THE DUALITY OF HOPE: A CRITICAL EXAMINATION OF CONTROVERSIAL ANNOTATIONS IN HOPEEDI

Mohammad Aflah Khan*, Neemesh Yadav*, Diksha Sethi*, Raghav Sahni*
Indraprastha Institute of Information Technology
New Delhi, India
{aflah20082, neemesh20529, diksha20056, raghav20553}@iiitd.ac.in

ABSTRACT

This study investigates the HopeEDI hope speech dataset, revealing a significant number of potentially controversial annotations, notably tied to the 'All Lives Matter' movement. We have also identified instances where hateful/toxic/implicitly controversial content was wrongly marked as hopeful. The implications for deploying models trained on this dataset are profound, risking biases and stigmatization. We advocate for thoroughly examining the HopeEDI dataset, cautioning against biased models. We reannotate the hope speech and non-english labelled text, introducing a new class, 'Potentially Controversial', providing reasons for why the label was changed. This updated dataset aims to promote balance and mitigate ethical concerns in real-world applications.

1 INTRODUCTION

Recent years have witnessed a growing interest in the study of online communication, with a specific focus on fostering positivity and support. This paper delves into the HopeEDI hope speech dataset, a widely acclaimed resource for analysing positive online interactions. The scope of our study is directed towards the hope speech class within the HopeEDI dataset, guided by several key considerations. We're concerned about the risk of mistakenly promoting biased or controversial content due to false positives. Our decision to reannotate a subset of the dataset is motivated by the need for a focused and thorough analysis, enabling a nuanced understanding of hope speech annotations. Furthermore, we address the challenge of identifying concealed negativity within hopeful expressions, acknowledging the intricacies of discerning hidden meanings behind ostensibly positive language. We want to highlight the complexities of distinguishing between genuine hope speech and subtly veiled sentiments that usual classification methods might overlook.

2 RELATED WORK

This investigation delves into the concept of hope speech, specifically focusing on the dataset as it forms the backbone of any model built on it. Previous research, exemplified by (Palakodety et al., 2019), concentrated on analyzing trends in YouTube comments amid political tensions between India and Pakistan. However, the definition of hope speech in this research was confined to "web content which plays a positive role in diffusing hostility on social media triggered by heightened political tensions." Our study adopts a more extensive and inclusive definition of hope speech proposed by (Chakravarthi, 2020), characterizing it as "YouTube comments/posts that offer support, reassurance, suggestions, inspiration, and insight." The dataset employed in our study is the English subset of the HopeEDI dataset provided by the same source. This dataset has since been used in several shared tasks (Chakravarthi & Muralidaran, 2021; Chakravarthi et al., 2022; Kumaresan et al., 2023) as well as other works (Yadav et al., 2023). Discovering flaws in existing datasets is not unprecedented. As an example, (Blodgett et al., 2021) identified several issues with the CrowSPairs dataset (Nangia et al., 2020). Subsequently, these concerns were addressed in (Névéol et al., 2022), which not only corrected the dataset but also translated it into French while rectifying several annotations. Popular vision benchmarks (Luccioni & Rolnick, 2022; Northcutt et al., 2021) also find the same.

* Equal Contribution

Table 1: Samples from the Dataset: 0 indicates Hope Speech, 1 indicates Non Hope Speech, 2 indicates Non-English and 3 indicates Potentially Controversial

Statement	Old Label	New Label	Reason
All lives matter no matter color gender or religion bernies an effing idiot fishing for minority votes .another two faced demonrat	0	3	Could be in the backdrop of using all lives matter to downplay BLM. This is further validated by the remarks about Bernie
We all bleed the same. Stop with the Radical Fascist Liberal Left Political Pandering. ALL LIVES MATTER. We are all created by one true God in his image. We need to stop with all the hate. Period.	0	3	The emphasis on liberal and left politics in the statement indicates that the statement is highly likely to have come from a conservative. The 'All Lives Matter' slogan used in that context was often used to downplay the black lives matter slogan
@Vernice Davis I'm not white and I believe that all lives matter. Isn't it a bit racist to assume that all black people agree with your values?	0	3	Seems like an opposing comment to 'Black Lives Matter' and isn't actually intended to be hope speech.
I know that homosexuality exists	0	1	Presenting a fact, rather than hopeful or encouraging statement.
there's a fine line between disagreeing with them and being against them. she literally said she's alright with lgbt.	0	1	Nothing hopeful here. It seems like a response to someone

Prior work have also shown that biases in datasets propagate downstream and hence ensuring data quality becomes paramount (Hirota et al., 2022; Feng et al., 2023)

3 Annotation Methodology

Upon reviewing the original annotations, it became apparent they were poorly done, seeming almost random. Comments without hope were mislabeled, including those expressing resentment or anger. To establish a common taxonomy, all four annotators agreed on definitions and individually annotated 100 comments for pairwise inter-annotated agreement. The initial iteration had subpar agreement due to some annotators lacking context, but after several rounds, the final average agreement improved to 0.55 – considered moderate for such a subjective task. With satisfactory values, we proceeded to annotate the entire dataset. Annotators were also asked to provide reasoning for flipped labels in each sample. The annotators were allowed to research on topics they might be unfamiliar with and even discuss topics & samples amongst themselves to come up with fair annotations.

4 Observations and Discussion

Following the completion of annotations, the annotators engaged in discussions regarding the prevalent themes and common mistakes observed. We showcase some of these changes in Tables 1, 3. We also report the statistics of our annotations in Appendix A.2. Notably, many controversial annotations were associated with the 'All Lives Matter' movement, often perceived as undermining the 'Black Lives Matter' movement, likely mislabeled due to insufficient context. We, therefore, mark these as 'Potentially Controversial' as having them in the dataset and training models on them can make the models adverse to the idea of 'Black Lives Matter'. Additionally, instances of genuine hope speech were identified, particularly in comments discussing Madonna as an icon. Furthermore, several explicitly non-hopeful statements were incorrectly labelled as hopeful, lacking any basis for being considered as such. This shows that scrutiny of such annotations is paramount. We exclusively reannotate the hope-speech class (and the small set of non-English samples), recognizing that a false positive in this category has much more significant consequences. Using such mislabeled instances to promote comments could inadvertently amplify hate. The extensive size of the dataset, especially the non-hope class, makes it impractical to reannotate within our current setup. We tried to be as fair as possible, given the limited and incomplete contexts, while also keeping in mind the lack of external context about topics (as highlighted above).

5 Conclusion

We addressed inconsistencies in hope speech detection datasets by updating annotations for the 'Hope Speech' and 'Not English' classes, providing detailed rationales for each change. We also introduced a new 'Potentially Controversial' class highlighting ambiguous samples. These contributions improve annotation quality, raise awareness of inconsistency challenges, and promote transparency, paving the way for more robust models and future research.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

Bharathi Raja Chakravarthi. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pp. 41–53, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.peoples-1.5.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 61–72, Kyiv, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.ltedi-1.8.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José García-Díaz. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 378–388, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ltedi-1.58. URL https://aclanthology.org/2022.ltedi-1.58.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models, 2023.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Gender and racial bias in visual question answering datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, June 2022. doi: 10.1145/3531146.3533184. URL http://dx.doi.org/10.1145/3531146.3533184.

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Subalalitha Cn, Miguel Ángel García-Cumbreras, Salud María Jiménez Zafra, José Antonio García-Díaz, Rafael Valencia-García, Momchil Hardalov, Ivan Koychev, Preslav Nakov, Daniel García-Baena, and Kishore Kumar Ponnusamy. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In Bharathi R. Chakravarthi, B. Bharathi, Joephine Griffith, Kalika Bali, and Paul Buitelaar (eds.), *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 47–53, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2023.ltedi-1.7.

Alexandra Sasha Luccioni and David Rolnick. Bugs in the data: How imagenet misrepresents biodiversity, 2022.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8521–8531, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.583. URL https://aclanthology.org/2022.acl-long.583.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. Hope speech detection: A computational analysis of the voice of peace. *arXiv preprint arXiv:1909.12940*, 2019.

Neemesh Yadav, Mohammad Aflah Khan, Diksha Sethi, and Raghav Sahni. Beyond negativity: Re-analysis and follow-up experiments on hope speech detection. *CoRR*, abs/2306.01742, 2023. doi: 10.48550/ARXIV.2306.01742. URL https://doi.org/10.48550/arXiv.2306.01742.

A APPENDIX

A.1 DATASET DOWNLOADING

The dataset can be obtained at `https://github.com/aflah02/HopeEDI-Fix`. You can then use the README to understand how to merge this with the original dataset.

A.2 STATISTICS

We re-annotate a total of ‘2260‘ samples using the train and validation subset provided on HuggingFace Hub [1]. One important thing we wish to point out is that we only use the Hope Speech and Non-English labels for annotation, for reasons mentioned in Section 4.

We provide a graphical representation highlighting changes in the composition of the dataset after the reannotation process in Fig 1. We can see that around 16.24% of the Hope Speech class has now shifted to the Potentially Controversial class, and 12.54% has been relabeled as Not Hope Speech.

We believe this is a significant change, as during the process, we noticed a lot of false positives, which were originally labelled as Hope Speech, without looking into the proper context in which such language was used. However, we also note that one of the biggest challenges during reannotation was due to some of the incomplete and truncated contexts in the text.

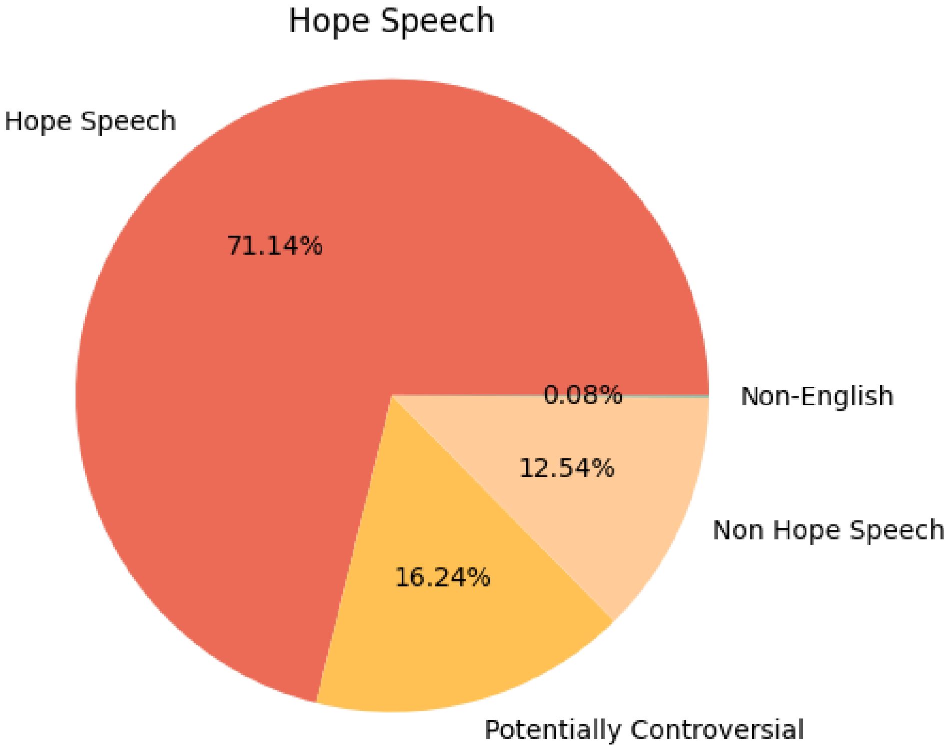

Figure 1: The change in labels between the original and the new annotation for the Hope Speech class

[1] `https://huggingface.co/datasets/hope_edi`

A.3 Dataset analysis

We provide a visual representation of an abstract representation of the newly annotated classes in our dataset[2]. Figures 2, 3, 4 show word-clouds for the 3 classes.

Figure 2 and 3 look very similar, hinting at the difficulty and subjective nature of the task. Simple presence/absence of words does not make something hopeful/non-hopeful. On the other hand, for the potentially controversial class, Figure 4 shows clear hints of topics related to BLM and Racism. Table 3 also shows a bunch of topic-wise samples to show some over-arching themes of the dataset.

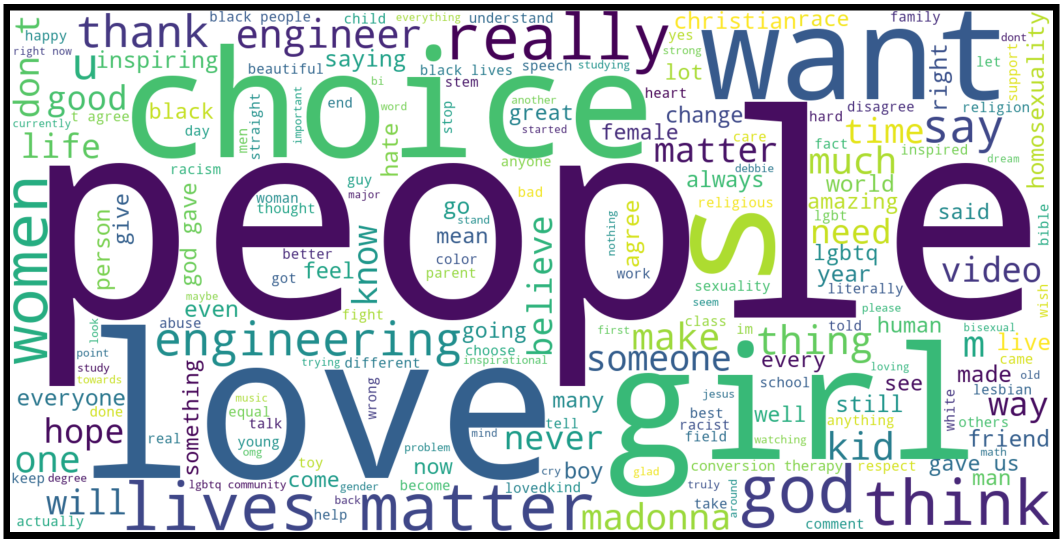

Figure 2: Word-cloud for the “Hope Speech” class

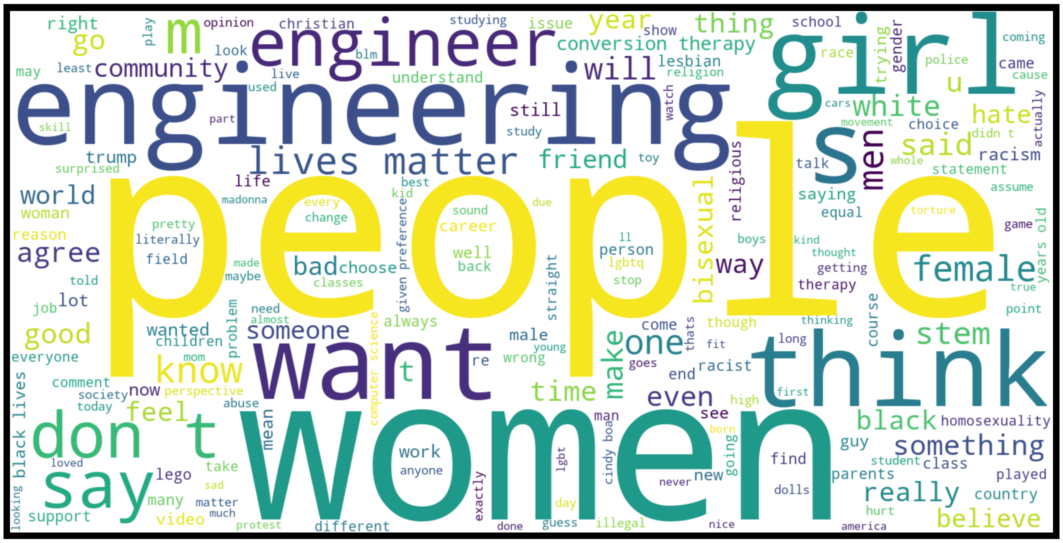

Figure 3: Word-cloud for the “Not Hope Speech” class

[2] Note that we only do this for the 3 arguably most important classes: Hope Speech, Non Hope Speech, and Potentially Controversial

Statement	Old Label	New Label	Reason
Racism is everywhere.Though Many of the BLM people are doing some retarded shit. You wouldnt understand how they felt unless you were black.	0	1	Not hopeful, goes onto abuse the people beinig talked about.
I find it difficult to say that I hate Islamic culture but not the individual people	0	1	Clearly not hope, talking about hating a certain culture.
When I was 12 I figured out that I was lesbian my mom said that I'm not her daughter anymore and sent me to my auntie in Newzeeland	0	1	Not hope clearly, highlights how parents can be disaproving.
I can't believe that people like me are put through such abuse. It breaks my spirit.	0	1	Not Hope, talks about the effects of abuse.
All lives matter . Including every race . Whites are the minority on this planet . Fac	0	3	Controverial, might be downplaying BLM.

Table 2: Some additional samples

Statement	Old Label	New Label	Reason	Overarching Theme
Racism is everywhere.Though Many of the BLM people are doing some retarded shit. You wouldnt understand how they felt unless you were black.	0	1	Not hopeful, goes onto abuse the people beinig talked about.	BLM
All lives matter . Including every race . Whites are the minority on this planet . Fac	0	3	Controversial, might be downplaying BLM.	BLM
Because being straight is the norm and people assume that everybody is straight. Because a lot of people are homophobic you grow up hating yourself and being fearful to say who you truly are	0	1	Sad as the person talks about hating themselves.	LGBTQIA+
When I was 12 I figured out that I was lesbian my mom said that I'm not her daughter anymore and sent me to my auntie in Newzeeland	0	1	Not hope clearly, highlights how parents can be disapproving.	LGBTQIA+
my parents threatened to throw me into conversion therapy for being trans but were too ashamed to. thankfully.nI'm in Canada btw–despite all our LGBT rights	0	1	Nothing hopeful rather seems like a negative experience	Conversion Therapy
i will never come out as bi to my grandparents because they are very proper and against LGBTQIA+. I came out to my dad and my sister. Im afraid that if i come out to them they will force me to go to conversion therapy.	0	0	Not hopeful, an unpleasant experience	Conversion Therapy
She comes across in this interview as relatable and human. Beautiful and creative. Thank you Madonna for sharing yourself with us. Xx	0	0	N/A	Madonna
I love Madonna. She truly is a freedom fighter and opened up doors for change. She was such a bright light for many people and gave them hope in a time where being different seemed not even an option at times.	0	0	N/A	Madonna

Table 3: Some overarching themes in the dataset

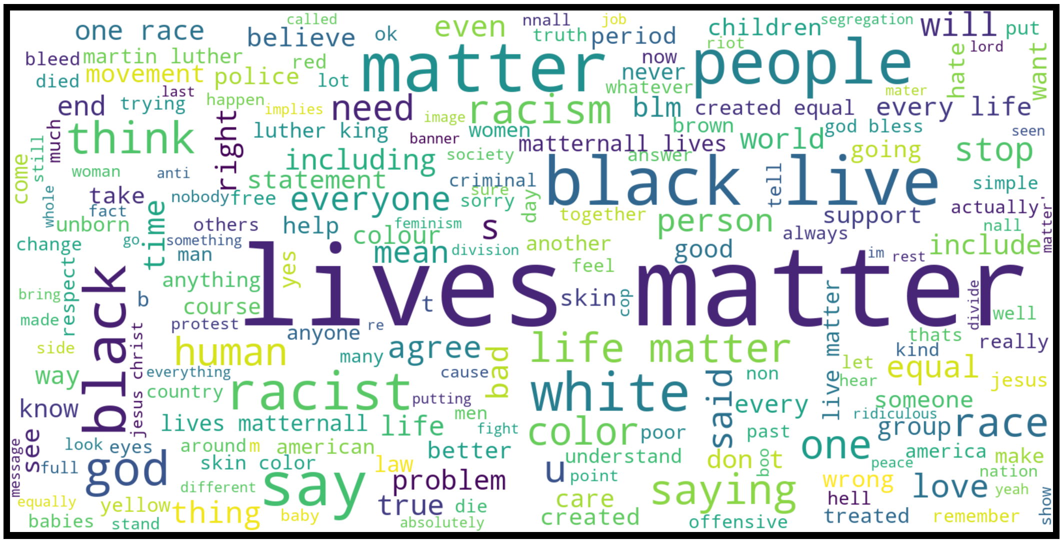

Figure 4: Word-cloud for the “Potentially Controversial” class

A.4 Limitations

While this study endeavors to shed light on the potential challenges associated with the HopeEDI hope speech dataset, it is essential to acknowledge certain limitations in our study:

1. Dynamic Nature of Language: The ever-evolving nature of language poses a challenge in accurately capturing and categorizing nuanced expressions. As language evolves, the dataset may become outdated, potentially leading to misinterpretations or misclassifications in the future.
2. Implicit Bias in Reannotation: While we aim to rectify issues in the original annotations, the reannotation process itself is susceptible to implicit biases. Despite our best efforts to mitigate subjectivity, some degree of bias may persist, influencing the revised labels.
3. Limited Linguistic Scope: Our reannotation efforts primarily focus on hope speech and English labeled text. Such issues may exist in other subsets too which require further investigation.