CS 541 - Artificial Intelligence Project

Warning: this paper discusses and contains content that can be offensive or upsetting

Akshay Atam
CWID: 20016304
aatam@stevens.edu

Surya Giri
CWID: 10475010
sgiri2@stevens.edu

Wilson Jeng
CWID: 10466128
wjeng@stevens.edu

Abstract

The goal of this project is to create an annotated hate speech dataset, which can be used for various research purposes, such as training machine learning models for hate speech detection. We used HateModerate, a custom dataset using hate speech examples from various datasets as well as leveraging the power of GPT-3 for generation. A well-defined set of guidelines from Facebook was used for categorizing hate speech across 3 tiers with a total of 41 sub-tiers. The annotation process will involve a team of human annotators, who will review each example and assign labels based on whether the example is in line with the section it falls. The resulting dataset will include 41 hate speech categories, relating to violent speech, dehumanizing speech based on filth, subhumanity, and more, social exclusion, economic exclusion, and more. This project has the potential to contribute to the development of more accurate and effective hate speech detection algorithms, which can help to combat hate speech and promote a more inclusive and respectful online environment.

1 Introduction

All that follows on this task's background we gained from Stevens Institute of Technology's CS 541 Artificial Intelligence course guided by Professor Xueqing Liu. This project falls under the Data preparation part of Artificial Intelligence. This domain involves human intelligence and expertise in labeling or annotating data for training AI models. Specifically, it relates to the subfield of AI called supervised learning, which focuses on training models using labeled data. The annotations provide the ground truth or reference labels necessary for supervised learning algorithms to learn from the data and make predictions on new, unseen data. These annotations might involve labeling objects in images, categorizing text, assigning sentiment scores, or identifying specific features or characteristics.

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.

The rest of the paper is broken into following sections:

- Section 2 encapsulates our background work, looking at various annotations and their methods.
- Section 3 provides detailed answers the research questions showcased in the project proposal.
- Section 4 outlines our methodology of annotation.
- Section 5 showcases our results with examples of annotations that were fit for the category and unfit for the category.
- Section 6 concludes our paper, and

- Section 7 deals with the future scope of the paper.

2 Background and Related work

The process of annotating datasets plays a crucial role in various domains of research and application, ranging from natural language processing to computer vision and beyond. Annotated datasets provide valuable labeled examples that enable the development and training of machine learning models. These models, in turn, play a vital role in tasks such as sentiment analysis, object recognition, spam detection, and more. The quality and accuracy of annotations directly impact the performance and generalization capabilities of these models. However, annotating datasets is a complex and labor-intensive task that requires human expertise and domain knowledge. It involves carefully reviewing data instances and assigning appropriate labels or annotations based on predefined guidelines. Challenges may arise due to subjective interpretations, ambiguous cases, or evolving societal norms. Therefore, understanding the nuances and best practices in dataset annotation is essential to ensure reliable and representative training data, fostering the development of robust and accurate AI models.

Annotation of hate speech text could vary based on the content that is available to us. The authors in (Mulki et al., 2019) classified hate speech based in Arabic lexicon on the basis of Normal, Abusive or Hate while the works of (Albadi et al., 2018) had 234 different annotators classifying hate speech based on religious groups that are targeted.

While a single dataset can be used for annotation, (Gomez et al., 2020) evaluted four different hate speech datasets to create a benchmark dataset, MMHS150K dataset, containing an image with associated text. Real time tweets from September 2018 to February 2019 were used. A similar work by (Yang et al., 2019) explored different multimodal feature fusion strategies to generate a dataset with images and text. A majority voting was used as classification among five different classes.

Annotation of one dataset can have biases as some annotators might consider one example as a strong case of hate speech while another might have a milder response. To mitigate this, authors of (Haddad et al., 2019) used inter-annotator agreement on Tunisian hate speeh dataset to classify between Abusive Speech (AS) and Hate Speech (HS). The authors also used metrics of Cohen's Kappa (κ) and Krippendorff's alpha (α) to have consistency with the annotations.

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 Research Questions

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. nonhate 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?

With our research, we are able to answer to the above mentioned research questions. The answers for the first and second questions comes down to assigning a predefined category for each of our labels. Fortunately, the dataset, to annotate, is classified based on the Facebook Community Guidelines (FB) 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.

For the third question, to ensure high-quality annotations, we incorporated iterative feedback loops and inter-annotator agreement assessments to minimize subjectivity and enhance consistency. Furthermore, periodic meetings and discussions among the annotators were held to address challenges and clarify annotation guidelines, promoting a shared understanding of the task.

Finally, for the last question, the annotated hate speech dataset serves as a valuable resource for training and evaluating Large Language Models, which can be employed to automatically detect and filter out hate speech in online platforms. The relationship between the size of dataset and model performance has been left out as future work.

4 Experimental Setup

The dataset that needs to be annotated is **Hate-Moderate**. 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 Guidelines mentioned in the previous section. Our task was to label the examples as hate speech if they are in line with the category they are assigned.

The filtering of examples as part of a certain category was performed individually as well as facilitating inter-annotator agreement for difficult examples. In addition, we also leveraged the use of ChatGPT if we needed a broader and simplified definition of that guideline. The training examples were divided into three domains,

- **Easy:** These were the examples which were the easiest to annotate since they followed the exact meaning of guideline they were assigned.
- **Medium:** Such examples were not directly related to the guideline they were assigned because of a deeper sentiment or outright incorrect labeling.
- **Hard:** These examples used inter-annotator agreement and the most work to label.

Such categorization was necessary to avoid incorrect labeling of examples. Consider the following hate speech examples labeled under Dehumanizing speech - Feces. Examples of this category represent and attack to a person/group based on the words commonly used to describe fecal matter (i.e. shit, crap, etc.)

1. "*buddhists don't give a shit about other religions*"
2. "*Imagine being a middle class woman in America. Literally the easiest shit you could ask for.*"
3. "*So go ahead, go nuts, go apeshit, fucking nigs*"
4. "*Every black dude I've ever met was a piece of shit*"
5. "*those shitty windows need repairing, I'll get immigrants - will do a bad job but they're dirt cheap*"

Every statement uses a hate word used to describe fecal matter. However, we can see the meaning of all the statements and come up with a decision to evaluate and label the correct examples.

1. **Statement 1** uses the phrase "*don't give a shit*" as a vulgar expression to indicate a lack of concern, it is not directly related to the guideline it is referred to.
2. **Statement 2** uses the word "*shit*" as a general term to refer to something easy or trivial, it is not directly related to the guideline it is referred to.
3. **Statement 3** uses the word "*apeshit*" along with offensive language ("fucking nigs") to target and dehumanize people based on their race, it is not directly related to the guideline it is referred to.
4. **Statement 4** uses the phrase "*piece of shit*" to dehumanize black individuals, making it a derogatory statement. However, since the phrase could be viewed literally as well as figuratively, it needs to be discussed with the annotators.
5. **Statement 5** uses the term "*shitty*" to describe windows. The term used might refer to windows being non-functional. This statement is a derogatory statement as it refers to immigrants.

5 Results

Filtering for specific types of hate speech involving insects, bacteria, disease, feces, and certain objects was successful in locating some hate speech, but only enough that can be counted on two hands. The majority of the remainders were all hate speech, and yet they remain undetected and unmarked as such. Individual filters for hate speech may be futile, because there are limitless options to insult and hurt people, and no way to mark them all as such.

When it comes to filtering out hate speech in language, one can't simply find specific words like bug or crap, because they might be lacking in context. At the time, they might be talking about actual insects or excrement, instead of comparing people to them. In this case, the in depth analysis of the statement was used along with inter-annotator meetings.

Based on the project description, the task was to label examples of hateful speech based on whether they meet the criterion of Facebook's community standards guidelines for hate speech. The labeling was done on a Google spreadsheet, assigning a value of 1 if the example met the criterion and 0 if it did not. Here are the results for the given examples:

1. **Dehumanizing speech** or imagery in the form of comparisons, generalizations, or unqualified behavioral statements (in written or visual form) to or about **insects** (including but not limited to: cockroaches, locusts)

    - **Valid Examples (labeled as 1)**
        (a) "*sarah do you realise that we live in australia where flies, mozzies and numerous other subhuman insects live and exist? I know at my house if i leave the door open for 10 mins all hell breaks loose and their prayer music fucking swarms in*"
        This example targets a person or group of people (insects) using dehumanizing speech and imagery. It matches the criterion and is labeled as 1.
        (b) "*Calling immigrants 'coackroaches' is the most descriptive term - I feel like once you get rid of one there's always more*"
        This example targets immigrants and uses the term "cockroaches" to dehumanize them. It matches the criterion and is labeled as 1.
        (c) "*Keep dreaming you idiot!! And I'm NOT a Jew, I'm a realist. They're scattering like roaches,....afraid of all the BitChute vids and comments.*"
        This example targets a religious group (Jews) and dehumanizes them by comparing them to roaches. It meets the criterion and is labeled as 1.
    - **Invalid Examples (labeled as 0)**
        (a) "*Cockroaches are the lowest of the low, and should be treated as such.*"
        Although this example dehumanizes cockroaches, it does not target a person or group of people as specified in the criterion. Therefore, it does not meet the criterion and is labeled as 0.
        (b) "*This country is being overrun by a swarm of locusts, we need to control the borders.*"
        While this example uses the term "locusts" to metaphorically refer to a group of people, it does not target a specific person or group of people based on their protected characteristic or immigration status. It does not meet the criterion and is labeled as 0.
        (c) "*The streets are infested with vermin like cockroaches, it's not safe to go out at night.*"
        Similar to the previous example, this statement refers to cockroaches metaphorically as vermin, but it does not target a person or group of people based on their protected characteristic or immigration status. It does not meet the criterion and is labeled as 0.

2. **Dehumanizing speech** or imagery in the form of comparisons, generalizations, or unqualified behavioral statements (in written or visual form) to or about **animals** in general or specific types of animals that are culturally perceived as intellectually or physically inferior (including but not limited to: Black people and apes or ape-like creatures; Jewish people

and rats; Muslim people and pigs; Mexican people and worms)

- **Valid Examples (labeled as 1)**
  (a) "*n*ggers belong in the wild with the other apes*"
  This example dehumanizes a specific racial group by comparing them to apes and using a racial slur. It meets the criterion of dehumanization and is labeled as 1.
  (b) "*no matter what people say, i refuse to accept those black apes are equal to whites*"
  This example also dehumanizes a specific racial group by referring to them as "black apes" and asserting their inequality compared to another racial group. It meets the criterion of dehumanization and is labeled as 1.
  (c) "*peta campaigns against keeping wild animals in zoos are right, let's switch them for some n*gger apes! they are fucking abhorrent*"
  In this example, there is a dehumanizing comparison between animals and a specific racial group. It uses a racial slur and expresses contempt towards that group. It meets the criterion of dehumanization and is labeled as 1.
- **Invalid Examples (labeled as 0)**
  (a) "*if you think smart, blacks are the real danger to society*"
  Although this statement expresses a negative opinion about a racial group, it does not involve dehumanization by comparing them to animals. It does not meet the criterion of dehumanization and is labeled as 0.
  (b) "*blacks are irrelevant*"
  This example expresses a negative opinion about a racial group but does not involve dehumanization by comparing them to animals. It does not meet the criterion of dehumanization and is labeled as 0.
  (c) "*black people cannot be trusted*"
  While this statement makes a negative generalization about a racial group, it does not involve dehumanization by comparing them to animals. It does not meet the criterion of dehumanization and is labeled as 0.

3. **Expressions of disgust**, Expressions that suggest the target causes **sickness**, including but not limited to: vomit, throw up OR Expressions of **repulsion or distaste**, including but not limited to: vile, disgusting, yuck.

   - **Valid examples (labeled as 1)**
     (a) "*jewish people are vile*"
     This example expresses repulsion towards a specific religious group, referring to them as "vile." It meets the criterion of expressing disgust and is labeled as 1.
     (b) "*trans people are vile*"
     In this example, there is a statement expressing repulsion towards a specific gender identity group, labeling them as "vile." It meets the criterion of expressing disgust and is labeled as 1.
     (c) "*transvestites are vile*"
     Similarly, this example expresses repulsion towards a specific group, using the term "vile" to describe transvestites. It meets the criterion of expressing disgust and is labeled as 1.
   - **Invalid examples (labeled as 0)**
     (a) "*you are so vile*"
     While this statement expresses repulsion or distaste, it does not target a specific person or group based on their protected characteristic or identity. It does not meet the criterion of hateful speech and is labeled as 0.
     (b) "*he is so vile*"
     This example also expresses repulsion, but it does not target a specific person or group based on their protected characteristic or identity. It does not meet the criterion of hateful speech and is labeled as 0.
     (c) "*she is so vile*"
     Similarly to the previous examples, this statement expresses repulsion but does not target a specific person or group based on their protected characteristic or identity. It does not

meet the criterion of hateful speech and is labeled as 0.

6 Conclusion

This project focuses on the crucial task of annotating a hate speech dataset which can be further used for training Large Language Models for hate speech detection. We used **HateModerate**, which is a custom dataset with 41 different categories of hate speech taken from the Facebook Community Guidelines. Our team of 3 annotators used a classification scheme to divide the examples based on their direct meaning towards a certain topic and label each example as part of the category. To ensure high-quality annotations, we incorporated iterative feedback loops and inter-annotator agreement assessments to minimize subjectivity and enhance consistency. Finally, the annotated hate speech dataset serves as a valuable resource for Large Language Models for automatic hate speech detection.

7 Future Works

With this project, we also come across certain domains in which this work can be modified. We could make a system that links possibly insulting words and a target of harassment together. Paired with an automatic hate speech detector, the model can notify the user in real-time if they are writing a text that is against the guidelines. Moreover, we can also observe and evaluate the relationship between a model's performance and the dataset it is trained on to capture insights on the metrics we observe. Finally, an article on Science Direct (Charitidis et al., 2020) used a system for detecting hate speech in social media accounts. Something also notable is the fact that hate speech is detected in multiple languages, all of which are accounted for.

References

Facebook community guidelines. https://transparency.fb.com/policies/community-standards/hate-speech/.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. 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)*, pages 69–76.

Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karakeva. 2020. Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, 17:100071.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A tunisian hate speech and abusive dataset. In *Arabic Language Processing: From Theory to Practice*, pages 251–263, Cham. Springer International Publishing.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 11–18, Florence, Italy. Association for Computational Linguistics.