On the Importance of Nuanced Taxonomies for LLM-Based Understanding of Harmful Events: A Case Study on Antisemitism

Anonymous ACL submission

Abstract

Large language models (LLMs) can help elucidate hate, violence, and other toxicity. However, labeling harmful events is challenging due to the subjectivity of labels such as "toxicity" and "hate." Motivated by the rise of antisemitism, this paper studies the capability of LLMs to discover reports of antisemitic events. We pilot the task of hateful event classification on the AMCHA Corpus—a continuously updated dataset with expert-labeled instances of fine-grained types of antisemitism—and show that incorporating domain knowledge from fine-grained taxonomies is needed to make LLMs more effective. Our experiments find that providing precise definitions from a taxonomy can steer GPT-4 and Llama-3 to somewhat improve on tagging antisemitic event descriptions, with GPT-4 achieving up to a 14% increase in mean weighted F1. However, LLMs are still far from perfect at understanding antisemitic events, suggesting avenues for future work on LLM alignment and precise definition of antisemitism.

1 Introduction

Understanding hateful or harmful events from news reports can reveal broad societal trends (Pontiki et al., 2020) and harms toward marginalized communities.[1] However, harm is a subjective concept that annotators operationalize differently (Breitfeller et al., 2019; Sap et al., 2022; Alkomah and Ma, 2022; Kansok-Dusche et al., 2023; Yin and Zubiaga, 2021; Fleisig et al., 2023). LLMs may thus operationalize an "average" perspective when in reality one of two annotators sees a harmful stereotype, erasing valuable disagreement (Pavlovic and Poesio, 2024; Richardson, 2021).

This work investigates approaches to address these challenges by adding fine-grained prior knowledge to LLM prompts. We stress-test LLMs' ability to perform nuanced classification for descriptions of *antisemitic events*. The case of antisemitism is fit for this investigation because of its frequently debated definitions (Klug, 2023; Harrison and Klaff, 2021; Feldman and Volovici, 2023; Herf, 2021; Penslar, 2022; Nexus, 2023; Jerusalem, 2021). Despite its controversial nature,[2] studying antisemitism is important due to increased hate crimes against Jewish people[3] as well as the general harmful consequences that online hate can have both online and offline (e.g. harassment, mental distress, hate crimes, Räsänen et al., 2016; UN, 2018; Byman, 2021).

To study this task, we scrape and release the AMCHA Corpus, a growing challenge set of 6,748 English-language contextualized descriptions of antisemitic events that occurred on higher education campuses, annotated for coarse- and fine-grained categories of antisemitism. The typology and dataset were created by the AMCHA Initiative through continuous monitoring, screening, and consensus-coding of events according to their coarse-grained categories and fine-grained types of antisemitism.[4]

Our work asks the following questions:

1. How well do LLMs label the coarse-grained categories and fine-grained types of antisemitism included in the AMCHA Corpus?
2. To what extent can we steer LLMs to use various definitions of antisemitism?
3. Within texts labeled as antisemitic, which types and categories of antisemitic events are harder for LLMs to predict?

[1] https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm

[2] While antisemitism is controversial due to events in the Middle East, we take a descriptive stance to accommodate disagreement on definitions of antisemitism.

[3] https://www.fbi.gov/news/press-releases/fbi-releases-2022-crime-in-the-nation-statistics

[4] https://amchainitiative.org/categories-antisemitic-activity. Note that we recognize that differing taxonomies exist (JDA, 2021), but we employ this taxonomy due to the corpus' uniquely rich content, labels, and metadata for event classification.

4. How much can in-context learning improve LLMs' antisemitic event classification?

2 Methods

We experiment on gpt-4-1106-preview[5] and llama3-8b-instruct.[6] Our classification task is set up as follows: Given an input event description along with the date and university of the event (collectively, input d), model M must classify the coarse-grained categorical label c of the event, the set of n fine-grained type labels $t = t_1, \ldots, t_n$, as well as, optionally, the binary antisemitism label l. In some experiments, we provide additional inputs such as definitions (Def), in-context examples (IcE), and the antisemitism label l (As).

The M-noCtx setup asks the model M for labels l, then c, then t. M-As modifies M-noCtx by only prompting the model for c and t. M-As-IcE has the same task presentation as M-As but prepends one randomly selected entry corresponding to each potential value of t from the corpus. M-Def and M-As-Def use the same task setup as M-noCtx and M-As, respectively, but we also supply the definitions of each candidate for c and t, as well as Wikipedia's definition of antisemitism.

For M-noCtx and M-Def, we first compute the binary detection rate of antisemitic events, defined as the percentage of entries where the model predicts that the text describes an antisemitic event. Then, since not all types and categories have equal frequency in the dataset, we compute a **WF1** metric, representing an F1 score weighted by category or type frequencies within the corpus' gold labels.

3 Results

Overall, we find that LLM categorization of our events as antisemitic is quite poor (<0.4 for both LLMs in the zero-shot M-noCtx setup), and certain inputs improve performance more than others. From binary and coarse-grained perspectives, GPT-4 is less aligned with gold labels than Llama-3, but GPT-4 is more aligned on fine-grained labels (28.16% mean WF1 across types for GPT-4 vs. 25.52% for Llama-3). Notably, two particularly poorly aligned types are those that are (a) contentious (e.g. describing **BDS** as antisemitic) or (b) reliant on historical knowledge (e.g. *Historical antisemitism* with swastikas painted on buildings) are least aligned with AMCHA's labels. Providing category definitions (M-Def) improves detection of antisemitic **BDS** events, suggesting the need for clarity and definitions for what falls under nuanced concepts such as antisemitism.

The fine-grained *Denigration* and *Destruction of Jewish property* types have lower precision than recall, indicating that models mistake incidents involving *Historical antisemitism* tropes as *Destruction of Jewish property* and mistake incidents targeting institutions or organizations as incidents targeting and denigrating individuals, suggesting the need for infusion of historical knowledge that would help differentiate them. M-Def corrects several cases of *Historical antisemitism* and *Genocidal expression* that GPT-4 initially mistakes for *Denigration* or *Destruction of Jewish property*, indicating that adding definitions helps models operationalize historical knowledge. However, comparing M-As-IcE to M-As, we see that Llama-3's fine-grained type alignment is significantly worse across the board, showing that few-shot learning hurts fine-grained classification alignment.

4 Conclusion and Discussion

In this work, we extracted and released the AMCHA Corpus and studied LLMs' abilities to detect fine-grained harmful event types in the context of antisemitism. Our findings show that while Llama has higher binary detection rates and can be steered to improve coarse-grained category alignment, GPT-4 appears to be more steerable toward aligned fine-grained classification. We also observe that definitions tend to improve WF1 scores more than in-context examples. Our findings suggest that LLMs show promise for understanding harmful events at scale, decreasing the human burden of exposure to distressing news, and better grasping real-world manifestations of harm toward marginalized communities. However, our findings showcase that models struggle with detection and fine-grained categorization of nuanced concepts. Future work should explore how to better set up detection of categories that are contentious or that rely on deep historical or group-specific knowledge. Future work can also generalize our study to other forms of hate with multiple stakeholders who have differing perspectives, possibly through creating annotator-specific taxonomies with definitions that can steer LLMs to actively represent different annotators' stances as in Deng et al. (2023).

[5] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4
[6] https://llama.meta.com/llama3/

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Daniel L Byman. 2021. How hateful rhetoric connects to real-world violence. https://www.brookings.edu/articles/how-hateful-rhetoric-connects-to-real-world-violence/. Accessed: 2024-4-29.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

David Feldman and Marc Volovici. 2023. *Antisemitism, Islamophobia and the Politics of Definition*. Palgrave Critical Studies of Antisemitism and Racism. Springer International Publishing.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Bernard Harrison and Lesley Klaff. 2021. The IHRA definition and its critics. In Alvin H Rosenfeld, editor, *Contending with Antisemitism in a Rapidly Changing Political Climate*, pages 9–43. Indiana University Press.

Jeffrey Herf. 2021. IHRA and JDA: Examining definitions of antisemitism in 2021. *Fathom*.

JDA. 2021. The jerusalem declaration on antisemitism.

Jerusalem. 2021. The jerusalem declaration on antisemitism. https://jerusalemdeclaration.org/. Accessed: 2024-5-3.

Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, Anke Zeißig, Lisanne Seemann-Herz, Sebastian Wachs, and Ludwig Bilz. 2023. A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, violence, & abuse*, 24(4):2598–2615.

Brian Klug. 2023. Defining antisemitism: What is the point? In David Feldman and Marc Volovici, editors, *Antisemitism, Islamophobia and the Politics of Definition*, pages 191–209. Springer International Publishing, Cham.

Nexus. 2023. The nexus project - israel and antisemitism. https://nexusproject.us/. Accessed: 2024-5-3.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Derek Penslar. 2022. Who's afraid of defining antisemitism? *Antisemitism Studies*, 6(1):133–145.

Maria Pontiki, Maria Gavriilidou, Dimitris Gkoumas, and Stelios Piperidis. 2020. Verbal aggression as an indicator of xenophobic attitudes in Greek Twitter during and after the financial crisis. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 19–26, Marseille, France. European Language Resources Association.

Pekka Räsänen, James Hawdon, Emma Holkeri, Teo Keipi, Matti Näsi, and Atte Oksanen. 2016. Targets of online hate: Examining determinants of victimization among young finnish facebook users. *Violence and victims*, 31(4):708–725.

Sharon Richardson. 2021. Against generalisation: Data-driven decisions need context to be human-compatible. *Business Information Review*, 38(4):162–169.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

UN. 2018. Hate speech and real harm. https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.