Detecting Antisemitic Hate Speech using Transformer-based Large Language Models

Dengyi Liu and Minghao Wang, M.S. in Data Analytics and Visualization

Faculty Advisor: Andrew G. Catlin

Katz School of Science and Health

ABSTRACT

Identifying hate speech by both academic institutions and social media organizations can pose as a challenge given the large scale of data and changing patterns of hate speech content. Advances since 2020 in both transformer-based and Generative AI served as a platform to create novel solutions that help address both of the aforementioned challenges. The research team created a data labeling method and built a proof of content for antisemitic hate speech that allowed for the comparing and contrasting of these state-of-the-art analytical approaches. This research can play a pivotal role in upholding social harmony by refining content moderation processes to quickly and effectively mitigate the spread of harmful speech.

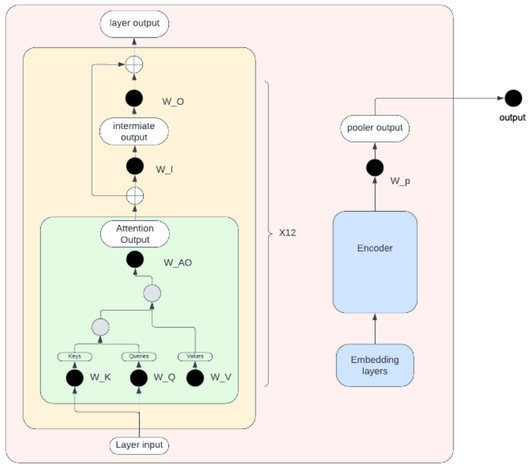
INTRODUCTION

- The increase in online antisemitic hate speech poses significant challenges, necessitating effective detection and mitigation strategies.
- To address the critical issue of antisemitic hate speech detection online, this project explored utilization of advanced natural language processing (NLP) techniques, particularly transformer-based models such as: **BERT** (Devlin et al., 2019; Mozafari et al., 2019); **DistillBERT** (Sanh et al., 2020; Jiao et al., 2019) and **LLaMA-2** (Touvron et al., 2023).
- The project centers on evaluating the effectiveness of transformer-based NLP models in accurately identifying antisemitic content, thereby enhancing content moderation processes.
- The study leveraged recent advancements in machine learning and NLP to develop a nuanced detection framework.
- Models such as BERT, DistillBERT, and LLaMA-2, known for their deep linguistic understanding, were applied to detect nuanced and overt antisemitic expressions.
- This project aimed to contribute to the broader field of online safety and hate speech detection, offering insights into the practical application of transformer-based models in combating antisemitic expressions online.

METHODOLOGY

The approach combined data labeling, model training, and performance evaluation:

- **Data Preparation**: Researchers collected thousands of Twitter/X posts and employed a systematic voting algorithm by a dedicated team to classify each text as ‘antisemitic' or 'neutral’.
- **Model Training**: Researchers utilized the pre-trained BERT model for the transformer-based model, fine-tuning LLaMA-2 for the Large Language Models, both on the curated dataset for optimal hate speech detection.
- **Evaluation**: The models’ accuracy, precision, recall, and F1 score were measured to ensure robust detection capabilities against unseen test data.



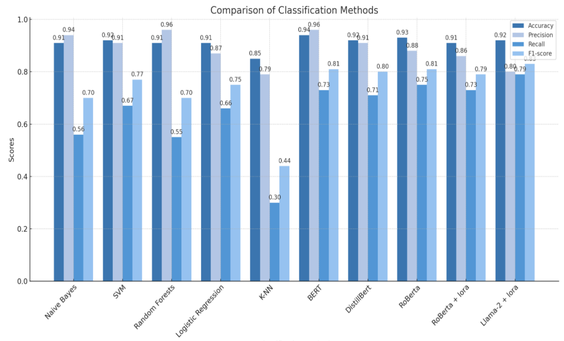
RESULTS

The performance of several machine learning and NLP models were applied to the dataset. The models included Naïve Bayes, SVM, BERT, DistillBERT, LlaMa-2 and RoBERTa (Liu et al., 2019).

Each model was selected based on its relevance to text classification and its unique approach to understanding language. The evaluation aimed to determine the most effective model in terms of accuracy, efficiency, and processing time.

Table 1. Comparison of Classification Methods

Method	Accuracy	Precision	Recall	F1-score
Naive Bayes	0.91	0.94	0.56	0.70
SVM	0.92	0.91	0.67	0.77
Random Forests	0.91	0.96	0.55	0.70
Logistic Regression	0.91	0.87	0.66	0.75
K-NN	0.85	0.79	0.30	0.44
BERT	0.94	0.96	0.73	0.81
DistillBert	0.92	0.91	0.71	0.80
RoBerta	0.93	0.88	0.75	0.81
RoBerta + lora	0.91	0.86	0.73	0.79
Llama-2 + lora	0.92	0.80	0.79	0.83



DISCUSSION & CONCLUSIONS

Models like BERT effectively identify antisemitic hate speech, underscoring the value of recent advances in NLP for content moderation.

Challenges: Contextual hate speech detection remains complex, highlighting the need for advanced linguistic analysis and bias mitigation.

Recommendations: Ongoing model refinement will be crucial due to the evolving nature of online hate speech.

Ethics and Expansion: Ethical deployment and model transparency are essential, with future work to include multilingual capabilities.

Impact: This study is a step forward in automated hate speech detection, with significant implications for online safety and inclusivity (Warner et al., 2012; Davison et al, 2017; Yuan et al., 2019). These findings contribute to safer digital spaces, requiring collective efforts for sustained progress.

REFERENCES

Devlin, J., Chang, M., Lee, K., and Kristina Toutanova. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. https://arxiv.org/abs/1910.01108

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. https://arxiv.org/abs/1907.11692

Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288, 2023.

Warner, W., and Hirschberg, J. (2012) Detecting hate speech on the world wide web. Association for Computational Linguistics, Proceedings of the Second Workshop on Language in Social Media.

Davison, T., Warmsley, D., Macy, M., and Weber, I. (2017) Automated hate speech detection and the problem of offensive language. International AAAI Conference on Web and Social Media, 11.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017) Attention is all you need. NeurIPS.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2019). A bert-based transfer learning approach for hate speech detection in online social media. arXiv preprint arXiv:1910.12574

Yuan, L., Wang, T., Ferraro, G., Suominen, H., and Rizoiu, M. (2019). Transfer learning for hate speech detection in social media. arXiv preprint arXiv:1906.03829

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F. and Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351