SYNTHETIC DATA GENERATION USING LLMS FOR HATE SPEECH DETECTION IN POLITICAL POSTS

Completed Short Paper

Tintu Koshy, Toronto Metropolitan University, Toronto, Canada, tintu.koshy@torontomu.ca

Omar Ahmed, Toronto Metropolitan University, Toronto, Canada, omar1.ahmed@torontomu.ca

Zeinab Noorian, Toronto Metropolitan University, Toronto, Canada, znoorian@torontomu.ca

Amira Ghenai, Toronto Metropolitan University, Toronto, Canada, aghenai@torontomu.ca

Abstract

Hate speech detection in political discourse is hindered by the scarcity of domain-specific hate examples and severe class imbalance in election-related data. To address this challenge, we develop a topic-aware synthetic data generation pipeline that uses large language models to produce contextually grounded hate-speech samples aligned with discourse from the 2024 U.S. election. We manually annotate 6,499 tweets, apply BERTopic to identify thematic structure, and generate synthetic hate tweets conditioned on representative examples and topic-level cues. These synthetic samples are combined with the original dataset to fine-tune transformer-based classifiers. The augmented dataset yields significant improvements in hate-speech detection, with the best-performing model increasing its Hate-class F1 score from 0.67 to 0.88 after augmentation. These findings demonstrate that LLM-generated synthetic data can effectively enrich rare hate expressions and substantially enhance classifier performance in politically charged contexts.

Keywords: Political Hate Speech, Synthetic Data Generation, Large Language Models

1 Introduction

Recent estimates indicate that more than two-thirds of the global population uses social media, and this number continues to grow each year (Kemp, 2025). While these platforms enable open expression and public engagement, they also create environments where hateful and offensive content can spread rapidly. Posts targeting individuals or groups based on religion, race, nationality, gender, or other identities often trigger cascades of similar comments and replies, amplifying harmful discourse (Mathew et al., 2019). In politically sensitive periods, such as elections, the prevalence and visibility of hate speech can intensify, with documented cases showing that online hostility may escalate into offline tensions and real-world consequences (Grimminger & Klinger, 2021). Automated hate-speech detection systems play a central role in moderating such content at scale, but their effectiveness is limited by data scarcity, particularly for hate expressed in specific contexts such as political discussions. Existing hate-speech datasets tend to be small, imbalanced, or misaligned with the linguistic characteristics of political discourse, making it difficult for machine learning models, especially transformer-based classifiers, to learn reliable patterns (Grimminger & Klinger, 2021). Although data augmentation using external corpora has been explored, prior studies show that generic hate-speech datasets often fail to capture domain-specific language, reducing their impact on model performance (Arango et al., 2019; Li et al., 2023; Vidgen & Derczynski, 2020).

To address this gap, our research investigates whether synthetic hate-speech data generated by large language models (LLMs) can help strengthen detection models in politically charged settings. Because

hateful posts form only a small fraction of election-related discourse, we use synthetic generation to expand the pool of domain-specific hate examples while preserving political context. We develop a topic-aware pipeline that uses GPT-3.5-turbo to generate synthetic hate tweets grounded in real examples from the 2024 U.S. election. Unlike prior work that focuses on zero-shot or few-shot classification using LLMs (Bang et al., 2023; García-Díaz et al., 2023; Liang et al., 2023), our approach leverages LLMs as data generators to enrich underrepresented hate categories and improve downstream classifier robustness. This paper makes the following key contributions:

- A topic-aware synthetic generation pipeline grounded in real political discourse.
- A systematic comparison of baseline and augmented BERT-based hate-speech classifiers.

More broadly, the study contributes to IS research on sociotechnical system design by showing how generative AI can support moderation pipelines in data-scarce settings, and it offers practical guidance for organizations seeking to build more robust content-governance workflows in politically sensitive environments.

2 Related Work

Hate speech on social media, particularly on Twitter, has been widely studied across platforms and contexts (Paz et al., 2020). Early detection work relied on traditional machine-learning methods, but deep learning models soon demonstrated clear performance advantages (Subramanian et al., 2023). With the introduction of transformer architectures, models such as BERT and RoBERTa became the strongest performers due to their attention mechanisms and contextual representations (Vaswani et al., 2023). Empirical studies show that BERT outperforms earlier neural models on general and domain-specific hate-speech datasets (Alatawi et al., 2021), while ensembles can further improve performance (Kovács et al., 2022). More recent work has explored zero and few-shot LLM classification (García-Díaz et al., 2023), but these approaches still underperform fine-tuned transformers.

However, general advances in hate-speech detection do not fully reflect the unique challenges posed by political communication, particularly during election periods. Specifically, previous research examining hate and offensive speech during U.S. election cycles shows that political events can shape online hostility in complex ways. Studies of the 2016 and 2020 elections (Ali et al., 2022; Grimminger & Klinger, 2021) highlight that while overall levels of hate speech remain relatively low, spikes often occur in response to major campaign moments or policy announcements. These works also emphasize the difficulties of detecting hate in political discourse due to subjective language, topic-specific variation, and highly imbalanced datasets; for example, Grimminger and Klinger (2021) report that only 11.7% of their annotated election tweets were hateful. More recent analysis of violent political rhetoric on Twitter during the 2020 election (Kim, 2023) similarly finds extremely low prevalence rates but sharp increases around key events such as the Capitol Riot. Overall, this literature highlights the challenges of reliably identifying hate rhetoric in political contexts and signals the need for improved, domain-specific models.

In addition to these domain-specific challenges, hate-speech detection is further complicated by severe class imbalance, leading researchers to explore data augmentation approaches. Kovács et al. (2022) addressed this by supplementing training data for a hybrid CNN–LSTM model, while more recent work has explored LLM-based augmentation. Khullar et al. (2024) used BLOOM with few-shot prompting to generate hate speech in low-resource languages, and Zhuoyan Li et al. (Li et al., 2023) applied zero-shot and few-shot prompting with GPT-3.5 to create synthetic samples for BERT fine-tuning, though synthetic data performed below real data. Girón et al. (2025) used unmoderated models such as Mistral-7B-Instruct-v0.2 to generate toxic text aligned with seed examples, achieving modest gains in low-resource settings.

Prior synthetic augmentation studies largely overlook temporal shifts in political hate speech; our method addresses this through topic- and time-aware conditioning based on election-period data. Synthetic augmentation in political discourse may also amplify bias and over-represent explicit hate.

3 Methodology

Our methodology examines whether LLMs can generate synthetic hate-speech data that reflects real political discourse. We begin by constructing a domain-specific dataset using the publicly available X/Twitter election corpus released by Balasubramanian et al. (2024), which contains over 255 million posts from the 2024 U.S. presidential election cycle. To capture a period of heightened political tension and rhetorical variability, we extract tweets posted between 17 October and 10 November 2024, a three-week window covering both the final campaign phase and immediate post-election reactions. Prior work shows that political polarization and hostile rhetoric intensify around election events (Howard et al., 2018; Muñoz et al., 2024), making this interval well suited for examining whether LLMs can reproduce real-world linguistic patterns.

After cleaning the dataset, we manually annotate a subset of tweets, apply BERTopic to identify major themes, and observe sparsity in several hate-related clusters. We also test HateXplain-based augmentation, but stylistic mismatch limits its suitability. We therefore develop a topic-aware synthetic generation pipeline and evaluate several transformer-based classifiers on the original and augmented datasets. Figure 1 provides an overview of the full topic-aware synthetic data generation and validation pipeline used in this study.

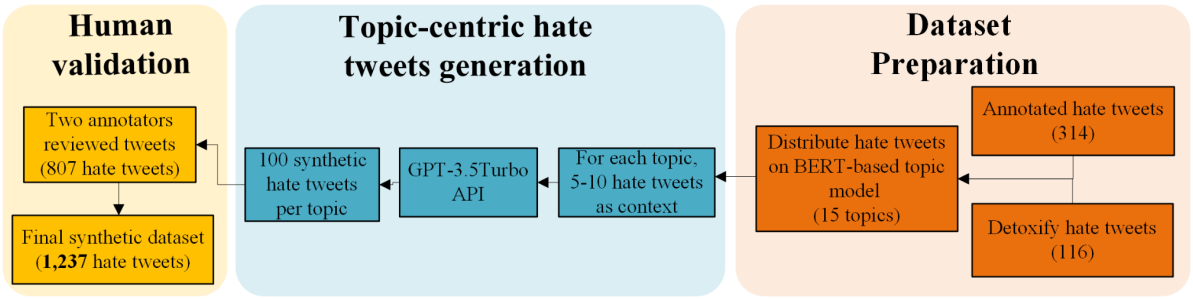

Figure 1. Overview of the topic-aware synthetic hate-tweet generation pipeline.

3.1 Dataset and Preprocessing

The dataset underwent several preprocessing and filtering steps to prepare it for analysis. First, we cleaned the tweet text by removing URLs, hashtags, user mentions, non-alphabetic characters, extra whitespace, and newline symbols. We also discarded duplicates, non-English posts, and tweets containing only emojis. To retain meaningful political discourse, we compiled a custom list of election-related keywords[1], including candidate names, party references, and common political terms. Tweets containing at least one of these keywords were retained. The final dataset contained 701,868 tweets from 365,886 users, with highly skewed activity (1.92 tweets/user), mean length of 169 characters, 23.2% containing URLs, and average engagement of 9.72 retweets and 4.52 replies.

3.2 Manual Annotation for Ground Truth

A total of 6,499 tweets were randomly sampled from the dataset. The tweet text was manually annotated to identify hate, normal (non-hate), and offensive posts. Three annotators labelled each tweet as *Hate*, *Normal*, or *Offensive.* Table 1 lists the definition of each label with an example.

[1] Keywords = [2024 Elections, 2024 Presidential Election, Biden, Biden2024, conservative, CPAC, Donald Trump, GOP, Joe Biden and Kamala Harris, Joe Biden, Joseph Biden, KAG, MAGA, Nikki Haley, RNC, Ron DeSantis, Snowballing, Trump2024, trumpsupporters, trumptrain, US Elections, thedemocrats, DNC, Kamala Harris, Marianne Williamson, Dean Phillips, williamson2024, phillips2024, Democratic party, Republican party, Third Party, Green Party, Independent Party, No Labels, RFK Jr, Robert F. Kennedy Jr., Jill Stein, Cornel West, ultramaga, voteblue2024, letsgobrandon, bidenharris2024, makeamericagreatagain, Vivek Ramaswamy]

Label	Definition	Sample tweet from dataset
Hate	Attacks or dehumanizes a protected group or individual based on identity	*Stop being a white supremacist boot licking Nazi loving MAGA*
Normal	News, support,personal opinion, or respectful disagreement, without offensive language	*Because he's supporting Donald Trump in this election*
Offensive	Insults or uses strong negative language, mainly targeting an individual's behaviour	*Kamala Harris cant stop lying about Trump Pray for her soul*

Table 1. Definition of labels along with sample tweets.

The labeling resulted in a Fleiss' Kappa and Krippendorff's Alpha score of 0.55, indicating moderate agreement among annotators. This level of agreement reflects the interpretive difficulty of distinguishing hate, offensive, and non-hate content in political discourse, where expressions are often implicit, context-dependent, or rhetorically charged. The majority label, when at least two annotators agreed, was assigned as the final label for each tweet. Tweets for which all three annotators disagreed were labelled as *No Majority* and were removed from further analysis. The annotated dataset contains 314 hate tweets (4.83%), 3,201 offensive tweets (49.25%), 2,902 normal tweets (44.65%), and 82 tweets (1.26%) with no majority agreement.

3.3 Topic Modeling for Context Extraction

To capture the thematic structure of political discourse and ensure that synthetic hate-speech generation remained contextually grounded, we applied BERTopic[2] to all tweets. We chose BERTopic because it produces coherent, interpretable clusters using transformer-based embeddings and class-based TF–IDF, which is particularly effective for short, noisy political text on Twitter.

Before finalizing the model, we experimented with several topic configurations, testing both smaller solutions (8–12 topics) and larger ones (20–25 topics). Models with too few topics merged distinct political conversations into overly broad themes, while models with too many topics fragmented the discourse into unstable, low-coherence clusters. The most balanced and interpretable outcome emerged from the 15-topic solution, which we generated using the following parameters: the all-MiniLM-L6-v2 embedding model, UMAP for dimensionality-reduction settings (15 neighbors and 5 components), HDBSCAN for clustering, and the default c-TF–IDF settings for keyword extraction. We fit BERTopic on the full preprocessed corpus and used the resulting topic assignments to label each tweet with a topic ID. We then joined these topic IDs with the manually annotated dataset using tweet-level identifiers, which enabled us to map the original hate tweets onto the topic clusters and calculate the hate-tweet distribution across topics.

In Table 3, we present the top 5 topics, the distribution of tweets across these topics, and the assigned topic labels. Topic labels were drafted with ChatGPT and verified by the research team. When we linked the original hate tweets onto these topics, only five topics contained more than 20 hate tweets, several contained fewer than 10, and Topic 11 contained none. Because our synthetic-generation pipeline relies on original hate tweets as seed examples to maintain contextual grounding, these sparse clusters highlighted where additional synthetic samples were required to strengthen topic diversity across the dataset.

Topic #	Tweet Count	Suggested Label
1	73,695	Biden Criticism & Economic Performance
2	73,091	Kamala Harris, Identity Politics & Campaign Messaging
3	68,842	Trump 2024 Campaign & Supporter Narratives
4	39,834	Democratic vs. Republican Party Politics

[2] BERTopic library: https://github.com/MaartenGr/BERTopic (A transformer-based topic modeling technique that uses document embeddings, UMAP, HDBSCAN, and class-based TF-IDF to generate interpretable topics.)

5	20,643	General Political Reactions & Public Commentary

Table 3. Summary of BERTopic-derived top 5 topics and their tweet frequencies for the 2024 U.S. election dataset.

3.4 Data Augmentation Using External Corpora

Our manually annotated dataset contained fewer than 5% hateful tweets (Table 1), resulting in a substantial class imbalance that limited the number of real, in-domain hate examples available for model training, augmentation, and prompt conditioning. To increase the pool of hate samples, we first incorporated additional labeled tweets from the HateXplain corpus (Kovács et al., 2022). HateXplain provides high-quality annotations for hate, offensive, and normal content; however, its linguistic style differs from election-related content, raising questions about its suitability for in-domain augmentation. Because external corpora may introduce stylistic or contextual drift, we next explored synthetic data generation as a targeted strategy for expanding the hate-speech class while preserving domain alignment. In this step, we used a large language model to generate new hate tweets conditioned on real examples selected from our dataset. Conditioning was applied to maintain topic relevance, rhetorical framing, and political vocabulary. The following subsection describes our synthetic generation framework, including prompt and topic constraints, and quality-control procedures.

3.5 Topic-Aware Synthetic Tweet Generation

The labeled corpus contains less than 5% hateful tweets, which constrains model learning and underrepresents the linguistic diversity of political hate. To mitigate the scarcity of hateful content while preserving domain context, we generated synthetic tweets conditioned on in-domain examples (a.k.a tweets) and topic context. Specifically, we prompted GPT-3.5-turbo using representative examples and keywords from our BERTopic clusters to better preserve topic relevance, lexical style, and rhetorical cues observed in election discourse.

3.5.1 Prompting and context

For each of the 15 topics, we supplied 5–10 original hate tweets (pre-cleaned to remove URLs, mentions, and symbols) alongside the topic's keywords. Ten tone-style combinations were defined as follows: incendiary-conspiracy-laced, threatening-violent metaphors, aggressive-stereotype-driven, dismissive-hashtag mockery, dehumanizing-emoji-sarcasm, explicit-slang-heavy, fear-mongering-racial-codewords, angry-ALL CAPS + excessive punctuation, coded-xenophobic insinuations, provocative-ethnic slurs. Furthermore, three tweet length options [Short (5-15 words), Medium (15-30 words), Long (30-50 words)] were also specified in the prompt.

3.5.2 Expanding seeds for sparse topics

As is discussed in section 3.3, some topics contained few or no hate tweet examples. To avoid overfitting the generator to a tiny seed set, we randomly selected 500 tweets from the entire dataset for the topics with fewer than 20 original hate tweets. We applied Detoxify model to these tweets to identify hateful tweets with a high confidence (threshold set as 0.9 . Two annotators further manually reviewed these newly classified hate tweets resulting in the identification of an additional 116 hate tweets which expanded the hate tweet count to 430.

3.5.3 Generation process

We aimed for 100 synthetic tweets per topic (10 tone-style -pairs x 10 tweets each) using Open AI's GPT-3.-turbo API where seeds were limited in a particular topic (e.g., topic with <~40), and context examples were necessarily reused across multiple style and tone profiles. However, when seeds were abundant (e.g., topics with ~70 hate tweets), unique examples were distributed to diversify outputs. This resulted in a generation of a total of 1,600 synthetic hate tweets.

3.5.4 Quality control and human validation

We applied a two-stage filtering approach using Sentence-BERT[3] to remove duplicated synthetic tweets. In the first stage, within each topic, we eliminated tweets with high similarity (cosine scores >0.75) to their seed examples or to other synthetics in that topic. In the second stage, we concatenated all topics, and further identified and removed synthetic tweets that are similar (threshold 0.75) across different topics. The resulting pool comprised **1,178** synthetic tweets. Two annotators then independently labeled each item as hateful or non-hateful; we retained **807** tweets for which annotators reached substantial agreement, as indicated by a Cohen's Kappa score of 0.68, and used them as the final synthetic hate tweet set.

3.6 Model Fine-Tuning and Evaluation

To assess the impact of synthetic data augmentation on hate-speech classification, we fine-tuned three BERT-based models commonly used for offensive-language detection. The first model, Hate-speech-CNERG/bert-base-uncased-hatexplain[4], is pretrained on the HateXplain corpus and provides a baseline trained on general hate-speech data. The second and third models, twitter-roberta-base-offensive[5] and twitter-roberta-base-dec2021-offensive, are RoBERTa variants pretrained on large-scale Twitter corpora for offensive and abusive language detection. These models were selected to compare performance across architectures with varying degrees of domain alignment.

All models were fine-tuned using our annotated dataset, both in its original form and in an augmented version that incorporates synthetic hate-speech samples[6]. Fine-tuning followed standard transformer training procedures, using a three-class classification schema (Normal, Offensive, Hate). We used an 80/20 train–test split and applied the same preprocessing steps across all experiments to ensure comparability. Evaluation was conducted using precision, recall, and F1 score for each class, along with overall macro-averaged metrics.

4 Results

To evaluate the impact of synthetic tweet augmentation, we compared the performance of three transformer-based classifiers (described in section 3.6) before and after the inclusion of topic-aware synthetic hate tweets (a.k.a Original dataset vs augmented dataset). As shown in Table 4, all three models exhibited substantial improvements in detecting hate speech, particularly in the underrepresented "Hate" class. The F1 scores for HateXplain (CNERG), twitter-roberta-base-offensive, and twitter-roberta-base-dec2021-offensive increased from 0.46, 0.60, and 0.67 to 0.86, 0.88, and 0.88, respectively. These gains were driven by improvements in both precision and recall, indicating better identification of hateful content and fewer false positives. The results suggest that the inclusion of synthetic tweets—generated using in-domain examples and topic-specific keywords—enabled the models to learn more nuanced patterns of hate expression that are contextually grounded in political discourse. This augmentation approach improved classifiers robustness and enhanced the models' ability to generalize beyond the limited examples present in the original dataset.

Model	Metric	Original Dataset	Augmented Dataset
Hate speech CNERG model	Precision	0.51	0.98

[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT networks," *EMNLP*, 2019. https://arxiv.org/abs/1908.10084

[4] S. Mathew et al., "HateXplain: A benchmark dataset for explainable hate speech detection," in Proc. AAAI, vol. 35, no. 17, 2021, pp. 14867–14875. https://arxiv.org/abs/2012.10289

[5] F. Barbieri, J. Camacho-Collados, L. Neves and L. Espinosa-Anke, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in Proc. EMNLP, 2020. https://arxiv.org/abs/2010.12421

[6] GitHub repository: TTKoshy/synthetic_data_generation_using_llms: Synthetic Data Generation using LLMs for Hate Speech Detection in Political Posts

	Recall	0.42	0.77
	F1 score	0.46	0.86
Twitter-roberta-base-offensive	Precision	0.56	0.90
	Recall	0.65	0.86
	F1 score	0.60	**0.88**
Twitter-roberta-base-dec2021-offensive	Precision	0.67	0.89
	Recall	0.67	0.88
	F1 score	0.67	**0.88**

*Table 4. Performance of BERT-Based models to identify **hate** tweets before and after synthetic data augmentation.*

5 Conclusion

This study examined whether synthetic hate-speech data generated by large language models can improve hate-detection in political discourse. Using a manually annotated subset of 2024 U.S. election tweets, we developed a topic-aware pipeline to create synthetic hate samples aligned with real political language. Incorporating these samples into transformer-based classifiers led to substantial improvements in detecting hate speech, with Hate-class F1 scores rising from 0.46–0.67 to 0.86–0.88 after augmentation. These gains indicate that domain-aligned synthetic data can effectively address class imbalance and enhance model robustness in politically sensitive contexts. The study also contributes to IS research and practice by showing how generative AI can support moderation workflows in data-scarce settings, while underscoring the need for careful validation in politically sensitive environments. Limitations include the use of a single election-related dataset, a limited set of classifiers, and the risk that synthetic hate samples may introduce bias or be misused.

References

Alatawi, H. S., Alhothali, A. M., & Moria, K. M. (2021). Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding With Deep Learning and BERT. *IEEE Access*, *9*, 106363–106374. https://doi.org/10.1109/ACCESS.2021.3100435

Ali, R. H., Pinto, G., Lawrie, E., & Linstead, E. J. (2022). A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00633-z

Arango, A., Pérez, J., & Poblete, B. (2019). *Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation*.

Balasubramanian, A., Zou, V., Narayana, H., You, C., Luceri, L., & Ferrara, E. (2024). *A Public Dataset Tracking Social Media Discourse about the 2024 U.S. Presidential Election on Twitter/X* (arXiv:2411.00376). arXiv. https://doi.org/10.48550/arXiv.2411.00376

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity* (arXiv:2302.04023). arXiv. https://doi.org/10.48550/arXiv.2302.04023

García-Díaz, J. A., Pan, R., & Valencia-García, R. (2023). Leveraging Zero and Few-Shot Learning for Enhanced Model Generality in Hate Speech Detection in Spanish and English. *Mathematics*, *11*(24), 5004. https://doi.org/10.3390/math11245004

Girón, A., Huertas-Tato, J., & Camacho, D. (2025). LLM synthetic generation to enhance online content moderation generalization in hate speech scenarios. *Computing*, *107*(8), 164. https://doi.org/10.1007/s00607-025-01518-8

Grimminger, L., & Klinger, R. (2021). *Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection* (arXiv:2103.01664). arXiv. https://doi.org/10.48550/arXiv.2103.01664

Howard, P. N., Kollanyi, B., Bradshaw, S., & Neudert, L.-M. (2018). *Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States?* (arXiv:1802.03573). arXiv. https://doi.org/10.48550/arXiv.1802.03573

Kemp, S. (2025). *Digital 2025: The state of social media in 2025—DataReportal—Global Digital Insights*. https://datareportal.com/reports/digital-2025-sub-section-state-of-social

Khullar, A., Nkemelu, D., Nguyen, V. C., & Best, M. L. (2024). Hate Speech Detection in Limited Data Contexts Using Synthetic Data Generation. *ACM Journal on Computing and Sustainable Societies*, *2*(1), 1–18. https://doi.org/10.1145/3625679

Kim, T. (2023). Violent political rhetoric on Twitter. *Political Science Research and Methods*, *11*(4), 673–695. https://doi.org/10.1017/psrm.2022.12

Kovács, G., Alonso, P., Saini, R., & Liwicki, M. (2022). *Leveraging external resources for offensive content detection in social media*. https://journals-sagepub-com.ezproxy.lib.torontomu.ca/doi/10.3233/AIC-210138

Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). *Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations* (arXiv:2310.07849). arXiv. https://doi.org/10.48550/arXiv.2310.07849

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … Koreeda, Y. (2023). *Holistic Evaluation of Language Models* (arXiv:2211.09110). arXiv. https://doi.org/10.48550/arXiv.2211.09110

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2019). Thou Shalt Not Hate: Countering Online Hate Speech. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 369–380. https://doi.org/10.1609/icwsm.v13i01.3237

Muñoz, P., Bellogín, A., Barba-Rojas, R., & Díez, F. (2024). Quantifying polarization in online political discourse. *EPJ Data Science*, *13*(1), 39. https://doi.org/10.1140/epjds/s13688-024-00480-3

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate Speech: A Systematized Review. *Sage Open*, *10*(4), 2158244020973022. https://doi.org/10.1177/2158244020973022

Subramanian, M., Easwaramoorthy Sathiskumar, V., Deepalakshmi, G., Cho, J., & Manikandan, G. (2023). A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, *80*, 110–121. https://doi.org/10.1016/j.aej.2023.08.038

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, *15*(12), e0243300. https://doi.org/10.1371/journal.pone.0243300