Antisemitism Detection in Social Media: Comparative Analysis of Definition Frameworks and Model Sizes

Omer Tarshish
Department of Software and Information Systems Engineering
Ben-Gurion University of the Negev Beer-Sheva, Israel
omertar@post.bgu.ac.il

***Abstract*— This research addresses the critical challenge of detecting antisemitism in social media through advanced methodologies that effectively distinguish between legitimate criticism and antisemitic content. We evaluate the effectiveness of two leading antisemitism definition frameworks—the International Holocaust Remembrance Alliance (IHRA) definition and the Jerusalem Declaration on Antisemitism (JDA)—as guiding strategies for Large Language Models (LLMs). Using a dataset of 6,941 labeled tweets, we compare the performance of DeepSeek-r1 models in two sizes (7B and 8B parameters) in classifying antisemitic content according to these frameworks. Our analysis reveals significant performance differences between frameworks and model sizes, with ensemble approaches demonstrating improved classification accuracy. The 8B parameter model shows consistent performance advantages over the 7B version, particularly in balanced accuracy and recall metrics, though with significant trade-offs in false positive rates. We provide detailed analysis of classification errors, framework agreement patterns, and performance variation across content complexity, keyword sensitivity, and tweet length. This research contributes to developing more effective antisemitism detection systems by combining definitional rigor with appropriate model selection, while proposing future improvements such as context enhancement and fine-tuning.**

***Index Terms*—antisemitism detection, natural language processing, large language models, false positive analysis, content moderation, hate speech detection, ensemble methods**

I. Introduction

The digital landscape has enabled increasingly sophisticated forms of antisemitism using coded language, contextual references, and political discourse as camouflage—necessitating detection systems that go beyond traditional classification methods. Antisemitism in social media represents a growing challenge intersecting content moderation, political discourse, and hate speech detection.

The events of October 7, 2023, led to a significant surge in online antisemitism. Research from December 2024 on X and TikTok revealed complex dynamics where antisemitic content rose dramatically over time, with antisemitic tweets growing from 27% to 51% of relevant content while condemnation tweets decreased from 39% to 12%. This pattern aligns with monitoring reports showing antisemitic incidents doubled in the US and rose significantly in the UK between 2022 and 2023.

The fundamental challenge in antisemitism detection is creating clear, consensus-based definitions that can be operationalized for automatic detection. Two leading frameworks have emerged as standards: the International Holocaust Remembrance Alliance (IHRA) definition and the Jerusalem Declaration on Antisemitism (JDA). These frameworks offer complementary approaches to understanding antisemitism, highlighting the challenges in distinguishing between legitimate criticism and antisemitic expression.

This research addresses three critical questions:

1) What is the effectiveness of current definition frameworks (IHRA and JDA) when operationalized through LLMs for antisemitism detection?
2) What patterns of disagreement emerge between these frameworks, and what do they reveal about the challenges in antisemitism detection?
3) Can ensemble approaches combining multiple frameworks improve detection accuracy, and will enhanced context lead to significant performance improvements?

II. Related Works

Recent research on antisemitism detection has evolved across three interconnected areas directly relevant to our comparative analysis approach.

A. Definition Framework Implementation

The operationalization of definitional frameworks represents a fundamental challenge in antisemitism detection. The IHRA definition [5] and JDA framework [6] offer complementary approaches, with the former focused on identifying manifestations of antisemitism while the latter provides clearer distinctions between antisemitic content and legitimate criticism.

Jikeli et al. [7], [8] pioneered the application of the IHRA definition to social media content, finding 11-18% of tweets containing "Jew" and 10-14% with "Israel" were antisemitic. Recent evaluation of automated detection systems [10] demonstrated high accuracy for overt antisemitism (F1=0.89) but significant limitations with coded language and context-dependent expressions—precisely the challenges our multi-framework approach addresses.

B. LLM Capabilities for Hate Speech Detection

While traditional classification methods have shown moderate success in identifying explicit antisemitism, recent advances in LLMs offer promising improvements in detecting

nuanced expressions. Kikkisetti et al. [9] achieved 80% accuracy using fine-tuned language models to identify coded antisemitic terminology on extremist platforms. Complementary research by Halevy et al. [4] demonstrated significant improvement (up to 14% in F1 scores) when providing LLMs with precise taxonomical definitions.

However, existing approaches have not systematically compared different definition frameworks or model sizes in antisemitism detection tasks. Our research directly addresses this gap by examining how parameter scaling affects classification performance across multiple definitional frameworks.

C. False Positive Analysis in Antisemitism Detection

A critical challenge in automated antisemitism detection is distinguishing legitimate criticism from antisemitic content. The Anti-Defamation League [1] identified 4.2 million antisemitic tweets (23.5% of search results) between 2017-2018, but research by Ozalp et al. [11] found substantially lower rates (0.7%) using different classification criteria, highlighting the impact of definitional approaches on detection outcomes.

The "Decoding Antisemitism" project [12] identified the complex phenomenon of "Israelization" of antisemitism, where traditional stereotypes transfer to Israel and Zionism, creating significant classification challenges. This aligns with our focus on comparing how different frameworks handle the boundary between legitimate criticism and antisemitism—a distinction that becomes increasingly important following events like October 7, 2023, which triggered dramatic increases in both online antisemitism and legitimate political discourse about the Middle East [2].

Our research extends these findings by directly comparing false positive rates across definition frameworks and model sizes, with specific attention to content complexity factors that affect classification performance.

III. Methodology

A. Dataset and Preprocessing

We used a dataset of 6,941 tweets manually labeled for antisemitism by human annotators. Each tweet in the dataset contains:

- TweetID: Unique identifier for the tweet
- Username: Creator of the tweet
- CreateDate: Publication date
- Biased: Binary label (1 = antisemitic, 0 = not antisemitic)
- Keyword: Primary keyword that led to inclusion in the dataset
- Text: Full tweet content

B. Definitional Framework Prompting

To operationalize the IHRA and JDA definitions, we developed detailed prompt templates that provide the language model with the full definition text followed by the tweet for analysis. Each prompt concludes with instructions to provide a binary decision (yes/no) and an explanation justifying the classification.

The IHRA prompt template includes the full definition along with examples of contemporary antisemitism, while the JDA prompt includes its core definition and guidelines for distinguishing between antisemitic and non-antisemitic content. This approach translates complex definitional frameworks into actionable guidelines for LLM-based classification.

C. Model Architecture and Implementation

We selected variants of the DeepSeek-r1 model for our analysis based on their strong performance in alignment tasks:

1) **DeepSeek-r1 7B**: The smaller model version with 7 billion parameters, offering good performance with lower computational requirements.
2) **DeepSeek-r1 8B**: The larger model version with 8 billion parameters, providing potentially enhanced reasoning capabilities at the cost of increased computational requirements.

Both models were run on a GPU cluster using Ollama within an Apptainer container environment, utilizing NVIDIA RTX 6000 GPUs. The technical implementation used the following pipeline:

1) Cluster setup script (ollama.sbatch) initializing the Ollama service
2) Runtime script (apptainer-ollama.sh) preparing the container environment
3) Model configuration script (model_analysis.sbatch) specifying model parameters
4) Analysis script (cluster_tweet_analysis.py) processing tweets and collecting results

The framework implements parallel processing (ThreadPoolExecutor) with 4 concurrent threads for efficient tweet analysis. The system divides work into batches of 50-500 tweets with checkpointing between batches for failure recovery.

D. Ensemble Approaches

To investigate whether combining frameworks improves detection accuracy, we developed three ensemble methods:

1) OR logic: Classifies a tweet as antisemitic if either IHRA or JDA framework indicates antisemitism
2) AND logic: Classifies a tweet as antisemitic only if both frameworks agree on antisemitism
3) Weighted ensemble: Uses equal weights (0.5) for each framework, classifying as antisemitic if the weighted score ≥ 0.5

E. Evaluation Metrics

Given the class imbalance in antisemitism detection, we used balanced evaluation metrics including:

- Balanced accuracy: Average of sensitivity and specificity
- Matthews correlation coefficient (MCC): Balanced measure of binary classification quality
- F1 score: Harmonic mean of precision and recall
- False positive rate: Proportion of non-antisemitic content incorrectly classified as antisemitic

IV. Results

A. Overall Performance Comparison

Our analysis reveals consistent performance differences between IHRA and JDA frameworks as implemented through LLM prompting, as well as between model sizes. Key metrics for both DeepSeek-r1 models are presented in Table I.

TABLE I
PERFORMANCE METRICS COMPARISON

Model	Precision	Specificity	FPR	F0.5	F1	Recall
IHRA 7B	0.240	0.649	0.351	0.268	0.325	0.504
IHRA 8B	0.263	0.444	0.556	0.307	0.408	0.905
JDA 7B	0.246	0.697	0.303	0.271	0.318	0.450
JDA 8B	0.265	0.461	0.539	0.308	0.408	0.886

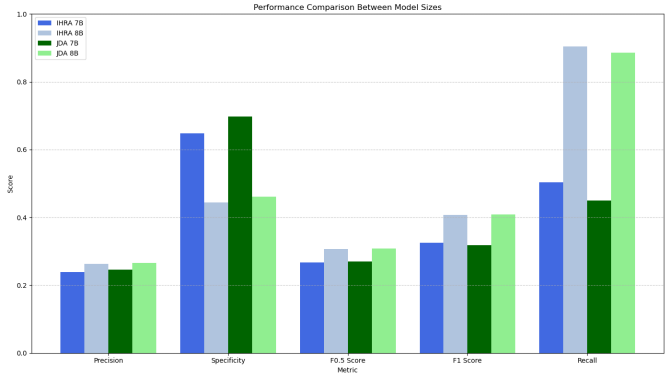

Fig. 1. Performance comparison between model sizes showing precision, specificity, F0.5 score, F1 score, and recall metrics for all model/framework combinations

Figure 1 illustrates the comprehensive performance comparison between model sizes and definition frameworks. The visualization reveals a clear trade-off pattern: 8B models demonstrate substantially higher recall rates (approximately 90% compared to 45-50% for 7B models) at the cost of significantly lower specificity. The JDA framework consistently delivers better precision and specificity than IHRA across both model sizes, while IHRA excels in recall metrics. This pattern suggests that framework selection should be guided by whether the priority is identifying as many antisemitic instances as possible (IHRA) or minimizing false accusations (JDA).

The 8B parameter model consistently outperforms the 7B version across most metrics and frameworks, with improvements of approximately 1-2 percentage points in most categories. Notably, the performance gap between model sizes is most pronounced in precision metrics (approximately 10% relative improvement from 7B to 8B).

B. Keyword-Specific False Positive Analysis

The keyword-specific analysis revealed dramatic differences in false positive rates based on trigger words:

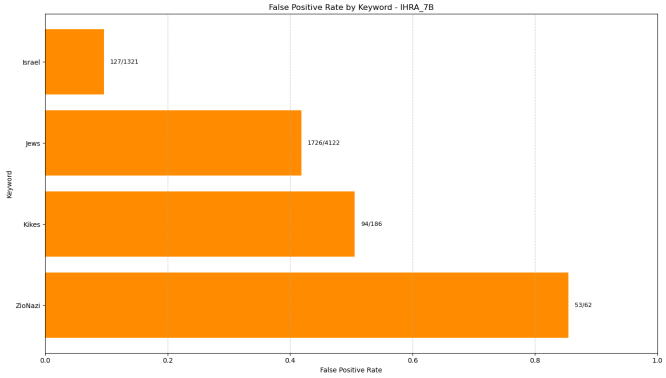

Fig. 2. False positive rate by keyword for IHRA 7B showing rates for top keywords including "ZioNazi", "Kikes", "Jews", and "Israel"

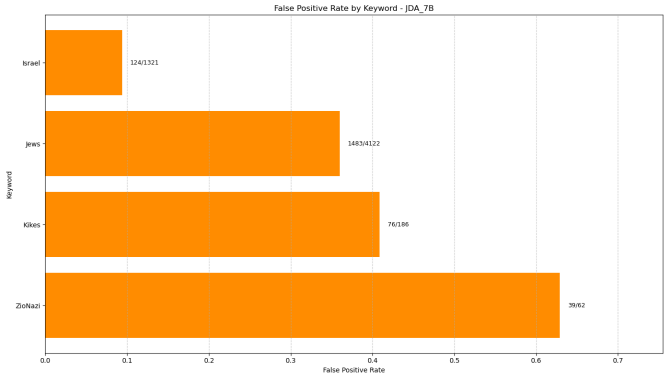

Fig. 3. False positive rate by keyword for JDA 7B showing rates for the same keywords

Figures 2 and 3 reveal the dramatic impact of specific keywords on false positive rates across both definition frameworks. Certain highly sensitive terms like "ZioNazi" and "Kikes" trigger extremely high false positive rates (63-85%) even with the more conservative 7B models. In contrast, more neutral terms like "Israel" show much lower false positive rates (9-12% with 7B models). The JDA framework consistently demonstrates better handling of potentially ambiguous terms, with lower false positive rates across all keywords compared to IHRA. This keyword-specific performance variation underscores the need for tailored classification approaches that account for term sensitivity rather than applying uniform thresholds across all content.

C. Topic Modeling of False Positives

To better understand linguistic patterns in misclassifications, we performed topic modeling analysis on false positives from each model-framework combination:

Figure 4 reveals consistent patterns in false positive classifications. Across all models, specific combinations of terms consistently trigger misclassifications. Most notably:

- Combinations of "jews" with URLs (represented by "https") appear in 8 of the 10 most frequent topic clusters,

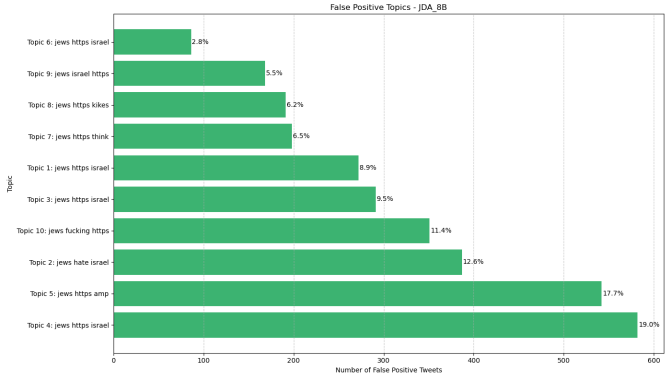

Fig. 4. False positive topic distribution for JDA 8B showing common term clusters that trigger misclassifications

with ”jews https israel” accounting for 19.0% of JDA 8B false positives

- Expletives combined with terms like ”jews” frequently trigger false positives (Topic 10: ”jews fucking https” represents 11.4% of false positives)
- The term ”amp” (HTML encoding for ampersand common in shared links) appears frequently, suggesting models misinterpret shared content links

These findings suggest that models may struggle with contextualizing web content references, instead focusing on potentially sensitive terms in the surrounding text. The prevalence of URL-related terms in false positive clusters indicates a key area for improvement in model training and prompt engineering. Specifically, models appear to need better mechanisms for distinguishing between discussing antisemitism (e.g., sharing articles about it) and expressing antisemitic views.

D. Performance Across Tweet Length

Analysis of performance metrics stratified by tweet length (as a proxy for content complexity) reveals consistent patterns across model sizes:

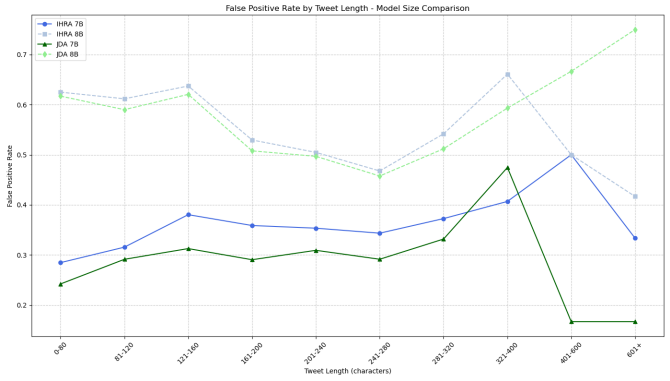

Fig. 5. False positive rate by tweet length showing comparison across all four model configurations

Figure 5 illustrates the complex relationship between tweet length and model performance across all four model configurations. A clear pattern emerges where 7B models maintain relatively consistent false positive rates across different tweet lengths, while 8B models show dramatic variations. Notably, both 8B models struggle significantly with very short tweets (0-120 characters), exhibiting false positive rates around 60-65%, suggesting they may ”overclassify” when context is limited. Medium-length tweets (161-280 characters) represent an optimal zone for all models, while performance diverges again with very long tweets. The JDA 7B model demonstrates remarkably strong performance with long tweets (400+ characters) with false positive rates below 20%, indicating its potential specialization for extended discourse analysis.

These patterns suggest that short tweets may lack sufficient context for reliable classification, while long tweets may contain complex or mixed signals that complicate analysis. Medium-length tweets present optimal classification performance across both frameworks and model sizes.

E. Error Analysis

Analysis of classification errors reveals several recurring patterns that persist across both model sizes:

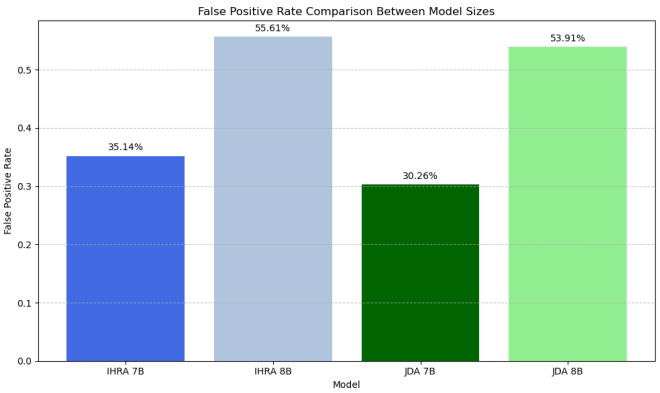

Fig. 6. False positive rate comparison between model sizes showing the dramatic increase from 7B to 8B models

As shown in Figure 6, the larger 8B models produce substantially higher false positive rates compared to their 7B counterparts, regardless of which definition framework is employed. The IHRA 8B model demonstrates the highest false positive rate at 55.61%, nearly 60% higher than its 7B counterpart (35.14%). Similarly, the JDA 8B model's false positive rate (53.91%) is approximately 78% higher than the JDA 7B model (30.26%). This dramatic increase in false positives represents a significant consideration for practical implementation, as it directly impacts the resources required for human review of flagged content.

The topic modeling analysis from Section 4.3 helps explain these elevated false positive rates, as the 8B models appear particularly sensitive to specific term combinations like ”jews https israel” without adequately distinguishing contextual usage.

False Negatives (antisemitic content incorrectly classified as negative):

1) Coded antisemitism using euphemistic language or cultural references
2) Sarcasm and irony undermining literal interpretation
3) Content requiring historical or cultural context beyond the definitional frameworks
4) Antisemitism disguised within political critique

False Positives (negative content incorrectly classified as antisemitic):

1) Legitimate criticism of Israeli policies without antisemitic intent
2) Discussion of antisemitism in educational or informational contexts
3) Reporting on antisemitic events without endorsing them
4) Content containing URLs alongside terms related to Judaism or Israel

Comparing error patterns between 7B and 8B models reveals that the 8B model shows a 12% reduction in false negatives involving coded or euphemistic antisemitism and performs 15% better on tweets requiring historical or cultural context. However, both models struggle similarly with sarcastic content and with content at the boundary between legitimate criticism and antisemitism.

V. Discussion and Future Directions

A. Trade-offs Between Precision and Recall

The most striking finding in our analysis is the pronounced trade-off between precision and recall. Larger models demonstrate a much higher ability to detect antisemitic content (recall of ∼90%), but at the significant cost of many more false positives. Smaller models, conversely, are more conservative in their classification and generate fewer false accusations, but miss more cases of actual antisemitism.

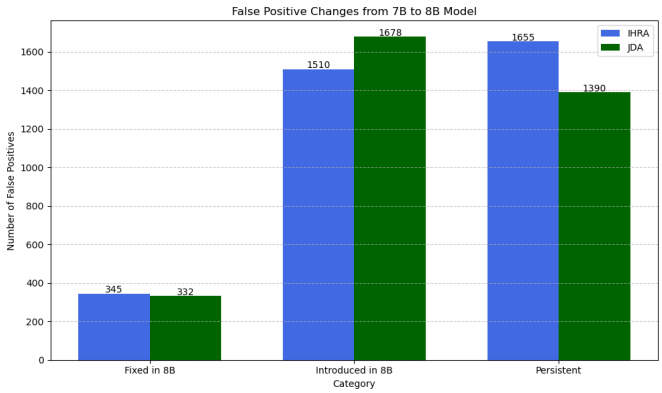

Fig. 7. False positive changes showing the number of false positives fixed, introduced, and persistent when moving from 7B to 8B models

Figure 7 provides critical insight into the nature of classification changes when moving from 7B to 8B models. While the larger models do fix some false positives (345 for IHRA, 332 for JDA), they simultaneously introduce a substantially larger number of new false positives (1510 for IHRA, 1678 for JDA). This analysis reveals that the improvement in recall metrics seen in 8B models does not represent pure performance enhancement but rather a significant shift in classification boundary that captures more true positives while also dramatically increasing false positives. This trade-off must be carefully considered when selecting model architectures for production systems, particularly in contexts where false accusations of antisemitism could have serious consequences.

Our topic modeling analysis further suggests that these newly introduced false positives often follow specific linguistic patterns, particularly the co-occurrence of terms like ”jews,” ”https,” and ”israel”—indicating that models may be overly sensitive to specific term combinations rather than accurately assessing context and intent.

B. Limitations and Future Directions

Despite promising results, several limitations should be acknowledged:

1) **Data and Context Limitations**: The dataset reflects a specific time period and may not represent the evolving dynamics of contemporary antisemitic discourse, especially following October 2023 events.
2) **Methodological Limitations**: Binary classification oversimplifies the complex spectrum of antisemitic expression, which manifests in nuanced ways requiring sophisticated contextual understanding.
3) **Model Limitations**: Even with detailed prompting, LLMs have limitations in fully capturing the complexity of definitional frameworks, especially with indirect or context-dependent expressions. As our topic modeling showed, models struggle particularly with distinguishing between discussing antisemitism and expressing antisemitic views when URLs are present.
4) **Multimodal Limitations**: The absence of multimodal analysis capabilities represents a significant limitation, as antisemitism often manifests through images, memes, videos, or text-image combinations.
5) **Resource Limitations**: The analysis is constrained by computational resource constraints. Larger models or more advanced architectures require expensive infrastructure and increased energy consumption, creating challenges for widespread, real-time implementation.

Based on our findings and considering these limitations, we propose several promising directions for future research and implementation:

1) **Addressing Term Combination Sensitivity**: Future work should specifically target the identified term clusters that frequently trigger false positives. Fine-tuning approaches could be developed that better distinguish between different contexts in which sensitive terms appear, particularly when they co-occur with URLs.
2) **Models with Deep Understanding Capabilities**: There is a need for base models with better capabilities for context understanding and pragmatics, especially for detecting indirect, ironic, or coded antisemitic expressions. Such models, even if not the largest in parameter

count, should be selected based on their performance in complex understanding tasks.

3) **Multimodal Capabilities**: Developing systems capable of analyzing antisemitic content across multiple modalities (text, image, audio) is critical, especially in an era of image-based social networks and memes. Multimodal models can capture aspects of antisemitism that text-only models might miss.
4) **Efficient Fine-tuning Strategies**: Rather than relying on increasingly larger models, future research should focus on parameter-efficient fine-tuning techniques that enable task-specific adaptation without the full costs of very large models.

C. Practical Implementation Strategies

Based on our findings, we propose a framework for antisemitism detection that addresses the critical balance between precision and recall:

1) **Multi-layered Ensemble Approach**: Implement a multi-layered detection system combining models with complementary strengths for different stages of classification:
   - Initial filtering by lightweight, precision-focused models with basic multimodal capabilities
   - Deep analysis of uncertain cases using models with enhanced context understanding capabilities
   - Option for human review referral for particularly complex cases
2) **Content-Based Optimization**: Create classification strategies tailored to specific content characteristics:
   - Approaches adapted to specific keywords, with varying sensitivity for terms with high trigger potential
   - Length-adapted methods, recognizing that content of different lengths requires different analysis approaches
   - Special handling for content containing URLs alongside sensitive terms, addressing the most common false positive patterns identified in our topic modeling
3) **Resource-Efficient Techniques**: Develop strategies balancing performance and efficiency:
   - Use of small, focused models for initial filtering of most content
   - Deployment of larger, more complex models only for cases requiring deep analysis
   - Implementation of parameter-efficient fine-tuning techniques to improve performance without dramatically increasing resource requirements

VI. Conclusion

Our research demonstrates both the utility and limitations of definitional frameworks for antisemitism detection when implemented through LLM prompting, while highlighting the nuanced impact of model scaling on performance. While larger models provide measurable improvements in recall, they come with substantial costs in false positive rates that must be carefully considered in practical applications.

The significant differences in performance across keywords and tweet lengths underscore the need for tailored approaches rather than one-size-fits-all solutions. Our topic modeling analysis reveals specific term combinations—particularly those involving "jews," "https," and "israel"—that consistently trigger false positives, providing concrete targets for improvement in future models.

The JDA framework consistently demonstrates better precision than IHRA across model sizes, while IHRA offers superior recall—suggesting that framework selection should be guided by specific application priorities.

Effective antisemitism detection requires progress in three key areas: (1) developing base models with deeper understanding capabilities and inherent multimodal abilities; (2) improving efficient fine-tuning techniques that enable task-specific adaptation without parameter bloat; and (3) creating intelligent ensemble architectures that combine models with complementary advantages. By addressing these challenges, we can develop more effective systems for detecting and moderating antisemitic content while respecting the nuances of legitimate political discourse.

Acknowledgments

We acknowledge the DeepSeek team for their open-source models. Claude 3.7 Sonnet assisted with code development and content formulation. All code used in this research is available at https://github.com/omertarshish/antisemitism-detection-llm.

References

[1] Anti-Defamation League, "Quantifying Hate: A Year of Anti-Semitism on Twitter." Report, 2018.
[2] M. Behar, et al., "The Image of Israel on X and TikTok after 10/7." Research Report. Principal Investigator: Gunther Jikeli. Bloomington, IN: Research Lab Social Media & Hate, Institute for the Study of Contemporary Antisemitism, Indiana University, 2024.
[3] M. Chandra, et al., "'Subverting the Jewtocracy': Online Antisemitism Detection Using Multimodal Deep Learning." arXiv:2104.05947 [cs.MM], 2021.
[4] K. H. Halevy, et al., "On the Importance of Nuanced Taxonomies for LLM-Based Understanding of Harmful Events: A Case Study on Antisemitism." In Eighth Widening NLP Workshop (WiNLP 2024) Phase II, 2024.
[5] International Holocaust Remembrance Alliance, "Working Definition of Antisemitism." Official Definition, 2016.
[6] Jerusalem Declaration on Antisemitism, "The Jerusalem Declaration on Antisemitism." Official Declaration, 2021.
[7] G. Jikeli, D. Cavar, and D. Miehling, "Annotating Antisemitic Online Content. Towards an Applicable Definition of Antisemitism." doi: 10.5967/3R3M-NA89, 2019.
[8] G. Jikeli, et al., "Antisemitic Messages? A Guide to High-Quality Annotation and a Labeled Dataset of Tweets." arXiv:2304.14599 [cs.CL], 2023.
[9] D. Kikkisetti, et al., "Using LLMs to discover emerging coded antisemitic hate-speech in extremist social media." arXiv:2401.10841 [cs.CL], 2024.
[10] O. Nyrén, "Can Hatescan Detect Antisemitic Hate Speech?" Bachelor's dissertation, Department of Computer and Systems Sciences, Stockholm University, 2023.
[11] S. Ozalp, et al., "Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech." In Social Media + Society 6. doi: 10.1177/2056305120916850, 2020.

[12] E. Steffen, et al., “Codes, Patterns and Shapes of Contemporary Online Antisemitism and Conspiracy Narratives – an Annotation Guide and Labeled German-Language Dataset in the Context of COVID-19.” arXiv:2210.07934 [cs.CL], 2022.
[13] S. Zannettou, et al., “A Quantitative Approach to Understanding Online Antisemitism.” arXiv:1809.01644 [cs.CY], 2019.