bins so as to capture major shifts in press infrastructures, including telegraphy and news agencies, without over-representing any single period. Because annotation is still ongoing, the present paper does not yet define a final train/dev/test partition; instead, it focuses on benchmark construction, annotation protocol, and an initial pilot subset of 200 annotated instances. The current pilot uses Gemini 2.5 Flash-Lite only.

4. Annotation targets and interpretive rationale

Each rumour instance is annotated along two axes: topical framing and evidential stance. Topical framing captures the primary social domain in which the rumour is presented, using seven labels: WAR_DIPLOMACY, MARKETS_COMMERCE, CRIME_JUSTICE, HEALTH_EPIDEMIC, POLITICS_PUBLIC_LIFE, SCIENCE_TECHNOLOGY, and SOCIAL_CULTURAL_LIFE. Evidential stance captures how the text positions the circulating claim, using four labels: ASSERTED, ATTRIBUTED, HEDGED, and DENIED_CORRECTED.

Each instance receives one label per axis. Annotators assign the topical-framing label that best matches the primary domain foregrounded in the passage rather than all potentially relevant themes. For evidential stance, ASSERTED is used when the claim is presented with minimal hedging, ATTRIBUTED when it is linked to a named or inferable external source, HEDGED when uncertainty or rumor-marking language is foregrounded without clear attribution, and DENIED_CORRECTED when a circulating claim is explicitly rejected, corrected, or countered.

To improve consistency, the guidelines prioritize local textual evidence over inferred background context. Formulaic expressions such as it is said, we hear, or we learn are treated as HEDGED unless the wording clearly attributes the claim to a specific source, in which case ATTRIBUTED is preferred. Likewise, passages that repeat a circulating claim only in order to reject it are labeled DENIED_CORRECTED rather than ASSERTED.

4.1 Mini-examples for label legibility

The protocol includes short representative examples to keep the labels intelligible for humanities readers and annotators. For evidential stance, ATTRIBUTED is illustrated by formulations such as “A dispatch from Vienna states that ...”; HEDGED by “It is rumored that cholera has appeared ...”; DENIED_CORRECTED by “The report of a bank failure is unfounded ...”; and ASSERTED by “The prisoner confessed ...”. These examples are not substitutes for historical reading, but scaffolding that helps annotators recognize period-appropriate evidential signals and avoid importing contemporary assumptions.

5. Human annotation protocol and adjudication

Because DH resources must be trustworthy and reusable, this paper specifies a lightweight but explicit human annotation protocol. Annotators receive (i) the cleaned rumour span, (ii) minimal metadata (year, region), and (iii) label definitions with decision rules. They do not receive full article context by default, in order to keep the unit comparable with model inputs; however, optional context lookup is permitted when OCR fragmentation or severe ambiguity makes the span difficult to interpret.

The benchmark is designed for full double annotation by two human annotators with backgrounds in digital history, including the author and a second annotator. Each instance receives one topical-framing label and one evidential-stance label. At the time of submission, 200 instances have been annotated by the author as an initial pilot subset, while full double annotation of the 800-instance benchmark is ongoing. After independent annotation, disagreements will be reviewed in adjudication sessions, with brief notes recorded for recurrent borderline cases. These notes are intended to form part of the released resource documentation and to make interpretive decisions transparent for future users.

Once the double-annotation pass is complete, I will compare the two annotators' decisions, revise guideline wording where necessary, and produce an adjudicated gold set for subsequent model evaluation. Because annotation is still in progress, I do not yet report inter-annotator agreement; instead, the present paper reports a preliminary model pilot on the singly annotated subset.

6. Pipeline overview and LLM prompting

I organize the project as a reproducible pipeline from corpus to benchmark construction, human annotation, and model evaluation. Phase A produces cleaned rumour candidates and structured metadata. Phase B evaluates ChatGPT, Claude, and Gemini as semantic annotators under historically informed prompting. My aim is not to construct an exhaustive leaderboard, but to compare how widely used general-purpose LLMs handle topical framing and evidential stance in nineteenth-century rumour discourse under controlled prompt conditions.

I test three prompt variants: P1 (minimal label definitions), P2 (definitions plus short historical examples), and P3 (definitions plus a caution against presentist reasoning and a requirement to cite local textual evidence). For the present pilot, I generate a single structured response per instance and retain full model outputs, including predicted labels and rationales, in order to support auditing and later re-analysis.

Each model takes as input a cleaned sentence span with minimal metadata (year, region), and produces as output a topical frame, a stance label, an optional audit flag, and a short justification grounded in quoted words or phrases. In the current submission, I use this setup for a preliminary pilot with Gemini 2.5 Flash-Lite on a singly annotated 200-instance subset. The pilot reported here uses the P3 prompt variant. For the current pilot, Gemini 2.5 Flash-Lite produces one structured response per instance, and pilot results are reported as diagnostic accuracy against the available single-annotator reference labels. Full comparative evaluation across models will follow once the adjudicated benchmark is complete.

7. Evaluation plan and DH-oriented reporting

I outline the full evaluation framework for ChatGPT, Claude, and Gemini once the adjudicated benchmark is complete. As an initial pilot, however, I run Gemini 2.5 Flash-Lite on a singly annotated subset of 200 instances in order to test prompt usability, inspect output rationales, and identify recurrent error types before full-scale comparative evaluation.

Standard metrics such as macro-F1 and confusion matrices remain necessary, but are insufficient for DH use on their own. I therefore also plan to report per-region and per-period breakdowns and audit-flag rates in order to capture temporal variation and unsupported or presentist rationalization.

Gemini returned valid structured JSON for all 200 cases, indicating that the prompt and output schema are operationally stable for batch annotation. On this preliminary subset, the model achieved 0.66 accuracy for topical framing and 0.745 accuracy for evidential stance when compared against the available single-annotator reference labels. The pilot used the P3 prompt variant and is intended as a diagnostic rather than benchmark-final evaluation, helping me test prompt stability, output structure, and recurrent confusion patterns before full adjudicated comparison. Because the current pilot subset is singly annotated and label distributions are uneven, I treat these scores as provisional evidence of task feasibility rather than as final performance claims.

In the pilot, topical framing is strongest for WAR_DIPLOMACY and MARKETS_COMMERCE, and weaker for broader categories such as SOCIAL_CULTURAL_LIFE. For stance, performance is strongest on the dominant

HEDGED class. Even so, the pilot is useful for validating the prompt design and identifying confusion patterns for later adjudicated evaluation.

I complement aggregate scores with qualitative error typologies that matter for interpretation. Recurrent categories include (i) collapsing WAR_DIPLOMACY into POLITICS_PUBLIC_LIFE when dispatches mention ministers or cabinet changes; (ii) confusing ATTRIBUTED with HEDGED in formulaic phrases such as it is said or we learn; and (iii) over-triggering HEALTH_EPIDEMIC for metaphorical uses of terms such as plague.

7.1 A DH use case: mapping credibility work

As an illustrative DH use case, I propose a credibility work map that combines topical framing with evidential stance to identify where newspapers perform verification, denial, or distancing. For example, a spike in DENIED_CORRECTED within MARKETS_COMMERCE during a financial panic can be read alongside editorials about speculation and information flows. Likewise, persistent ATTRIBUTED stance within WAR_DIPLOMACY may highlight reliance on telegraphed dispatches and named correspondents.

8. Resource plan (LRE Map)

I plan to release the benchmark and protocol as a DH-facing language resource with clear documentation. The release package will separate (a) benchmark text spans and metadata, (b) human annotations and adjudication notes, and (c) model outputs produced under documented prompt conditions. It will also include normalization notes, dependency-pattern templates used in retrieval, label definitions, and examples for ambiguous cases, so that the resource remains reusable beyond a single model or corpus.

9. Conclusion

I have described a DH-grounded resource-building pipeline that connects robust rumour retrieval in historical newspapers with an annotation framework and an evaluation design for LLM-based topical framing and evidential stance analysis. Rather than presenting a completed benchmark, the paper introduces a concrete protocol for constructing one in a transparent and historically sensitive way. The preliminary Gemini pilot shows that structured model output is operationally feasible, while also highlighting the need for fuller annotation and adjudication.

Next steps include completing double annotation and adjudication for the balanced 800-instance benchmark, releasing prompt templates and evaluation scripts, extending the pilot to additional models, and testing the resource with DH researchers.

10. Limitations and ethical considerations

Historical newspapers contain sensitive material, including racialized language, moral panics, and stigmatizing descriptions that may be reproduced or amplified by automated tools. I therefore recommend that downstream users treat the benchmark as an index for scholarly inquiry rather than as a stand-alone truth source, and that they document interpretive choices when using model annotations in publications. When releasing model outputs, I will include guidance on handling harmful content and on avoiding over-interpretation of noisy OCR spans.

Methodologically, the scheme is intentionally compact in order to suit a short paper and a first resource release; it does not yet capture richer discourse structure such as multi-step attribution chains, editorial genre differences, or networked rumour propagation across titles.

11. Reproducibility and planned release

To make the resource reusable for DH audiences, I will publish a compact “how to cite and reuse” guide alongside the benchmark, including provenance fields, a recommended citation, and a changelog. I will also release prompt templates and evaluation code so that other researchers can rerun the annotation experiment with different models or local systems. Because reproducibility in historical corpora is complicated by OCR post-processing and collection updates, I store both the cleaned span used for annotation and a pointer back to the raw source text whenever stable identifiers are available; otherwise, I provide hashed text fingerprints for alignment across releases.

12. Illustrative micro-reading: how the labels support interpretation

To demonstrate how the scheme supports interpretation rather than purely technical classification, I outline a simple micro-reading workflow. A researcher interested in epidemic rumours, for example, can filter the corpus for HEALTH_EPIDEMIC and compare stance patterns across periods and regions: are rumours mainly HEDGED, ATTRIBUTED, or DENIED_CORRECTED? A similar workflow applies to MARKETS_COMMERCE during financial panics, where correction labels may point to editorial interventions aimed at calming

markets or managing reputational risk. By linking each label to quoted evidence in the local span, the resource is designed to support exploratory mapping while keeping close reading central.

13. References

Allen, B., Sieczkiewicz, R., & Radev, D. (2020). What's hard about historical newspaper analysis? Technical Report, University of Michigan.

Bloch, M. (1921). Réflexions d'un historien sur les fausses nouvelles de la guerre. Revue de Synthèse Historique, 33, 13–35.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 964–971.

Mueller, M. (2014). Shakespeare His Contemporaries: Collaborative curation and exploration of early modern drama in a digital environment. Digital Humanities Quarterly, 8(3).

PleIAs. (n.d.). US-PD-Newspapers (Hugging Face dataset). https://huggingface.co/datasets/PleIAs/US-PD-Newspapers

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP 2019, 3982–3992.

Smith, D. A., & Cordell, R. (2018). A research agenda for historical and multilingual optical character recognition. NULab Working Paper.

biglam. (n.d.). hmd_newspapers (Hugging Face dataset). https://huggingface.co/datasets/biglam/hmd_newspapers