INFO 555 - Applied NLP Project 2

Kiwoon Hong

1 Introduction

Detecting hate speech on social media platforms is critical to maintaining a safe and inclusive online environment. Hate speech not only fosters discrimination and harm, but also leads to an imbalance in digital discourse by silencing the voices of certain groups. As social media has had a huge impact in modern society, hateful content can have far-reaching consequences, including social division and mental health impacts. Therefore, implementing effective hate speech detection is essential to encourage respectful interactions, protect communities, and ensure that social media remains a healthy platform.

This project aims to annotate hate speech by sampling tweets. However, detecting hate speech is a challenging task due to its complex and sensitive nature. Hate speech can take many forms, from obvious blames to subtle slang that only certain groups recognize. Cultural, linguistic, political, and regional differences also complicate the identification process, as what one group feels is hate speech may not be to another. This complexity often leads to problems with automated systems and human annotation, such as over-censorship or failure to detect offensive content.

Tweets Hate Speech Detection dataset[1] on Hugging Face contains a variety of tweets, including about 7 percent labeled hate speech, and 93 percent labeled non-hate speech.

2 Related Work

Hate speech annotation is an active area of research. In particular, the problem of hate speech ambiguity mentioned in the introduction has also been addressed. paper by Assimakopoulos et al. (2020), acknowledged this ambiguity and used 10 years of online commentary data and focused on context and hidden messages using Critical Discourse Analysis(CDA). As a result, they suggested that a multi-layer annotation scheme rather than a dichotomous distinction might be useful. Ron et al. (2023) decomposed the expressions into multidimensional factors. After filtering and sampling tweets about Jews, they decomposed the hateful expressions into five categories for annotation.

However, these two prior studies were characterized by a multilayered categorization of hate speech about specific groups. Therefore, it is difficult to generalize the classification of general tweets into specific groups and subcategories according to their respective distinction principles.

3 Procedure

3.1 Experiment

3.1.1 Data

This `tweets_hate_speech_detection` dataset is extracted from the Hugging Face dataset provided by Roshan Sharma, which is divided into a training set of 31,962 data and a test set of 17,197 data. The training set consists of English tweets from Twitter users and their hate speech labels (1 for hate speech, 0 for not hate speech), while the test set consists of tweets only. The distribution of labels in the train data is biased. Only 2,242 data points, about 7% of the data, are labeled as hate speech. Therefore, given that this may cause difficulties in annotation, the data was first undersampled to match the number of hate speech, and then resampled 300 random tweets. After checking the label distribution, the labels were removed for annotation, leaving only the tweet text for annotation.

3.1.2 Annotation Protocol

The project has a total of two annotators, with the annotator tagging the data themselves and one additional LLM acting as a reference. The first annotator is the researcher. He is a master's student

[1] https://huggingface.co/datasets/tweets-hate-speech-detection/tweets_hate_speech_detection

in data science, does not use Twitter, is not a native English speaker, and is not an American citizen, so he does not have expertise in American culture. The second annotator is Noah (an alias). He is an undergraduate physics student, has been on Twitter, is a native English speaker, and is a U.S. citizen. Both annotators shared the same guidelines for hate speech The data annotation was based on the definition of hate speech provided by the UN's `What is hate speech`[2] article. Only two labels were allowed: `'hate'` and `'nohate'`.
The LLM model used as a comparison group for data annotation is the HuggingFace's `roberta-hate-speech-dynabench-r4-target`[3] model. This model is from a paper by Vidgen et al. (2020), proposing a method for building dynamically trainable datasets and develops a model that can effectively handle new types of expressions and covert hate in hate speech detection.

3.2 Summary Statistics

The 300 tweets were tagged by two annotators and an LLM model. The researcher tagged 55 hate speech, Noah tagged 37, and the LLM model tagged 37. This means that 18.33%, 12.33%, and 12.33% of the total data was tagged as hate speech, respectively, which is significantly less than the previous 50% of undersampling.

3.3 Inter Annotator Agreement

The quality of the data was verified using IAA between the two annotators and between the annotators and LLM. Cohen's kappa score was used. In addition, tweets with tagging mismatches between two annotators and tweets with mismatches between annotators and the model were extracted for further analysis.

3.4 Demonstrating and Utilization

A simple Logistic Regression baseline model was created using the annotations of the two annotators, and the accuracy and the F1 score of the models were measured. Since there were no labels for the test part of this dataset, the test set was sampled by performing a sampling without replacement on the train part.

4 Results

The statistics and figures are attached in the appendix.

4.1 IAA

The Cohen's kappa score between the two annotators is 0.59, showing weak to moderate agreement. On the other hand, the researcher and LLM have a weak agreement of 0.46 and Noah and LLM have a weak agreement of 0.44.

4.2 Logistic Regression

Both models based on the data from the two annotators predicted the test data as 'nohate', resulting in low performance.

4.3 Error Analysis

Compared to Noah, the researcher was better able to identify non-American ethnicity discrimination and tag it as hate. For example:

```
#koreans &amp; joseon people in japan,
will abuse the  for claims of own
rights by rough demo.  #aberdeen
```

In contrast, Noah was better at recognizing hate speech related to US politics:

```
#intrumpsamerica extending your hand &amp;
uniting merica is unprecedented!
#rapist #egomaniac @user
```

5 Conclusion

Common hate speech had moderate matches. However, due to the the subcultural nature of social media and different sensitivities to specific category of hate speech even though they got the same guidelines, overall, the agreement was weak. Therefore, as mentioned in the related work above, detecting hate speech is a very nuanced task, and it is important to recruit annotators with diverse knowledge, cultures, environments, and memes.

5.1 Limitations

The dataset contained a large number of emojis, which had conversion issues due to Unicode and thus could not be fully interpreted. Also, both annotators could not understand tweets about cultures they did not belong to. Finally, a more appropriate demonstration model rather than a Logistic Regression model needs to be chosen based on less data.

[2] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[3] https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target

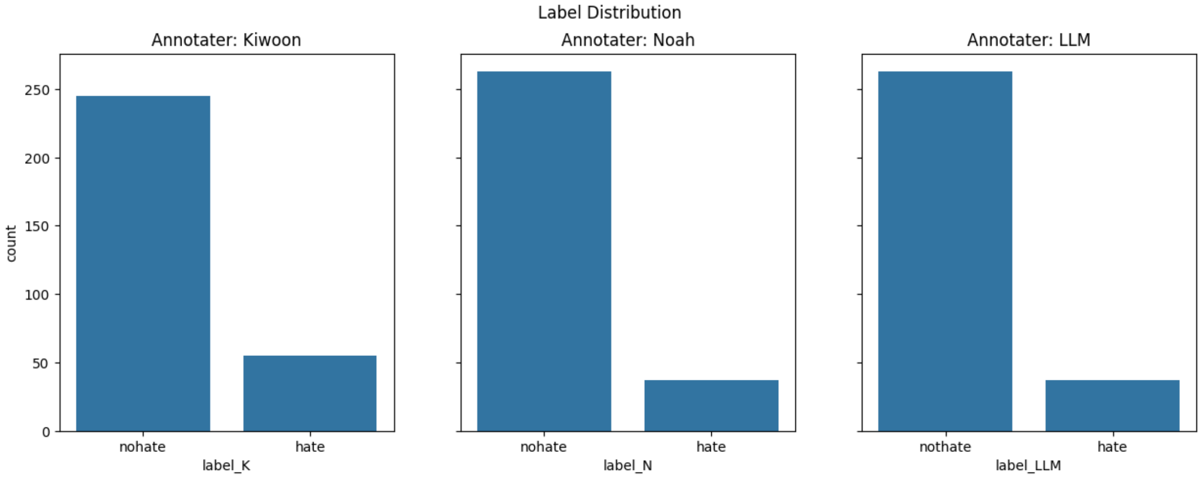
References

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5088–5097, Marseille, France. European Language Resources Association.

Gal Ron, Effi Levi, Odelia Oshri, and Shaul Shenhav. 2023. Factoring hate speech: A new annotation framework to study hate speech in social media. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 215–220, Toronto, Canada. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *CoRR*, abs/2012.15761.

A Appendix



Model	Accuracy	F1 score
Model_Kiwoon	55%	0.688
Model_Noah	52%	0.684

Table 1: Accuracy & F1 Score for each model based on two annotators.