NLP4CSS Final Project Proposal

Ashley Ren yren46@jh.edu
Swarali Mahimkar smahimk1@jh.edu
Yingfei Xu yxu238@jh.edu
Ruotong Zou rzou9@jh.edu

1 Introduction

As the internet has evolved, hate speech has become prevalent across various social media platforms(Gagliardone et al., 2015). It targets people based on their race, religion, sexual orientation, and other identities, and often causes real harm to individuals and communities(Gelber, 2021). Hate speech has far-reaching consequences for targeted groups, negatively affecting their physical and mental well-being, freedom of association, and autonomy(Brown, 2015). Given these serious harms, developing accurate and reliable methods for detecting hate speech is essential for creating safer online environments and protecting vulnerable communities.

Some studies suggest that hate speech targeting different groups uses different words, styles, and themes(Alkomah and Ma, 2022). For example, sexist hate speech often includes language aimed at demeaning women and reinforcing gender-based stereotypes. Women are frequently subjected to threats of sexual violence, online harassment, and moral shaming (Hardaker and McGlashan, 2016; Jane, 2016). However, there is still not enough research that clearly shows how these patterns vary between groups. Without this understanding, it is hard to improve detection systems or make them more accurate and inclusive.

To fill this gap, our first research question (**RQ1**) asks: How do the semantic patterns and topic clusters of hate speech differ across various target groups (such as race, religion, or gender) when analyzed using neural topic modeling techniques? This helps us explore what kinds of language and themes appear in hate speech directed at different communities.

Once we understand these differences, the next step is to ask whether current detection methods can handle them. While many tools rely on large amounts of computing resources, lightweight models like BERT and DistilBERT have shown strong performance even with little or no training data. These models are faster and easier to use, especially when time or data is limited. So our second research question (**RQ2**) asks: Are small pretrained models good enough to detect hate speech in zero-shot or few-shot settings, and how do they compare to traditional methods?

By answering these two questions, we aim to better understand how hate speech varies depending on the group being targeted, and whether existing language models can detect that variety in real-world settings. Our work brings together content analysis and machine learning evaluation, with the goal of helping researchers and developers build better, more reliable hate speech detection systems.

2 Related Work

Recent research has highlighted significant linguistic and thematic variation in hate speech targeting different identity groups, such as race, gender, or religion. (Yoder et al., 2022) empirically demonstrated that hate speech classifiers trained on specific identities often fail to generalize across different targets due to distinct semantic patterns and identity-specific references tied to historical prejudices and stereotypes. Such findings are reinforced by (Reyero Lobo et al., 2023), who emphasize the necessity of integrating explicit group-specific knowledge to improve detection accuracy and robustness against annotation biases. Similarly, (Casula and Tonelli, 2024) employed generative language models to augment data, significantly improving model performance for underrepresented hate targets like disabilities or religion, underscoring the need for target-aware dataset balancing.

In recent years, lightweight and low-resource hate speech detection methods have increasingly gained attention, with models such as DistilBERT and zero-shot classification showing promise due

to their efficiency and minimal training data requirements. Notably, zero-shot methods utilizing large language models like GPT-3 through carefully engineered prompts, enable classification without extensive task-specific training, although their reliability can be sensitive to prompt wording and inherent biases(Ye et al., 2024). Recent innovations further include combining few-shot learning with model distillation and incorporating domain-specific knowledge graphs, collectively enhancing model interpretability, efficiency, and adaptability across diverse hate speech contexts (Casula and Tonelli, 2024; Reyero Lobo et al., 2023).

3 Dataset and Methods

We use the **Measuring Hate Speech** dataset (Kennedy et al., 2020), which contains 39,565 unique social media comments with 135,556 total annotations. Annotations span 42 identity subgroups across 8 high-level target categories (e.g. race/ethnicity, religion, gender). Each comment is rated on 10 ordinal dimensions related to hatefulness (e.g. insult, dehumanization, violence, and genocide), which are then combined into a continuous hate speech score using Item Response Theory (IRT) adjustment.

To investigate our first research question (**RQ1**), we group comments according to their annotated target identity category and filter for likely hate speech (score > 0.5). We then apply neural topic modeling techniques to the filtered texts. Specifically, we experiment with BERTopic (Grootendorst, 2022) and Top2Vec (Angelov, 2020), two methods that combine transformer-based embeddings with clustering algorithms to identify common themes in text. For comparison, we also run traditional Latent Dirichlet Allocation (LDA) on the same subsets.

To address our second research question (**RQ2**), we evaluate the effectiveness of lightweight pre-trained models for hate speech detection. We test two types of classifiers: a zero-shot transformer-based model and a traditional machine learning baseline. For the zero-shot setup, we use the HuggingFace zero-shot-classification pipeline with BERT-base, assigning each comment a predicted label from a fixed candidate set (e.g., `["hate speech", "not hate speech"]`). As baselines, we train traditional Logistic Regression and linear SVM models on TF-IDF features, using the hate speech score threshold of 0.5 for binary labeling. Models are trained and evaluated on a stratified subset of the data to simulate few-shot scenarios.

4 Summary Statistics

The hate speech score in the dataset accounts for inter-annotator differences in interpretation and thus yields a more stable label. Scores range from -8.34 to 6.30, where values above 0.5 generally correspond to hate speech, values below -1 indicate counter-speech or support, and the range -1 to 0.5 captures ambiguous or neutral content. The mean hate speech score is -0.57, with a standard deviation of 2.38. Based on the > 0.5 threshold, 36.18% of comments are labeled as hate speech.

Figure 1 shows the distribution of the hate speech scores, which is roughly bimodel and centered around neutral speech, with a long rightward tail corresponding to highly hateful content.

5 Metrics

To evaluate our research questions comprehensively, we define specific metrics and evaluation criteria that align closely with our objectives.

For addressing **RQ1**, we measure the effectiveness of the neural topic modeling methods (BERTopic and Top2Vec) compared to traditional methods (LDA) using: **Topic Coherence Scores**: We use standard specifically the C_V *coherence* metric to assess the interpretability and semantic consistency of generated topics. **Qualitative Assessment**: We manually inspect representative comments to verify that the topics reflect meaningful semantic patterns relevant to target groups.

For **RQ2**, we evaluate the efficacy of lightweight pre-trained models in zero-shot and few-shot scenarios involves metrics including **Accuracy**: The overall proportion of correct classifications. **Precision and Recall**: Precision measures how accurately the models identify hate speech and recall measures how effectively they avoid missing actual hate speech instances. **F1 Score**: Particularly useful due to potential class imbalance, the F1 score balances precision and recall.

Overall, our research will be considered successful if it provides meaningful insights into the semantic and thematic differences in hate speech across targeted groups and shows that low-resource models can detect hate speech effectively in real-world scenarios.

References

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

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *Preprint*, arXiv:2008.09470.

Alex Brown. 2015. *Hate speech law: A philosophical examination*. Taylor & Francis.

Camilla Casula and Sara Tonelli. 2024. A target-aware analysis of data augmentation for hate speech detection. *Preprint*, arXiv:2410.08053.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

Katharine Gelber. 2021. Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Claire Hardaker and Mark McGlashan. 2016. "real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, 91:80–93.

Emma A Jane. 2016. Misogyny online: A short (and brutish) history.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Paula Reyero Lobo, Enrico Daga, Harith Alani, and Miriam Fernandez. 2023. Knowledge-grounded target group language recognition in hate speech. In *The Proceedings of SEMANTICS 2023, the 19th International Conference on Semantic Systems: Knowledge Graphs: Semantics, Machine Learning, and Languages*, volume 56 of *Studies on the Semantic Web*, pages 1–18. IOS Press.

Haotian Ye, Axel Wisiorek, Antonis Maronikolakis, Özge Alaçam, and Hinrich Schütze. 2024. A federated approach to few-shot hate speech detection for marginalized communities. *Preprint*, arXiv:2412.04942.

Michael Miller Yoder, Lynnette Hui Xian Ng, David West Brown, and Kathleen M. Carley. 2022. How hate speech varies by target identity: A computational analysis. *Preprint*, arXiv:2210.10839.

A Appendix

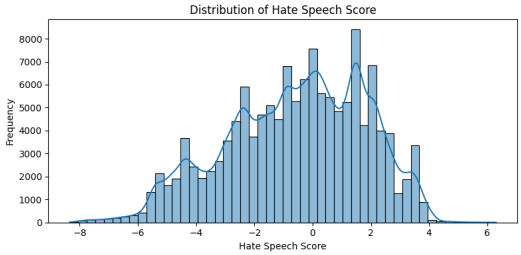

Figure 1: Distribution of hate speech scores.