Responsible NLP Checklist

Paper title: *Actors, Frames and Arguments: A Multi-Decade Computational Analysis of Climate Discourse in Financial News using Large Language Models*
Authors: *Ruiran Su, Markus Leippold, Janet B. Pierrehumbert*

How to read the checklist symbols:

☑ the authors responded ‘yes’

☒ the authors responded ‘no’

N/A the authors indicated that the question does not apply to their work

□ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ A. Questions mandatory for all submissions.

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☑ A2. Did you discuss any potential risks of your work?
In the Ethics Statement (following Section 6), the paper explicitly acknowledges that the methods presented "could also be used to craft more persuasive disinformation". Additionally, the authors discuss the environmental costs of large model inference and the risks associated with LLMs reproducing or amplifying biases.

☑ B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

☑ B1. Did you cite the creators of artifacts you used?
We cite the creators of the used artifacts in Section 3.1 (Dow Jones Intelligent Identifiers) , Section 3.2.2 (SBERT) , Section 4.1 (Gemini, LLaMA) , and Section 4.3 (GPT-4o, Claude, Qwen, Mixtral).

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
In Section 3.1, the authors describe the Dow Jones Intelligent Identifiers as a "proprietary subject taxonomy" and acknowledge Dow Jones for providing access in the Acknowledgments section. Furthermore, the References section explicitly classifies the models used as either "proprietary model" (e.g., Claude Sonnet 4) or "Open-weight" (e.g., Mixtral 8x22B).

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In the Ethics Statement (following Section 6), the authors specify the intended use of their work ("open academic inquiry," "monitor 'greenwashing'") and explicitly contrast this with incompatible uses such as crafting "persuasive disinformation". Additionally, the Acknowledgments section confirms that the use of the news corpus was facilitated by access provided by Dow Jones, implying use consistent with their terms.

☑ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice.

taken to protect/anonymize it?
In the Ethics Statement (following Section 6), the authors explicitly state: "Our analysis does not involve private or personally identifiable information".

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.1 and Appendix A detail the domain coverage (Core Climate Issues, Energy Transition, Climate-Affected Sectors) and the proprietary taxonomy used (DJID codes). Additionally, the Limitations section (following Section 6) explicitly documents that the study is based on a single, English-language news source.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Section 3.1 reports the total corpus size (980,061 articles) , and Table 1 breaks down the distribution of the 4,143 sampled articles across temporal strata. Furthermore, Section 5.1.1 explicitly details the data splits for the gold standard experiments: "1,400 train / 300 validation / 300 test".

☑ **C. Did you run computational experiments?**

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Section 4.1, the model sizes are reported (e.g., LLaMA-4 Maverick-17B, Mixtral-8x22B). The Acknowledgments section explicitly mentions the use of the "University of Oxford's Advanced Research Computing (ARC) facility" for computing infrastructure. Additionally, the Ethics Statement discusses the environmental costs of the large model training and inference.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1.1 details the experimental setup for the pipeline validation (comparison with RoBERTa baselines) . Appendix F.6 provides the inference settings (Temperature=0.2, Top-p=0.9). Appendix C.1 (Table 9) and Appendix D.1 (Table 10) explicitly list the hyperparameters for the argument detector and extraction models, including learning rate, batch size, and epochs.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.1.1 and Table 3 report macro F1 scores for model performance. Section 3.2.5 mentions using Jensen-Shannon divergence and cosine similarity to measure distribution fidelity. Section 4.2 reports inter-annotator agreement using Krippendorff's alpha and F1 scores. Appendix I.3 reports scores averaged over the 2,000-article gold standard.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Section 3.2.2 explicitly names the embedding model ("SBERT all-MiniLM-L6-v2") and the clustering algorithm ("agglomerative clustering with Ward linkage"). Appendix A.4 details the deduplication parameters (MinHash LSH with Jaccard similarity > 0.9) . Appendix F.6 lists inference parameters for the models (Temperature=0.2, Top-p=0.9).

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix G and Appendix H provide detailed excerpts of the annotation guidelines, including

specific rules for Actor-Stance , Frame Classification , and Argument Extraction . Appendix G.1 also describes the "interactive tutorial" provided to contributors.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.2 and Appendix G.1 specify that recruitment was done via the "Zooniverse citizen-science platform". The participants are explicitly described as "volunteers", implying no payment was involved, which is standard for Zooniverse projects.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Ethics Statement notes that the analysis "does not involve private or personally identifiable information".

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study relies on publicly available news data and volunteer contributions via an established citizen science platform.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix G.1 characterizes the annotators generally as "volunteers" who "did not have formal training in linguistics or climate communication". However, specific demographic (age, gender) or geographic statistics of the volunteer pool are not reported.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)