CS 541 - Artificial Intelligence
Optional Project

Akshay Atam[1], Surya Giri[1], Wilson Jeng[1]

1. Introduction

The Internet is the source of a huge variety of knowledge repositories (Wikipedia, Wordnet, etc.) and apps (YouTube, Reddit, Twitter, etc.) that everyone may access and use; it is also the communication platform of our time and the most crucial tool to protect free speech. It enables us to openly express and share with large audiences our opinions on any private or public subject. But regrettably, it also makes it easier to control the masses and defame certain people or groups of people. The spread of hate speech is one of these identified bad phenomena. Hate speech encourages violence against the persons, groups, or populations it targets as well as a bad self-image and social marginalization. A good illustration of use of hate speech could be seen in the 1994 Rwandan genocide. It is imperative to handle hate speech and annotation provides a useful tool for machine learning algorithms to detect and prevent hate speech in the online world.

2. Background and Related Work

There is a plethora of hate speech datasets across all languages and annotation of such datasets is crucial as it facilitates the use of large machine learning models to determine whether a text is hate speech or not.

Annotation of hate speech text could vary based on the content that is available to us. The authors in [1] classified hate speech based in Arabic lexicon on the basis of Normal, Abusive or Hate while the works of [2] had 234 different annotators classifying hate speech based on religious groups that are targeted, Thus, there are various ways in which we could annotate hate speech. However, the most important detail to keep in mind is the consistency and agreement between annotators choosing the classification of hate speech.

3. Dataset Description

The dataset that needs to be annotated is HateModerate. It is a custom dataset which comprises of hate speech from various datasets as well as examples generated by GPT-3. The dataset is classified based on the Facebook Community

[1] Stevens Institute of Technology, Hoboken, NJ 07030

Guidelines [3] which comprises of three tiers, each describing in detail of what constitutes as hate speech. These guidelines were chosen as they include detailed description of each tier and included sub-tiers, giving us a total of 41 different guidelines which can be used as fine-grained labels.

4. Research Question

Large Language Models trained on hate speech dataset are capable of distinguishing hate speech from regular speech. A good LLM requires appropriate annotation to achieve state-of-the-art performance. Annotation of dataset poses a lot of research questions most notably,

- What are the ethical considerations in annotating hate speech datasets, including issues related to bias, fairness, and privacy?
- How do different annotation strategies, such as binary (hate speech vs. non-hate speech) vs. fine-grained (different types of hate speech) annotation, impact the performance of hate speech detection models?
- What are the inter-annotator agreement levels when annotating hate speech in datasets, and what are the potential sources of disagreement?
- How does the size and diversity of annotated datasets impact the performance and generalizability of hate speech detection models, and what are the trade-offs between dataset size and model performance?

5. Plan for experiment

To achieve an accurate level of annotation, the work would be divided equally among the group members. Annotation of work will be performed individually followed by cross check between the members.

References

[1] H. Mulki, H. Haddad, C. Bechikh Ali, H. Alshabani, L-HSAB: A Levantine Twitter dataset for hate speech and abusive language, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 111–118. doi:10.18653/v1/W19-3512. URL https://aclanthology.org/W19-3512

[2] N. Albadi, M. Kurdi, S. Mishra, Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 69–76. doi:10.1109/ASONAM.2018.8508247.

[3] Facebook community guidelines, https://transparency.fb.com/policies/community-standards/hate-speech/.